Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Nr. 857

Herwig Unger,
Wolfgang A. Halang (Eds.)

Autonomous Systems 2017

Proceedings of the 10th GI Conference

FernUniversität in Hagen
Schriften zur Informations-
und Kommunikationstechnik

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Herwig Unger,
Wolfgang A. Halang (Eds.)

Nr. 857

Autonomous Systems 2017

Proceedings of the 10th GI Conference

FernUniversität in Hagen
Schriften zur Informations-
und Kommunikationstechnik

Herwig Unger, Wolfgang A. Halang (Eds.)
Autonomous Systems 2017 – Proceedings of the 10th GI Conference
Fortschr.-Ber. VDI Reihe 10 Nr. 857. Düsseldorf: VDI Verlag 2017.
276 Seiten, 91 Bilder, 26 Tabellen.
ISBN 978-3-18-385710-4, ISSN 0178-9627,
€ 95,00/VDI-Mitgliederpreis € 85,50.

Keywords: Autonomous Systems – Autonomous Driving – Big Data – Data Mining – Architectures – Applications – Security – Safety

Owing to the overwhelming amount of often sensitive data hitting contemporary users of networked devices almost everywhere, this volume's two keynote addresses deal with the security and ethical problems of the increasingly intelligent and autonomously acting devices in our environment. Then, light is shed on theoretical, algorithmic, technical and security aspects of big data analytics, data mining, information retrieval and machine learning. As no other area of computer science and engineering, autonomous systems accelerate the development of new systems' hardware and highly specialised applications with an unprecedented intensity. Thus, the corresponding topics addressed in this book range from computer architectures for embedded systems dedicated to safety-related automation applications and designed for verifiability and correct operation over approaches to suppress electromagnetic interferences and to interpret experimental data to questions related to their design and development, to building models and investigating their impact on society.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter http://dnb.ddb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at http://dnb.ddb.de.

Schriften zur Informations- und Kommunikationstechnik
Herausgeber:
Wolfgang A. Halang, ehemaliger Lehrstuhl für Informationstechnik
Herwig Unger, Lehrstuhl für Kommunikationsnetze
FernUniversität in Hagen

Als Manuskript gedruckt. Printed in Germany.

ISSN 0178-9627

ISBN 978-3-18-385710-4

Preface

The task is, not so much to see what no one has yet seen;
but to think what nobody has yet thought,
about that which everybody sees.

Erwin Schrödinger

Somehow, the life of a scientist may be compared with the one of a monk in a monastery: new ideas are usually born in solitude, after a longer meditation process inspired by an external event, but rarely by rational thinking. Once an idea is born, especially if it sounds whimsical, then it must be cared for (developed), it must be fought for its existence, and its correctness and feasibility must be proven in competition with others. Particularly in the very beginning, this process requires a unique atmosphere away from private troubles, jealousy, hurry, pressures and economic needs. And — differing from the practice of most other scientific conferences — it needs time and honest, non-egoistic and enduring talk among colleagues and friends. Nevertheless, like in nature or society, also in science there is competition of ideas for the best ones, and only a few will survive.

For ten years now, the former PhD seminar, workshop and today's conference on Autonomous Systems has cultivated exactly such an environment. A growing number of colleagues, in this year coming from three continents, enjoys quiet moments in the inspiring nature of beautiful Majorca Island as well as the open, censor-free, not always politically correct, but (almost) unlimited discussions with colleagues and friends. Again, the 20 contributions of these proceedings, which were intentionally not peer-reviewed, but only checked for technical soundness and plagiarism, exhibit a variety of aspects related to the conference topics.

Owing to the overwhelming amount of (sensitive) data hitting today's users of most networked devices almost everywhere in this world, we decided to commence this volume with two keynote addresses related to security and ethical problems of the more and more intelligent and autonomously acting devices in our environment. We hope that these contributions may trigger intense interdisciplinary discussions.

A section on Big Data and Data Mining follows, touching technical and algorithmic details of data processing. Since most problems cannot be tackled anymore without proper mathematical background, the border to this book's subsequent part on Theory becomes more and more blurred. As no other area of computer science and engineering autonomous systems accelerate the development of new systems' hardware and highly specialised applications with an unprecedented intensity. This may be the reason why the concluding section Architectures and Applications turned out the most voluminous one spanning topics from processor architecture to electrical engineering, and showing the need for extended interdisciplinary cooperation between scientists.

Again, a PhD session and a tutorial are conducted to attract our youngest group of attendants, to encourage them and to provide them with hints for their further research and publication work.

Last but not least, we want to extend a sincere Thank You to Jutta Düring and Barbara Kleine, who worked hard in the background preparing these proceedings as well as setting up every, maybe not immediately perceivable detail of our event in a perfect manner. In addition, we appreciate the support of Fern-Universität in Hagen given to publish this volume.

Hagen, August 2017

Herwig Unger
Wolfgang A. Halang

Contents

Keynote 1

Keynote 2

Big Data and Data Mining

Theory

Architectures & Applications

Autonomous Components – A Growing Security Issue for Critical (Information) Infrastructures

Gerald Quirchmayr

Multimedia Information Systems Research Group
Faculty of Computer Science, University of Vienna, Austria

Abstract: The increasing number of attacks on critical information systems infrastructures and the advanced nature of some of these attacks have led to an uneasy feeling about the security of autonomous components embedded in these systems. While awareness of the growing threat seems to be increasing, a passive attitude towards the problem still seems to prevail. While in traditional office management and management systems environments the issue of information security is to a large extent dealt with at a professional level, core manufacturing systems still remain poorly protected, causing a huge problem for modern smart supply chains and smart architectures. This paper does consequently look at the problem posed by autonomous and inadequately protected autonomous components highly connected environments. It then describes some selected relevant cases before moving on to best practice guides and legislative efforts aimed at effectively countering the increasing threat from criminal organizations.

1 Background and Motivation

According to recent cyber security studies of leading international police forces, such as EUROPOL, the threat posed by cyber-attacks on IT systems information systems infrastructures is increasing dramatically. According to the IOCTA 2016 report [1] the crime areas most prevalent are still traditional hackers and organized crime. There however is a new worrying trend in the form of so called ″hacktivism″, politically motivated attacks and the growing involvement of state actors. With cases such the attacks on Sony [2], supposedly originating from North Korea, APT1 [3], attributed to Chinas, the cyber operations in Ukraine [4], carried out by Russian separatists, and the now legendary NSA files [5], the situation has deteriorated rapidly. The worst resulting development is the weaponisation of cyber space, with a substantial portion of the arsenal not being protected well enough and ending up in the hands of criminals, as the recent WannaCry episode [6] has amply demonstrated.

At a time when the danger is mounting, the worldwide negligence of system and infrastructure providers is shocking [6]. Not only remain systems unpatched, even after the WannaCry attack causing quite serious damage around the world, the response to the increasing threat is a long way from where it should be [6].

Set against this background of a continuously worsening situation this paper examines some of the core threats to autonomous components and points at some possible counter measures that do exist but are, for whatever reason, not used with the intensity they should be.

2 Developing Threats

According to recent cyber security studies of leading international police forces and warnings released by these organizations [6, 7] "This type of crime is characterized – as hardly any other – by a continually rising crime rate" [8]. The threat vectors identified in the reports include in the key findings that Cryptoware (encrypting ransomware) has become the most prominent malware threat, DDoS attacks continue to grow in intensity and complexity. Data remains a key commodity for cybercriminals; however data is no longer just procured for immediate financial gain. Increasingly it is acquired for the furtherance of more complex fraud, encrypted for ransom, or used directly for extortion. When considering intellectual property, the illegal acquisition of this data can reflect the loss of years of research and substantial investment by the victim. The core conclusion that "Cybercriminals use whatever communication method they perceive to be sufficiently secure" is an indication of what needs to be strategically done to effectively combat cybercrime. According to [6], "While securing critical infrastructures remains a private sector responsibility, attention should be given by regulators to the compliance of IT systems and mandatory security-by-design".

It is this situation which needs to be brought under control, because as long as critical infrastructure is maintained by private companies being under enormous pressure to cut development and operational costs in the interest of shareholder value, investment in security will remain at the bare minimum prescribed by law. Whenever IT infrastructure that is critical for the operation of business and society comes into the picture, the importance of a functioning security approach cannot be underestimated.

3 Autonomous Components and Core Infrastructure as Critical Vulnerability

Autonomous components are in the long run the only commercially and technologically viable way of providing the high level of operational performance needed in a highly networked economy. [SG-CG/D] SG-CG/M490/D_Smart Grid Information Security, as referenced in CEN-CENELEC-ETSI Smart Grid Coordination Group Smart Grid Reference Architecture [9] shows how security needs to be included from design to implementation. The currently still used approach of retrofitting security can only be perceived as a temporary solution that needs to be phased out quickly. Reality however could not be more different with one of the world's most advanced and only recently launched aircraft carrier types still running Windows XP. With this story hitting the general media immediately [10], the current perception of the engineering community's attitude towards cybersecurity has hit a new low.

These two examples show the possibilities for solid security aware system development currently at hand and the sad fact of new technology with advanced security features often not being implemented.

Considering the example given in immediately [10], it becomes painfully clear that a new attitude needs to be developed. While old by now outdated systems do in many cases still form the backbone of smart infrastructures, at least new technology being brought in should have a very different level of security. With core network components being a prime target for attackers, the need to adequately protect them is painfully clear. Past examples of DNS attacks, even in the rather simple form of DDoS attacks gain new power through the ample availability of completely unprotected IoT devices [11]. With more and more critical hubs operating in fully automatic or in semi-automatic mode, such as Internet exchanges, cloud data centers, and transport management and control systems, the danger posed by an intruder is increasing dramatically. Additionally, smart manufacturing / Industrie 4.0 will introduce new vulnerabilities with the move to just in time manufacturing and lean supply chains resulting in the removal of redundancies that could previously serve as fallback infrastructures. The recent WannaCry episode [6] has on top of all other security fears made clear that neither basic system patching nor backup requirements are being fulfilled by system operators for cost saving reasons, the dire consequences of such negligence are obvious. While the effects of a potential large scale blackout have been amply discussed in the media and in numerous publications, the danger of fully autonomous network control systems is still underestimated. It has to be finally accepted that the security of almost every critical infrastructure depends on the security of the information flow as much as it does on the security of the energy supply and the security of the material flow.

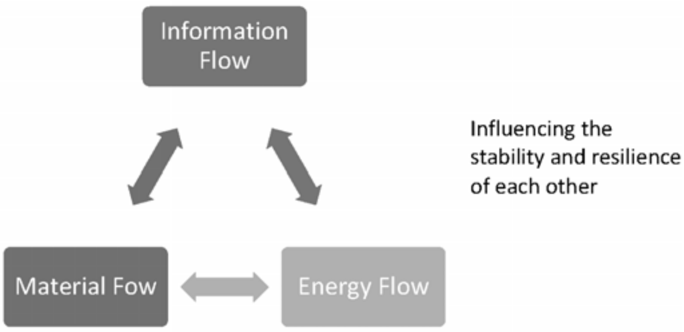


Fig. 1: Mutual dependence

Under the assumption of autonomous systems being at the core of these three mutually dependent flows, a major disaster becomes unavoidable, unless security in this highly connected environment is taken seriously and stops to be considered as yet another overhead type of requirement that should be met at the minimal possible cost.

4 Advanced Persistent Threats — the Beginning of a New Level of Attacks

Recently receiving more attention, APT attacks (Advanced Persistent Threats) form the spearhead of a new generation of attack patterns that employ sophisticated technology and are very well resourced. After the APT1 case received intension coverage, it was still almost two years until Switzerland should discover one of the worst cyber-attacks in Europe. Being based on the Turla family of malware [12] and better known under its codenames APT28 and APT29 the malware was used to infiltrate military and industrial systems in a neutral country, thereby crossing an internationally completely unacceptable border. It was the ensuing deep analysis of the case by MELANI:GovCERT (`https://www.GovCERT.ch`), the Swiss national CERT that sent a shock wave through Europe. One of the next small nations coming under attack, this time from far less professional aggressors, was Austria, having to suffer amongst other targets, from a severe cyber-attack against it major nation airport having a clearly political motivation. One of the latest cyber-attacks also targeting the nuclear power plant in Chernobyl [14] has again documented that cybercriminals show no limitations.

With military style campaigns being launched against civilian cyber infrastructures, in some extreme cases state actors engaging in outright criminal activities, the quick revelation of the modus operandi of the attackers is the best approach to prevent further success of the perpetrators. In that respect the highly professional Swiss reaction can serve as a model.

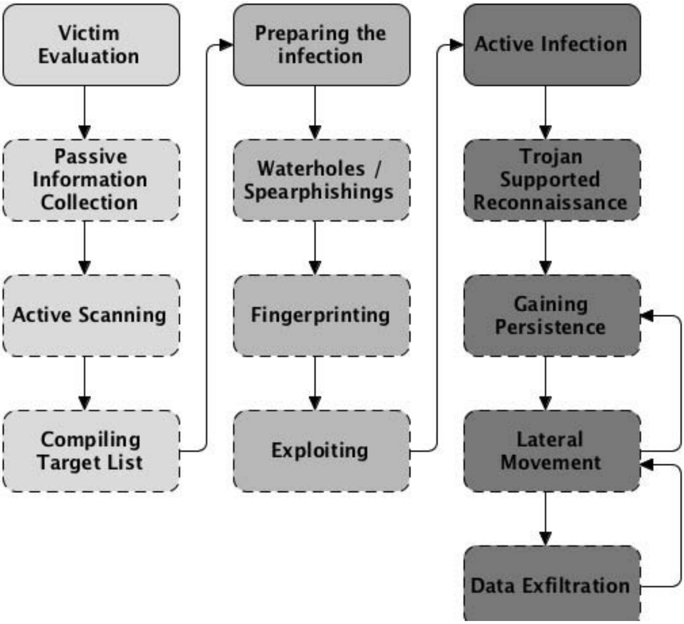


Fig. 2: Modus Operandi / Phases of the attack according to [12]

Knowing these three main phases and the task carried out in these phases now makes it possible for system operators to strengthen their environments and raises the chances to intercept this new dangerous form of attack.

Airports and nuclear plants coming under attack from madmen and criminals has once more made the necessity for protecting any embedded autonomous components absolutely clear. The increasing dependence on centralized cloud data centers, centrally maintained and controlled smart manufacturing production lanes and supply chains clearly necessitates a more serious and comprehensive approach to cyber security. That is why in a concerted European effort under the leadership of ENISA, the cyber security priorities ware identified and, based on the European Cyber Security Strategy [15], the NIS Directive [16] was launched as first piece of strategic legislation for making the European cyber space a safer environment. The areas of interest as identified by ENISA [17] comprise Critical Infrastructures and Services and IoT and Smart Infrastructures, with Critical Information Infrastructures, Internet Infrastructure, ICS SCADA, and Smart Grids being some of the most important focus areas.

5 The Way towards Effective Counter Measures

The way towards finally solving the problem will undoubtedly be a hard and long one, but the response is building up in the form of preparing for scenarios that were only a few years ago deemed as unrealistic. "While a new ransomware campaign (Petya) is still ongoing, and a few weeks only after the WannaCry outbreak, the report sheds light on the preparatory steps taken by authorities and industry to respond to such cyber-attacks." [18].

Legislation and best practice guides introduced over the past decade can today serve as an excellent basis for effectively countering cyber-attacks. The EU NIS Directive, being a legislative milestone, finally introduces a framework for CSIRT cooperation that enables and encourages the much needed exchange of information, which is the cornerstone for effectively countering threats [16].

The core role of CSIRTs/CERTs is best described in [15] as shown in Fig. 3.

Best practice guides, such as the Swedish Guide to Increased Security in Industrial Control Systems [19] and the NIST Guide to Supervisory Control and Data Acquisition (SCADA) and Industrial Control Systems Security [20] are some of the leading examples of guidebooks that have been around for some time and are now well established. The basic work however still needs to be done by every system operator of a critical infrastructure in the form of security policies

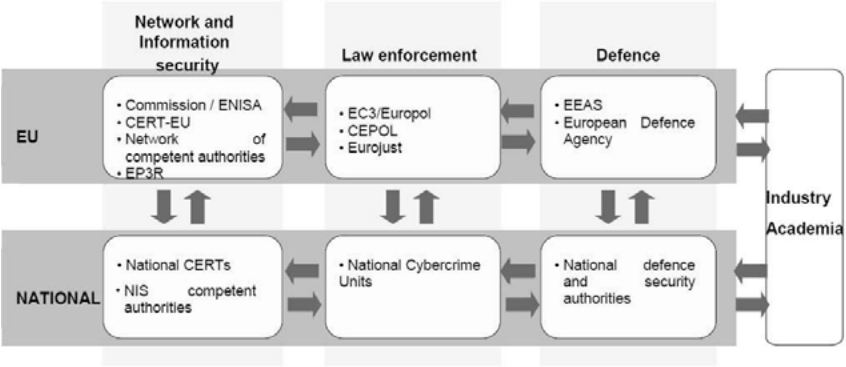

Network and Information security
Law enforcement
Defence
EU
NATIONAL
• Commission / ENISA
• CERT-EU
• Network of competent authorities
• EP3R
• EC3/Europol
• CEPOL
• Eurojust
• EEAS
• European Defence Agency
• National CERTs
• NIS competent authorities
• National Cybercrime Units
• National defence and security authorities
Industry
Academia

Fig. 3: The role of CERTs/CSIRTs

and security education and awareness (SETA) programs. Only these policies and programs can build the necessary foundation for a successful enterprise wide approach (as laid out in [21]).

6 Conclusion / Need for Action

While technology and best practice guides are in place, the current commercial pressures on system and service providers often stands in the way of equipping autonomous components with the necessary security elements. Given the recent worldwide attacks including critical infrastructures it is evident that cyber criminals do not respect any limits, running ransomware attacks even against hospitals and nuclear power plants. It is therefore a more than necessary reaction of the European Union to introduce obligatory standards after it has become clear that a purely voluntary approach does not work.

References

[1] The 2016 Internet Organised Crime Threat Assessment (IOCTA), `https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2016`

[2] Sony Pictures struggles to recover eight days after cyber attack, `http://www.reuters.com/article/us-sony-cybersecurity-investigation-idUSKCN0JG27B20141203`

[3] APT1 Exposing One of China's Cyber Espionage Units, http://www.mandiant.com/apt1

[4] Kenneth Geers (Ed.), Cyber War in Perspective: Russian Aggression against Ukraine, NATO CCD COE Publications, Tallinn 2015.

[5] The NSA files, https://www.theguardian.com/us-news/the-nsa-files

[6] Hackers Hit Dozens of Countries Exploiting Stolen N.S.A. Tool, https://www.nytimes.com/2017/05/12/world/europe/uk-national-health-service-cyberattack.html?action=click&contentCollection=World&module=RelatedCoverage&region=Marginalia&pgtype=article

[7] FBI IC3 2016 INTERNET CRIME REPORT, https://pdf.ic3.gov/2016_IC3Report.pdf

[8] BKA, https://www.bka.de/EN/OurTasks/AreasOfCrime/Cybercrime/cybercrime_node.html

[9] CEN-CENELEC-ETSI Smart Grid Coordination Group Smart Grid Reference Architecture, http://ec.europa.eu/energy/sites/ener/files/documents/xpert_group1_reference_architecture.pdf

[10] http://www.telegraph.co.uk/news/2017/06/27/hms-queen-elizabeth-running-outdated-windows-xp-software-raising/

[11] LizardStresser IoT botnet launches 400Gbps DDoS attack, http://www.computerweekly.com/news/450299445/LizardStresser-IoT-botnet-launches-400Gbps-DDoS-attack

[12] Technical Report about the Espionage Case at RUAG, https://www.melani.admin.ch/dam/melani/en/dokumente/2016/technical%20report%20ruag.pdf.download.pdf/Report_Ruag-Espionage-Case.pdf

[13] https://kurier.at/chronik/oesterreich/cyber-angriff-auf-flughafen-kam-aus-der-tuerkei/219.822.262

[14] http://www.independent.co.uk/news/world/europe/chernobyl-ukraine-petya-cyber-attack-hack-nuclear-power-plant-danger-latest-a7810941.html

[15] JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace, http://ec.europa.eu/information_society/newsroom/cf//document.cfm?doc_id=1667

[16] The Directive on security of network and information systems (NIS Directive), https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive

[17] EINSA, https://www.enisa.europa.eu/

[18] https://www.enisa.europa.eu/news/enisa-news/cyber-europe-2016-key-

lessons-from-a-simulated-cyber-crisis

[19] Guide to Increased Security in Industrial Control Systems, Published by the Swedish Civil Contingencies Agency (MSB), ISBN: 978-91-7383-089-8, Publ. no: MSB 0184-10.

[20] Guide to Supervisory Control and Data Acquisition (SCADA) and Industrial Control Systems Security, NIST Special Publication 800-82.

[21] Norm ISO/IEC 27001 Information technology – Security techniques – Information security management systems – Requirements, https://www.iso.org/isoiec-27001-information-security.html

Ethics in Autonomous Driving

The Report of the German Ethics Commission — And Questions to be Discussed

Theodor Tempelmeier

University of Applied Sciences, Rosenheim, Germany

Abstract: Autonomously driving cars require adaptations to the legal system, but also ethical considerations. This contribution presents some aspects of the recently published report of the German Ethics Commission on Automated and Interconnected Driving. Beyond this general report some additional questions — as specific for the computer science community — are proposed. The purpose of these questions is to foster discussion especially among the computer science community and from a computer science point of view.

1 Introduction

Autonomously driving cars are the current hype, so there is no need to emphasize the necessity of discussions concerning the legal and ethical implications of such systems. The German Ethics Commission on Automated and Interconnected Driving [1] has delivered its report recently[1]. Results of this report are presented in summarized form. And a series of questions — directly related to the report or beyond it — are posed. Mostly, no answers are given to these questions, yet; instead, it is the idea behind this contribution to foster discussions about ethics among all computer scientists.

2 The Report of the Ethics Commission

The work of the German ethics commission on automated and interconnected driving and the resulting report [1] are highly appreciated. Only a short overview of the report (referred to as "the ethics report" in the following) is given here. The reader is strongly encouraged to read the original report for himself or herself.

[1] In fact, the report was only released days before the deadline of this conference. As a consequence this contribution does not yet contain a full analysis of the report. A more thorough investigation will hopefully evolve during discussions at the conference and will possibly be published after the conference in an extended version of this contribution.

2.1 Levels of Automated Driving

Automated driving is commonly classified into different levels depending on the degree of automation [2]. The ethics report has its focus on levels 4 and 5 of automated driving [1, p. 14], that is on full automation and driverless systems. The difference between the two levels is that in level 4 automation is only capable of handling certain use cases, excluding for instance certain road types, speed ranges, and environmental conditions. In contrast, in level 5 all possible use cases and all driving modes have to be handled.[2]

In this contribution level 3, high automation[3], is also considered. In level 3, the human driver is expected to respond appropriately to a request to intervene, if the automated system cannot handle the situation. The German road traffic act, in the newly introduced paragraph 1a, requires for driving in high automation or full automation (that is levels 3 or 4) to signal to the driver with "sufficient" buffer time, if it is necessary for the driver to take over [5, para. 1a]. That would mean only level 3, at first glance. However with the end of a "use case" in level 4, the driver is also required to take over.

2.2 Ethical Rules

From the ethics report the following list has been excerpted and prioritized according to the personal view of the author; as such this shortened overview is also subject to discussion.

Protection of Human Beings. The protection of human beings is of highest priority, above any other considerations of usefulness, above any property damage, or damage to animals [1, rules 2 and 7].

No Distinction of Human Beings. The report forbids any distinction among human beings with respect to age, gender, physical or mental constitution in unavoidable accident situations.

Also forbidden is to offset the (number of?) injured or killed persons; but on the other hand, minimizing the number of injured or killed humans as a

[2] Note however that different terms are used in various contexts: In SAE terminology [3] levels 4 and 5 are called high automation and full automation, while in VDA terminology [2] high automation and full automation only mean level 3 and level 4. See [4] for a comparison of terminology. The difference in terminology is considered awkward and painful, but a decision has to be made, and in this contribution the VDA terminology [1, 2] is used.

[3] conditional automation in [3]

general programming principle may be allowed [1, rule 9]. Non-involved parties must not be sacrificed.

Transparency. Transparency of the technology in use has to be provided to the public [1, rule 12].

Privacy and Security. The report puts a very strong emphasis on privacy of personal data [1, rules 13 and 15, chap. 9 and 10], especially with respect to interconnection among vehicles or with a central infrastructure.

Freedom in the decisions of human beings. In a free society individuals should have the freedom to act on their own responsibility. An imposed obligation to use automated and interconnected driving systems may be ethically questionable [1, rules 4 and 6, chap. 5 and 7]. See also [1, chap. 3] for a discussion whether the human driver should be allowed to override the automated driving system.

Dilemma Situations. Dilemma situations, such as a decision between one human life and another, cannot be clearly standardized, nor can they be programmed such that they are ethically unquestionable [1, rules 8 and 5, chapters 1.3 and 1.4].

Handover in Emergency Situations. Highly automated vehicles (i.e. automated driving levels 3 and upward) must be designed such that the need for an abrupt handover of control to the driver (emergency) is virtually precluded. In an emergency situation the vehicle must autonomously enter into a safe state [1, rules 17 and 19].

3 Questions to be Discussed

Many questions arise from the ethics report and in general about automatic driving. The following list is given from a computer science point of view. Note that mostly no answers are given, although the author does have a firm personal opinion on all questions. However, a personal opinion is not an answer. To get answers to these questions, a discussion among all members of the computer science community is necessary. The arrangement of items in the following is prioritized according to the personal view of the author and is as such also subject to discussion.

Acquiring a Comprehensive Environment Model. Autonomously driving cars are dependent on a model of their environment. The various sensor signals are processed by a variety of algorithms, just described with a few keywords here: filtering, feature extraction, classification, pattern recognition, machine learning, neural networks, random forests, and many more. These algorithms are in many cases heavily dependent on *heuristic* or *probabilistic* methods. See e. g. [6, 7].

Q.: Do we really want to tie our lives to such heuristic, probabilistic methods?

Q.: Is it enough "proof" of concept, if x thousand hours of simulated driving have worked well? Or are 10x thousand hours of tests needed instead?

Hint: The ethics report states that self-learning systems "may be ethically allowed" and that they must meet the safety requirements regarding functions relevant to vehicle control. [1, rule 18 and chap. 11.3]

Transparency. The demand for transparency of the technology in automated driving must be set in relation with the preceding item (acquiring an environment model). Consequently this leads to the following questions.

Q.: Can the heuristic and probabilistic methods be explained to the public? To the judges? To computer scientists (when not specializing in this field)?

Q.: Will there be dilemma situations as often quoted (kill A or B) at all? Will there not rather be a different setting: some really obscure "cost function" with a number of constraints (computed somehow over the last few time frames and based on heuristic and probabilistic methods of object classification) forming the basis for decisions?

Note on the First Two Points: The need for transparency is widely acknowledged, see e.g. [6, 8]. The plethora of methods for pattern recognition, classification, for machine learning, etc. in combination with heuristic and probabilistic algorithms seems to make it difficult to achieve "transparency". *So this is the crucial central point about automated driving.*

Safety—Safe Hardware and Safe Software. Obviously safety in automated driving is a major concern.

Q.: Is there a safe state in driving? Stopping? Stopping on a German autobahn? Are fail-safe systems sufficient? Or are fail-operational systems needed? How much redundancy is necessary or economically feasible?

Q.: Is it ethically acceptable to continue using standard hardware, commercial off the shelf (COTS) systems, etc. instead of striving after inherently safe hardware? See for instance [9, 10] for aspects of safer hardware.

Q.: Are the currently available "safety" processors safe enough?
Is lock-stepping and error-correcting memory enough?

Q.: Is it ethically acceptable to continue using old, inapt, inherently unsafe programming systems?
Is tinkering with a language like C (resulting in MISRA-C) ethically acceptable?
Shouldn't safer languages such as Ada, SPARK-Ada, PEARL be a mandatory choice when thinking ethically?

Q.: Who defines the "state of the art"? Is MISRA-C state of the art?

Offsetting Human Beings against Each Other. The rules in the ethics report to give highest priority to protecting human beings (above any property damage etc.) and to forbid offsetting of human lives seem to be unquestionable at first, but ...

Q.: What should be done in a situation where one fatality has to be weighed against a huge property damage, which in turn may cause numerous fatalities (e.g. when hitting a big gas station ...)

Hint: The ethics report has mandated to give highest priority to protecting human life (see also rule 7). The reasoning is that this is a solution with the greatest potential to minimize harm in most cases, while foreseeing and weighing consequential damage of an alternate decision is technically not realizable at the current state of the art.

Q.: Is offsetting also to be forbidden when thinking of misbehaving parties? That is, could a road user ignoring the right of way have a lesser chance to survive as compared to others who strictly adhere to the traffic laws?

Q.: Is there a contradiction in [1, rule 9], when it says “no offsetting, but minimizing the number of injured or killed humans as a general programming principle may be allowed”?

Q.: Is it acceptable to assign higher rank to self-protection of the passengers of an autonomously driving car than to protection of non-involved persons? (See e.g. [11] and [6, chap. 4.2.3].)

Hint: The ethics report confirms that self-protection is not per se subordinate to protection of others; however, parties not involved in the generation of mobility risks must not be sacrificed [1, chap. 1.7].

Freedom of Decisions of Individuals. The ethics report emphasizes the freedom of the individuals (being allowed to act on their own responsibility, not being forced into automated driving systems), and this is a noble goal, but ...

Q.: Is it realistic to assume that the authorities and the (insurance) companies will adhere to these demands? Now and in the long term? See for instance [12] for a view of the German chancellor and the industry about the things to come.

Q.: Should overriding of the automatic driving mode be allowed?
Just for fun (because one has a car with lots of horsepower)?
For safety reasons (taking into account misbehaving parties, e.g. a driver on a head-on collision course, in combination with the idea “I can do better and I know how to survive”)?
For a smooth ride (switching off distance control to avoid emergency braking due to drivers cutting too close into one’s lane)?
To violate traffic laws (either because of the freedom to be reckless or in order to go to a safe place, e.g. taking an autobahn parking lot though it is officially closed)?

As an aside: Would the autonomous driving system ever violate the traffic laws, if this would presumably help to prevent fatalities or great property damage?

Handover in Emergency Situations. The ethics report is in a way fuzzy about the levels 3, 4, 5 of automated driving. While chapter 1.1 and figure 1 in the report seem to limit itself to levels four and five, rule 17 clearly addresses level three. At level three (high automation) the human driver must be able to take over control if so mandated by the autonomous system.

Q.: How is an emergency situation to be detected? Will there be graceful degradation, redundancy, fail-operational systems? Is there a safe state? Is a safe state reachable in due time after the autonomous driving systems has detected the emergency situation?

Q.: What is a reasonable time interval for allowing the driver to take over control in an emergency?

Hints: The ethics report states that handover must not be "abrupt".
Figures about how long it takes for a driver to fully comprehend the traffic situation and to physically take over control after an alarm vary: 2-3 seconds in [8] or 5-10 seconds in [6].
Another source states: "The idea of handing back control to a driver who has not been driving, in an emergency situation, is fundamentally a very bad idea" [13, video at 1:34].

Privacy of Personal Data. The demands to respect privacy of data in the ethics report is acknowledged, but ...

Q.: Is it realistic to assume that companies and authorities will adhere to these demands? Now and in the long term? Also in non-automatic driving mode?

Q.: Is it realistic to assume that suddenly the manufacturers will be able to provide waterproof secure systems?

4 Legal Aspects

The legal aspects of automated cars are not within the scope of this contribution. However, a few hints shall be given.

The problem of civil liability may be comparatively easy, as the insurance companies may probably accept to take the risk to provide coverage, because overall damage will probably be reduced by autonomously driving cars.

Another problem is criminal liability: Who is to be prosecuted in the case of a (fatal) accident? The car manufacturer, the overall system designer (whereby systems are comprising hardware and software), the software/hardware designer or programmer, the employer of the software/hardware programmer, the software/hardware subcontractor, the supplier of a particular sensor or actor?

The ethics report provides some guidance on liability in rules 11 and 10, and in chapter 11. Clearly must it always be possible to identify, who is (or was) the responsible driver—the human or the automated driving system [1, rule 16 and chapter 4].

5 Summary

The author hopes that intensive discussions in the computer science community will follow and answers to all open questions will arise, even if it may be only possible to admit that no answer can be given (yet). And the author does hope that everybody involved in construction autonomously driving cars will act in an ethically responsible way.

References [6, chap. 4 and 5] and [14–17] are recommended for further reading.

References

[1] Ethik-Kommission. Automatisiertes und Vernetztes Fahren. (Ethics Commission. Automated and Interconnected Driving. In German.) Bundesministerium für Verkehr und digitale Infrastruktur. Bericht Juni 2017. URL: https://www.bmvi.de/SharedDocs/DE/Anlage/Presse/084-dobrindt-bericht-der-ethik-kommission.pdf?__blob=publicationFile.
An English version of the twenty ethical rules of this report is also available. URL: http://www.germany.info/Vertretung/usa/en/__pr/P__Wash/2017/06/21-AutonomousVehicles.html. Accessed 2017 June, 26.

[2] VDA Verband der Automobilindustrie: Automation — From Driver Assistance Systems to Automated Driving. September 2015. URL: https://www.vda.de/dam/vda/publications/2015/automation.pdf. Accessed 2017 July, 14.

[3] SAE: Automated Driving. URL: https://www.sae.org/misc/pdfs/automated_ driving.pdf. Accessed: 2017 February, 27.

[4] Definition: Levels of AD URL: https://www.2025ad.com/latest/the-levels-of-automation/. Accessed 2017 July, 14.

[5] Bundesministerium der Justiz und für Verbraucherschutz: Straßenverkehrsgesetz. URL: https://www.gesetze-im-internet.de/stvg/BJNR0043 70909.html. Accessed 2017 June, 30.

[6] Autonomes Fahren – Technische, rechtliche und gesellschaftliche Aspekte. Markus Maurer, J. Christian Gerdes, Barbara Lenz, Hermann Winner (Hrsg.) (Autonomous Driving—Technical, legal, and social aspects. Mostly in german. Chapters 4 and 5 on ethics in English). Springer-Verlag, Berrlin Heidelberg 2015.

[7] Niemann, H.: Klassifikation von Mustern. 2. überarbeitete Auflage, 2003. URL: http://www5.informatik.uni-erlangen.de/fileadmin/Persons/Nie mannHeinrich/klassifikation-von-mustern/m00links.html. Accessed 2017 June, 6.

[8] Gallus, Giovanni Battista: The laws of robotics and autonomous vehicles may be much more than three, but don't panic... yet. Keynote at the 22nd International Conference on Reliable Software Technologies—Ada-Europe 2017, 12-16 June 2017, Vienna, Austria.

[9] Halang, W. A., Konakovsky, R. M.: Sicherheitsgerichtete Echtzeitsysteme. (Safety-oriented Real-Time Systems. In German) 2. Auflage, Springer, Berlin 2013.

[10] Section on "Safety-related and Real-Time Systems". In: Autonomous Systems 2016. Proceedings of the 8th GI Conference. Unger, Herwig; Halang, Wolfgang A. (Eds.): Fortschritt-Berichte. VDI, Reihe 10, Nr. 848. Düsseldorf, VDI Verlag 2016.

[11] Taylor, M.: Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians. October 7, 2016. URL: http://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/. Accessed July 10, 2017.

[12] Lenninger, R.: The end of human driving? Why Angela Merkel got it right. June 21, 2017. URL: https://www.2025ad.com/latest/angela-merkel-driverless-cars/. Accessed July 10, 2017.

[13] Gerdes, C. in: dpa: Wenn der Computer am Steuer sitzt. (When the computer is at the steering wheel. Video in English with German subtitles.) URL: https://www.msn.com/de-de/video/ansehen/dpa-story-wenn-der-computer-am-steuer-sitzt/vp-BBAJIdN. Accessed: 2017 June 30.

[14] Capurro, R.: Autonomous zombies are not an option. June 28, 2017. URL: https://www.2025ad.com/latest/rafael-capurro-driverless-cars/. Accessed July 10, 2017.

[15] Simon, H.: The ethics of autonomous driving: Quo vadis? May 16, 2017.

URL https://www.2025ad.com/latest/ethics-of-autonomous-driving/. Accessed July 10, 2017.

[16] Wallach, W.: Automated driving: Are we approaching a moral crossroads? January 22, 2016. URL: https://www.2025ad.com/latest/ethics-and-self-driving-cars/. Accessed July 10, 2017.

[17] Capurro, R., Degenhart, E.: Industry meets Philosophy. In driverless cars we trust? A debate (Part 1). Can robocars handle responsibility? A debate (Part 2). The hacking threat – how to make driverless cars safe: a debate (part 3). URL: https://www.2025ad.com/latest/degenhart-capurro-driverless-cars-debate/. Accessed July 10, 2017.

A Concept Supporting Resilient, Fault-tolerant and Decentralised Search

Mario Kubek and Herwig Unger

Chair of Communication Networks
FernUniversität in Hagen, Germany

Abstract: Decentralised search engines usually replace huge databases by distributed information connected via respective structures. These structures must be generated and maintained with a low overhead and shall come along with a high scalability and fault tolerance. The article will show that random walker are able to establish suitable hierarchic, *tree-like* structures. Due to their composition mechanism the exhibit self-maintaining and -healing properties as well as a high adaptivity to changing size and needs of the stored information.

1 Introduction and Motivation

In previous works, the fundamental functionality of centralised search engines has be criticised. It was figured out that nowadays the establishment of a more or less indexed copy of the web is no more a forward-looking concept [1].

Therefore, the main concept for a decentralised search engine has been worked out. In particular, centroids where defined and used to characterise and compare any documents on the basis of a single, representing term [5]. The properties of these centroids have been considered in detail by [7]. Furthermore, the centroid concept has been used to derive a hierarchical clustering method [6].

The concept is based on a peer-to-peer system similar to [4] or [3] which is built as an extension of existing web-servers [1]. Hereby, the bootstrap problem is solved through the use of the existing hyperlinks in the webpages itself (Fig. 1).

Nevertheless, the immediate application of the hierarchical clustering algorithm require the election of an initial first node, since later merging activities maybe difficult. This problem may be overcome by a button-up construction of the hierarchy, which will be described below in a new manner, since its origins goes back to a contribution of one of the authors from 2004 [2]. Using methods of self-organisation and self-healing, the developed approach may even significantly

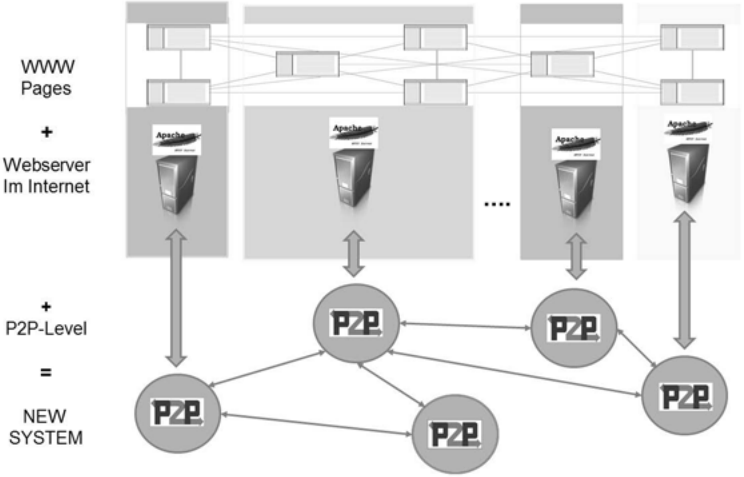


Fig. 1: Amending a Web-server into a peer-to-peer system

contribute to the fault tolerance of the solution caused by the weak connectivity in the so far used classical tree architecture.

2 Handling of Random Walker

Now, the peer-to-peer (P2P) system is considered, which is built by web-server extension as described in [1]. Every peer P with the IP address $IP(P)$ of this system has its own neighbourhood warehouse $N(P)$ containing initially the neighbours derived from the links of pages hosted by the web-server and later updated by the known, standard P2P ping-pong protocol.

For the intended hierarchy building procedures, random walkers shall be used. The needed information structure of each walker

$$RW = (IP[], D[], W[], c_{max}, c, z, ptr)$$

consists of

- An array $[IP_1, IP_2, .., IP_{c_{max}}]$ of positions to store IP addresses of peers belonging to a local cluster on a given level; whereby IP_1 always denotes and is initialised with the IP-address $IP(P)$ of the owning peer, i.e. the peer which has generated the random walker.

- A second array $[D_1, D_2, .., D_{c_{max}}]$ contains the respective, remaining path information (locator) to identify the document on the corresponding peer (needed if a peer may offer more than one document). Position D_1 is again reserved for the locator of the document of the owner/generator of the random walker.

- A third array $[W_1, W_2, .., W_{c_{max}}]$ containing an ASCII content description of the document represented by the locator in the corresponding positions of $IP[] : D[]$ (and later documents or sub-tree) managed by this random walker [1]. In our cases, the centroid term of the document will be used for this purpose.

- c_{max} the overall number of positions in the address array field which may be either a global constant or determined depending on the network conditions for each walker in an adaptive manner (e.g. depending on circulation times).

- c a counter indicating the number of filled positions in the IP array (initially assigned to 1). Note, that $c \leq c_{max}$ must be always fulfilled.

- z the number of additional, randomly to be determined positions.

- ptr the index of the current position of the walker with $1 \geq ptr \geq (c_{max} + z)$ and initialised to 1.

In the beginning, one random walker is generated for each document available from the respective web-server by the corresponding peer. Every document will be made identifiable beside $IP(P)$ by a local document locator D_h. An empty set $X(D_h)$ is stored, which later will contain IP addresses of peers with similar contents, i.e. which will built a completely connected sub-cluster.

Furthermore, each peer belonging to a web-server executes the following algorithm to process the generated and/or arriving random walkers.

1. $LOOP_1$**:**
   Receive a new random walker $R_n = recv()$.

2. Determine randomly the earliest departure time $\tau(R_n)$ for the random walker.

[1] Note, the random walker maybe more light-weighted, if this information is not a part of the random walker but retrieved from the peer of origin

3. Use a set C to keep all pairs of random walkers to computer and let $C = \emptyset$ at the start time of the algorithm

4. Let R be the set of random walkers, recently visiting the considered peer. Initialise $R := R_n$.

5. $LOOP_2$**:**
   Check $\forall i := 1..|R|$:
   IF $\forall j := 1..|R| \wedge (i \neq j)$

   - $(R_i, R_j) \notin C$ and
   - $(R_j, R_i) \notin C$ and
   - the recent time $t > \tau(R_i)$

   THEN

   - Set $R := R \setminus \{R_i\}$.
   - If $(ptr \leq c)$ then set $X(D_{ptr}) = \{IP_1, IP_2, .., IP_c\}$.
   - Increase $ptr := ((ptr + 1) mod(c_{max} + z)) + 1$.
   - Send out R_i to IP_{ptr} by $send(IP_{ptr}, R_i)$, if $ptr \leq c$ and otherwise to a randomly chosen successor out of the set of location in the neighbourhood set of the peer $IP_{next} \in N$ by $send(IP_{next}, R_i)$, if $c < prt \leq (c_{max} + z)$.

6. If another random walker R_s is received

   a) Determine randomly the earliest departure time $\tau(R_s)$ for this random walker.

   b) Set $C := C \cup (\{R_s\} \times R)$.

   c) Update $R := R \cup \{R_s\}$.

7. If $C \neq \emptyset$

   a) Take any $(R_i, R_j) \in C$.

   b) $Compute(R_i, R_j)$.

   c) Set $C := C \setminus \{(R_i, R_j)\}$

8. If $R = \emptyset$ GoTo $LOOP_1$ otherwise GoTo $LOOP_2$.

In order to ensure fault tolerance, any document D_i which is not checked by its random walker within a given timeout period T_{out} may take activities to synchronise a new random walker using a standard election mechanism with all nodes within $X(D_i)$.

3 Computation of Random Walker Data

This section describes the needed computation $Compute(R_i, R_j)$ for any pair of random walkers, which are meeting on any node in the P2P system. To distinguish both data areas, they are denoted by a leading first index in the formulas.

For the computation, two cases maybe distinguished.

1. $c_i + c_j \leq c_{max}$
   i.e. the two random walkers may be merged into one single. This case mostly appear at the start of the system or if a set of new documents and/or peers is added.

   The following updates are made for merging:

   a) $\forall k = (c_i + 1)..(c_i + c_j) : IP_{i,k} = IP_{j,k-c_j}$

   b) $\forall k = (c_i + 1)..(c_i + c_j) : D_{i,k} = D_{j,k-c_j}$

   c) $c_i = c_i + c_j$

   d) $c_{max,i}$, ptr_i and z_i remain unchanged

   e) The random walker R_j is cancelled by

      - $\forall(j := 1..|R|) \wedge (i \neq j)$
        let $C := C \setminus (R_i, R_j)$ and
        let $C := C \setminus (R_j, R_i)$.
      - $R := R \setminus \{R_j\}$

2. $c_i + c_j > c_{max}$
   i.e. the two random walkers met cannot be merged but the attached document links shall be sorted such that a maximum similarity is reached. Therefore the documents addressed via the set of URL's $(IP_n, m : D_n, m)$

$$U = \bigcup_{m=1}^{c_i} \{(IP_{i,m} : D_{i,m})\} \cup \bigcup_{m=1}^{c_j} \{(IP_{j,m} : D_{j,m})\}$$

are considered. Any clustering method or dichotomy building algorithm using document centroids in $W[]_i$ and $W[]_j$ as described in [6] shall be used to generate two subsets U_i and U_j from U with $|U_i| < c_{max,i}$ and $|U_j| < c_{max,j}$. The co-occurrence graph used for this purpose can be either obtained from the existing sub-cluster i or j or be a (temporary) combination of both.
The following updates are made for merging R_i and R_j using U_i and U_j:

a) $\forall k = 1..|U_i| : IP_{i,k} = IP_{i,k}(U_i)$

b) $\forall k = 1..|U_j| : IP_{j,k} = IP_{j,k}(U_j)$

c) $\forall k = 1..|U_i| : D_{i,k} = D_{i,k}(U_i)$

d) $\forall k = 1..|U_j| : D_{j,k} = D_{j,k}(U_j)$

e) $c_{max,i}$ and $c_{max,j}$ remain unchanged.

f) Set $c_i = |U_i|$ and $c_j = |U_j|$

g) z_i and z_j remain unchanged.

h) $ptr_i = c_i + z_i$ and $ptr_j = c_j + z_j$ to ensure that the updates are made in the fastest manner starting at the first node of the cycle in the next step.

In both cases the notification of all participating nodes on the made changes including an update of the local clusters is made within the next circulation of the random walker. In addition, the co-occurrence graph of all nodes newly merged to a random walker are merged and the respective resulting centroid term of all documents is determined under management of the peer owning the random walker (i.e. which is on $IP[1]$). Since this procedure as well as the future use of the owner (first peer in the sequence) of a random walker requires some computational performance of the respective machine, changes in the order of $IP[]$, $D[]$ as well as $W[]$ might be indicated and useful.

4 Building the Hierarchy

4.1 Structural Hierarchies

The methodology introduced in the before sections generate completely connected clusters of most similar documents, which are still isolated substructures.

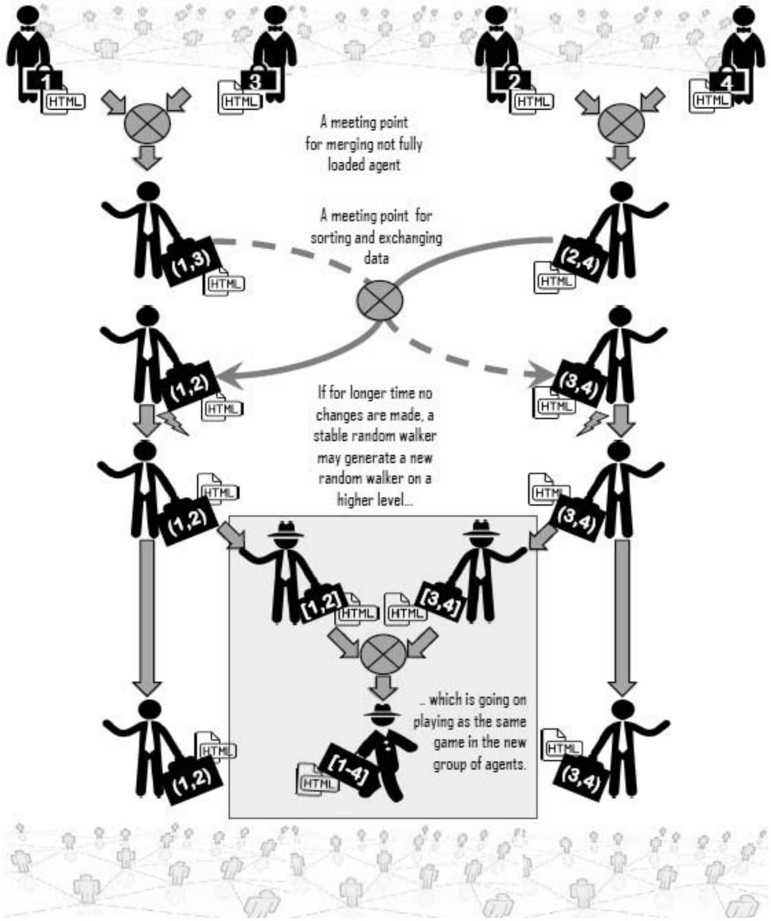

1
HTML
3
HTML
2
HTML
4
HTML
A meeting point for merging not fully loaded agent
(1,3)
HTML
A meeting point for sorting and exchanging data
(2,4)
HTML
(1,2)
HTML
(3,4)
HTML
If for longer time no changes are made, a stable random walker may generate a new random walker on a higher level...
(1,2)
HTML
(3,4)
HTML
[1,2]
HTML
HTML
[3,4]
... which is going on playing as the same game in the new group of agents.
(1,2)
HTML
[1-4]
HTML
HTML
(3,4)

Fig. 2: The agent's activities to build hierarchical, tree like structures

Now they shall be connected into a tree-like, hierarchical structure of completely connected sub-graphs.

Therefore two additions must be made to the definition of random walkers

$$RW = (IP[], D[], W[], c_{max}, c, z, ptr)$$

and changing it into

$$RW = (IP[], D[], W[], c_{max}, c, z, ptr, lev)$$

whereby lev, represents the level of the tree, on which the respective random walker is acting. Note that the leafs will have –differing to usual enumerations– level 1.

For the processing, the following updates are necessary:

1. In the beginning at the starting point of the above described algorithm, $lev = 1$ representing the lowest, the document level.

2. In W the centroid term of the associated document is contained, when random walkers are merged, it is replaced by the centroid term of all documents represented by the random walker.

3. The owner of the random walker, i.e. usually the peer represented in the first position of $IP[]$ observes all changes in the random walker.
   If for fixed, longer time interval Δ

   - no changes are observed in the $IP[]-$ and $D[]-$ area of the random walker and

   - $1 < c \leq c_{max}$,

   the $IP[1]$ peer is allowed *to launch another, new random walker*.

4. By doing so, it may happen that on any level a single random walker may not be connected to the community, since all existing other random walker have filled exactly all c_{max} positions.
   To avoid this case, position $c_{max} + 1$ is used as temporary (emergency) position. If $c = c_{max}$ it may merge with a random walker with $c = 1$ but must release this position at the next possible solution regardless any content aspects, i.e. as soon as either a random walker with $c < c_{max}$ is met or another random walker with $c_{max} + 1$ filled positions is found (in this case a new random walker at the same level is created).

5. This *new random walker* follow exactly the rules for the initial settings, as described above for the level 1-random walker, with the following updates:

   - The respective level information is the the level information of the generating walker increased by 1.

   - W contains the centroid term which is calculated as centroid term of all participating (represented) documents (level 2) or random walkers (in the higher levels), which are usually stored on the generating peers.

6. Also, the behaviour of the new random walker follows exactly the above described rules, however a computation, merging and sorting will only take place for random walkers having the same level lev.

7. For fault tolerance reasons, the $IP[]-$ and $D[]-$area may contain instead of a single information a sub-array, containing a small number (i.e. two or three) of information from other nodes of the represented sub-cluster. In such a manner, any break down or failure of the first peer in the list maybe tolerated by using a replacement peer.

8. The described hierarchy building mechanism is successively repeated for all higher levels $2, 3, ...$ and stops automatically, if the conditions formulated in 3 cannot be fulfilled any more.

Fig. 2 shows stepwise the run and the activities of the agents in the systems.

4.2 Hierarchies of Co-occurrence Graphs

Any decision making bases on co-occurrence graphs stored locally. On the lowest level 1 the respective co-occurrence graphs are built from one document. As soon as the positions $1..c_{max}$ of each random walker are filled, temporary and bigger co-occurrence graphs may be built and used for decision making but must be eventually re-organised, since every document may only represent its set of co-occurrences in one co-occurrence graph $R_{c,lev}$ on each level lev.

With the generation of a random walker for $lev + 1$, the co-occurrence graph of the respective represented region on level lev shall be built and be stable. At least with the complete compositions of the random walker on level $lev + 1$ also the next level co-occurrence graph $Rc, lev + 1$ shall be available and will be used in the same manner to assembly the next hierarchy level $lev + 2$. In addition, $Rc, lev + 1$ will be given down to all random walker until level 1.

Of course, the respective (but seldom necessary) updates may cause changes in the evaluation of the similarities of documents (sets of documents in the different regions) combined in a random walker. Since the remaining z randomly chosen positions in each random walker will allow meetings on any peer at any time, these discrepancies will be automatically recognised and the structure will be adapted. Exactly the same process will result in an automatic inclusion of newly appearing documents and their random walker.

Note that any change in a random walker require a new calculation of the co-occurrence graph on the respective level with the corresponding re-calculations on its upper and lower levels. Since this process is relatively seldom needed (last but not least to the stability of co-occurrence graphs as found in [7]), the overview shall be acceptable. Due to its size, all Rc, lev shall never be a part of the random walker but solely kept in a graph data base on the represented nodes, regions or sub-clusters. The complete connection of those structures (as described above) may make those updates more simple.

5 Implementation

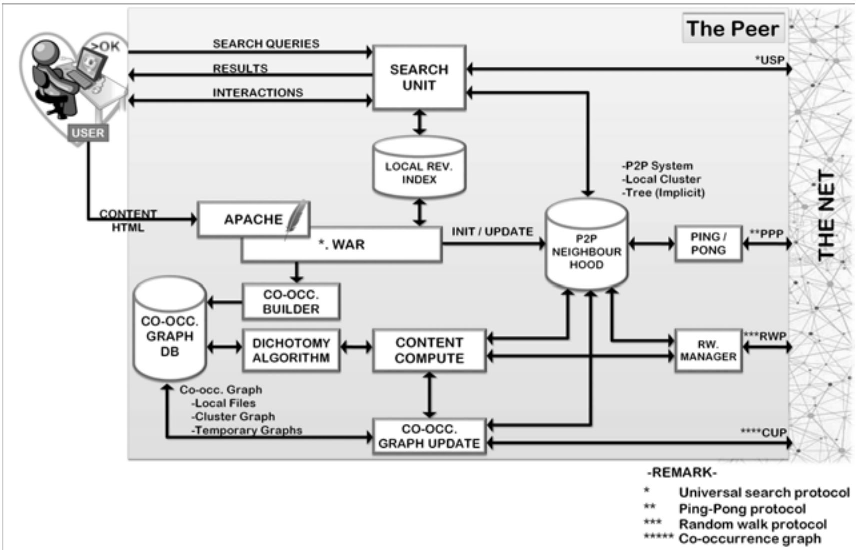


Fig. 3: The peer structure

Fig. 3 shows the possible extension of a (classic) peer as shown in Fig. 1. The blocks in the upper half of the scheme build up the functionality of a usual P2P search system with a storage of other peers and neighbourhoods already in a special built data base.

In a first step, the *Search Unit* will be extended to answer, forward and handle centroid based search requests made by users using a single message, nonbroadcasting, universal search protocol (USP), which guide the search request based on distance measures to the target node(s).

All other random walker management as described in section 2 is done by the added *RW-Management*, for which a special random walker protocol (RWP) is created. The processing of random walker data *Content Compute* unit, which may access the co-occurrence graph data base (build with NEO4J), which contain the co-occurrence graph data of

1. each HTML file offered from the local WWW server;
2. the combined, local cluster in which the node is a member;
3. temporary operations of the random walker to build or update the hierarchy structures.

In order to support these operation, a co-occurrence graph builder is implemented in a separate unit. The needed clustering is supported by the *Dichotomy Algorithm* unit. Last but not least, the exchange of co-occurrence information is supported by a co-occurrence update protocol (CUP) controlled by a respective separate unit.

6 Conclusion

The algorithmic fundamentals and implementation details of a fully decentralised P2P-search engine have been presented. It is working on the classification of documents as well as search requests by so called centroids, i.e. single terms. Agent based, a locally hierarchic, tree-like structure is built, which allows a routing of search requests without broadcasts. In addition, it is scalable and fault-tolerant as well.

References

[1] Eberhardt, R., Kubek, M., Unger, H.: *Why Google Isn't the Future. Really Not.* In: H. Unger and W. Halang: Proceedings of the Conference on Autonomous Systems 2015, Fortschritt-Berichte VDI, Series 10: Informatik/Kommunikation, VDI, Düsseldorf, (2015)

[2] Unger, H., Wulff, M.: *Cluster-building in P2P-Community Networks.* In: Journal on Parallel and Distributed Computing Systems and Networks, Vol. 5(4), pp. 172–177, (2002)

[3] Christen, M. et al.: *YaCy: Dezentrale Websuche*, Online Documentation, on `http://yacy.de/de/Philosophie.html`, downloaded on April 11th 2017, (2017)

[4] n.n.: *FAROO: Distributed Search.* White Paper, Faroo Limited, online version via `http://www.faroo.com/hp/p2p/whitepaper.html`, downloaded on April 11th 2017, (2017)

[5] Kubek, M., Unger, H.: *Centroid Terms as Text Representatives.*, In: DocEng 2016, Proceedings of the 2016 ACM Symposium on Document Engineering, Vienna, Austria, ACM, (2016)

[6] Kubek, M., Unger, H.: *Towards a Librarian of the Web.* In: Proceedings of the 2nd International Conference on Communication and Information Processing (ICCIP 2016), ACM, Singapore, (2016)

[7] Kubek, M., Böhme, T., Unger, H.: *Empiric Experiments with Text Representing Centroids.* In: Proceedings of the 6th International Conference on Software and Information Engineering (ICSIE 2017), Singapore, (2017)

Performance Evaluation of an Immune Genetic Algorithm

Pongsarun Boonyopakorn and Phayung Meesad

Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand

Abstract: This paper demonstrates a hybrid between two optimization methods which are the Artificial Immune System (AIS) and Genetic Algorithm (GA). The novel algorithm called the immune genetic algorithm (IGA), provides improvement to the results that enable GA and AIS to work separately which is the main objective of this hybrid. Negative selection which is one of the techniques in the AIS, was employed to determine the input variables (populations) of the system. In order to illustrate the effectiveness of the IGA, the comparison with a steady-state GA, AIS, and PSO were also investigated. The testing of the performance was conducted by mathematical testing, problems were divided into single and multiple objectives. The five single objectives were then used to test the modified algorithm, the results showed that IGA performed better than all of the other methods. The DTLZ multi-objective testing functions were then used. The result also illustrated that the modified approach still had the best performance.

1 Introduction

Optimization search research is an operation that refers to the procedure of finding the best solution to objective functions. In general, optimization searching considers the solution into local and global searches. It can be group into two categories which are single and population based. In recent years, the most popular technique used to solve the solutions was the Genetic Algorithm (GA) which is a population based search system. The GA is a metaheuristic method pioneered by Darwin's [1] which is based on the principle of natural genetics and natural biological selection. The process of GA's is iteratively the initial population of candidate solutions until the criteria have been met. The GA operation begin with the initial the number of population which are related to the solution that needed to be solve. The selection is a method that uses selection of parents from each individual in each iteration for regeneration. The crossover and mutation are operators used for regeneration to the next iterator. Finally, the evaluation of candidate solution will calculate then terminate the solutions once reached or iteration terminated.

The GA has been successfully applied to many research areas such as optimized tools, engineering, science, and management. In recent years, GAs have been proposed as hybrids by combining to various approaches such as PSO, Ant Colony, and Artificial Immune System which are effective for local and global searches aimed at improving the solution quality. Various problems have been solved by the hybrid GA which includes finding optimal traffic networks, job scheduling, stock markets, and data mining. Tarek A. [2], proposed a hybridization between Ant-based Algorithm and Genetic Algorithm. In their research, an Ant Colony was used to monitor the behavior of a genetic-local hybrid algorithm and dynamically adjusted its control parameters to optimize the exploitation exploration balance according to the fitness landscape. Jyoti [3], presented a hybrid combining the Particle Swarm Optimization algorithm (PSO) based on testing five functions. The idea behind the hybrid algorithm is that the total iterations have to be distributed between the genetic algorithm and particle swarm optimization algorithm. The proposed hybrid algorithm is proven to be more efficient than GA and PSO. In [4] Zhao, proposed a hybrid genetic algorithm for Bayesian network optimization. Their work used the Simulated Annealing technology to select children and used self-adaptive probabilities of crossover and mutation to conduct the local search. Finally, the Hill-climbing algorithm was employed to optimize the results. In [5], Wu and Lu studied the effects of hybrid optimization strategies by incorporating the metropolis acceptance criterion of Simulated Annealing (SA) into the crossover operator of GA. The algorithm was used to simultaneously optimize the input feature subset selection, the type of kernel function and the kernel parameter setting of SVR, namely GASA-SVR. In summary of the above, the study of hybrid Genetic Algorithms has yielded several successful approaches.

The Artificial Immune System (AIS), has been studied deeply in recent years which is a class of biologically inspired computation paradigm [6]. AIS approaches are used in various optimization applications and most of them show better efficiency in comparison with other population based algorithms. Various AIS models such as clonal selection, immune networks, and negative selection are also used in several applications such as optimization, clustering, pattern recognition and anomaly detection. In general, GA and AIS have been adopted as optimizers in the binary base which is categorized as NP-hard. Zhu [7], investigated two theories of AIS which are clonal selection and immune network theory, and integrated them with PSO to solve the job scheduling problem. In his research, the clonal selection theory is used to set up the framework which contains the processes of selection, cloning, hyper mutation and receptor editing, while the immune network theory is applied to increase the diversity of the potential solution repertoire. Barani [8], proposed an approach based on the genetic algorithm (GA) and artificial immune system (AIS), called GAAIS, for dynamic intrusion detection in AODV-based MANETs. His approach was able to adapt itself to network topology changes using two updating methods: partial and total. Each normal feature vector extracted from network traffic was represented by a hypersphere with fix radius. Ali et al. [9], improved the results of performance in the hybrid AIS and GA. The hybrid included two processes; firstly, AIS enables it to develop local searching ability and efficiency although the convergence rate for AIS is preferably not precise compared to the GA. Secondly, a Genetic Algorithm is typically initializing population randomly. The last generation of AIS will be the input to the next process of the hybrid which is the GA in this hybrid AIS-GA. A hybrid can ensure that a GA enters the stage of standard solutions more rapidly and accurately compared to GA initialized population at random.

As mention above, the hybrid AIS and GA have been applied to difference optimization application areas in recent years. The object of this paper is to describe the modified Genetic algorithm (GA) which is a combination of an Artificial Immune System (AIS) to form an Immune Genetic Algorithm (IGA) to reduce the search space and achieve efficient searches. Performances of the IGA and two other techniques will be compared. This paper is divided as follows: Section 2 presents the research method of the evolutionary algorithm. Section 3 presents the proposed of the algorithm and testing functions. Section 4 covers the results and analysis. Finally, the conclusion will be presented in session 5.

2 Research Method

This section discusses and analyzes the aim of the hybrid immune genetic algorithm concepts to utilize the locally characteristic information to seek out the ways and means of discovering the optimal solution when dealing with difficult problems. One must first generate a random detector, and then the initial population. Next perform selection, crossover, and mutation upon the population for a number of generations, until termination criterion is met.

2.1 Negative Selection

Negative selection inspired from the T cell maturation process has been developed for self-nonself detection in computer systems. In this technique, the first information is represented in a suitable form such as string form, real valued vector form, and hybrid form are considered as self-data. Then additional data are created in the same form as the self-data, in such a way that any of the newly created data does not match the self-data. The matching is done according to a matching rule which is selected depending on suitability. These newly created data which are used to distinguish between self-data and nonself-data are called detectors. If any of the detectors matches the data, then that data is considered nonself-data. Whereas, if no detector matches the data then that data is considered self-data. The detectors are created in such a way that they do not match any of the self-data. In negative selection, the T cell is presented to the self-body cells. If the T cell recognizes any of the self-body cells, then the cell is rejected. Remaining T cells are considered matured T cells and are used for the self-nonself detection [10].

```
Input: SelfData
Output: Repertoire
Repertoire $\leftarrow \Phi$
While ($\neg$StopCondition())
    Detectors $\leftarrow$ GenerateRandomDetectors()
    For ($Detector_i \in Repertoire$)
        If (NotMatches($Detector_i$, $SelfData$))
            $Repertoire \leftarrow Detector_i$
        End
    End
End
Return ($Repertoire$)
```

Fig. 1: Pseudocode for detector generation

Figure 1 describes the major steps in such an algorithm. In the generation stage, the detectors are generated by a few random process and censored by trying to match self samples. Those candidates that match are eliminated and the rest are kept as detectors. In the detection stage, the collection of detectors (or detector set) are used to check whether an incoming data instance is self or nonself. If it matches any detector (referred to figure 2), it is claimed as nonself or an anomaly. This description is limited to a few extents, but conveys the essential idea.

```
Input: InputSamples, Repertoire
For ($Inputi_{class}$ □ InputSamples)
    $Inputi_{class} \leftarrow$ "non-self"
   For ($Detector_i$ □ Repertoire)
      If (Matches($Input_i$, $Detector_i$))
         $Inputi_{class} \leftarrow$ "self"
         Break
      End
   End
End
```

Fig. 2: Pseudocode for detector application

2.2 Matching Rules

Matching rule is an important part in detector generation. There are different matching rules such as Hamming distance, Binary distance, Edit distance, and Value difference metric to match strings. In this paper, focus is on the R-Contiguous Bits (RCB) matching rule and R-Chunk matching rule [11]. The RCB matching rule is defined as follows: If x and y is equal-length strings defined over a finite alphabet, match (x, y) is true if x and y agree in at least r contiguous locations. As in the RCB matching rule, a detector is specified by a binary string c and parameter r.

2.3 Detector Generation

The detector generation technique can be divided into two parts. i) The value of the length of the chunk is taken from the user. Let the chunk length be x then from the first bit of a self-string x, none of the continuous bits are taken to form a chunk. Then to form the second bit x, none of the continuous bits are taken from another chunk and this goes on as long as x has none of the continuous bits taken to form a chunk. So, if the length of self-string is y, then $y - x + 1$ none of the chunks are formed from each self-string. ii) Each self-chunk set is taken one by one to the detector sets separately. As chunks are already created from self-strings two strings are considered the same only if all the bits of the two strings exactly match each other. Next, detectors are to be created such that newly created detectors do not match previously generated detectors or the self-chunk strings even-though the randomness of the detector generation process are maintained.

2.4 Chromosome Representation

The chromosome representation depends on the nature of the problem variables. The value of a bit string can be an integer number or binary number. For example, the representation choice of timetabling schedules for a few objects. A possible number of 15-bit strings can be used to represent a possible solution to a problem. In this case bits or subsets of bits might represent a choice of a few features: subject, section, instructor, time, and room.

110011100110011

Fig. 3: Chromosome Representation

where bit 1–3 represents subject, 4–6 represents course section, 7–9 represents instructor or professor, 10–12 represents times, and 13–15 represents room.

2.5 Initial Population

The chromosome's fitness value is assessed during the initial population process. Each individual contains its own fitness value. One possible way to assign a fitness value to individuals is by the following formula

$$fitness_i = \sum_{i=0}^{n} \alpha_i \tag{1}$$

where α is an element of bit-string and i is a number of bit-string that's contained in each individual.

2.6 Selection

The selection process chooses the next generation of the best individual. It stochastically allocates a higher number of copies in the following generation to highly fitting strings in the present generation. Four common methods for selection are Roulette Wheel selection, Stochastic Universal sampling, Normalized Geometric selection, and Tournament selection. For example, figure 4 shows Tournament selection which provides a chance to all individuals to be selected and thus it preserves diversity, although keeping diversity may degrade the convergence speed. In tournament selection, n individuals are selected randomly from the larger population, and the selected individuals compete against each other. The individual with the highest fitness wins and will be included as one of the next generation population. The number of individuals competing in each tournament is referred to as tournament size, commonly set to 2 (also called binary tournament).

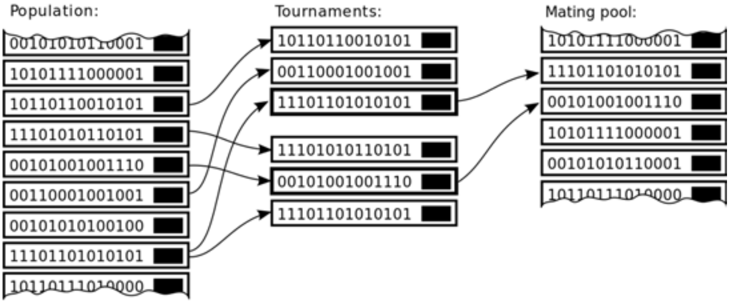


Fig. 4: Illustration of Tournament Selection in Size of 2

2.7 Crossover

The crossover process produces better chromosomes, two of the strongest are picked to produce a new chromosome of offspring. Figure 5 shows the example of single point crossover. Three types of crossover are applied in this process including Single point crossover, Double point crossover, and Uniform crossover.

2.8 Mutation

Mutation is the occasional random alteration of a value of a string position. The purpose of mutation in GAs are to preserve and introduce diversity. For dif-

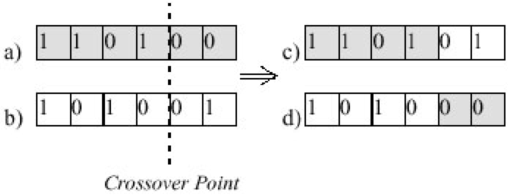


Fig. 5: Illustration of Single Point Crossover

ferent genome types, different mutation types are Bit string mutation, Flip Bit, Boundary, Uniform, and Gaussian. A randomly selected element of the string is altered or mutated when a string is chosen for mutation. Normally, mutation ranges around 0.1% – 0.2%. Figure 6 shows an example of bit string mutation.

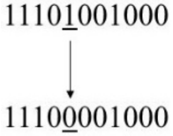


Fig. 6: Illustration of Bit String Mutation

3 Proposed Algorithm and Testing Functions

The proposed algorithm begins with initialize detector D, each of which fails to be a random value. The next step is to calculate the fitness of each cell in the population and rank them. In this case, the best candidate will be chosen to be detector D. Next, initialize a population P of gene, each set will have a random value then perform negative selection in any P which matches D. Calculate fitness of each chromosome in P and rank them and perform crossover and mutation. Loop if termination condition is not met then stop. The pseudocode for the Immune Genetic Algorithm is shown in figure 7.

3.1 Single Objective Test Functions

In order to compare and evaluate different algorithms, various benchmark functions with various properties have been suggested. Five single objective

```
d ← 0;
InitDetector[$D(d)$];    {Initializes the detector}
EvalDetector[$D(d)$];    {Evaluates the detector}
$t$ ← 0;
InitPopulation[$P(t)$];    {Initializes the population}
EvalPopulation[$P(t)$];    {Evaluates the population}
Matches[$P(t), D(d)$];    {Matches between the population and detector}
while (not terminatation) do
    $P'(t)$ ← Variation[$P(t)$];    {Creation of new solutions}
    EvalPopulation[$P(t)$];    {Evaluates the new solutions}
    $P(t+1)$ ← ApplyGeneticOperators[$P'(t) \cup Q$];    {Next generation pop.}
    $t \leftarrow t+1$;
end while
```

Fig. 7: Pseudocode for immune genetic algorithm

test functions are used in this paper to compare between GA, AIS, IGA, and PSO. The following are the test functions.

Ackley function

$$f(x) = -a\ \exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(cx_i)\right) + a + \exp(1) \qquad (2)$$

subject to $-35 \leq x_i \leq 35$.

The global minima is located at origin $x = (0,\ldots,0), f(x) = 0$ where $a = 20$, $b = 0.2$ and $c = 2\,\pi$.

Bohachevsky functions

$$f_1(x) = x_1^2 + 2\,x_2^2 - 0.3\ \cos(3\,\pi\,x_1) - 0.4\ \cos(4\,\pi\,x_2) + 0.7 \qquad (3)$$

subject to $-100 \leq x_i \leq 100$.
The global minimum is located at $x = f(0,0), f(x) = 0$.

$$f_2(x) = x_1^2 + 2\,x_2^2 - 0.3\ \cos(3\,\pi\,x_1)\ \cos(4\,\pi\,x_2) + 0.3$$

subject to $-100 \leq x_i \leq 100$.
The global minimum is located at $x = f(0,0), f(x) = 0$.

$$f_3(x) = x_1^2 + 2\,x_2^2 - 0.3\,\cos(3\,\pi\,x_1 + 4\,\pi\,x_2) + 0.3$$

subject to $-100 \leq x_i \leq 100$.
The global minimum is located at $x = f(0,0), f(x) = 0$.

Sphere function

$$f(x) = \sum_{i=1}^{d} x_i^2 \tag{4}$$

subject to $0 \leq x_i \leq 10$.
The global minima is located at $x = f(0,\ldots,0), f(x) = 0$.

Rastrigin function

$$f(x) = 10\,d + \sum_{i=1}^{d} \left[x_i^2 - 10\cos(2\,\pi\,x_i)\right] \tag{5}$$

Fifth function of De Jong

$$f(x) = \left(0.002 + \sum_{i=1}^{25} \frac{1}{i + (x_1 - a_{1i})^6 + (x_2 - a_{2I})^6}\right)^{-1} \tag{6}$$

where a_1 and $a_2 = -32$ to 32

3.2 DTLZ Multi-Objectives Test Functions

The DTLZ suite of benchmark problems, created by Deb et al. [12], is unlike the majority of multi-objective test problems in that the problems are scalable to any number of objectives. This is an important characteristic that has facilitated several recent investigations into what are commonly called ″many″ objective problems. DTLZ1 to DTLZ6 are scalable with respect to the number of distance parameters but have a fixed number of $M - 1$ position parameters, where M is the number of objectives. Note also that the objective functions of DTLZ have multiple global optima since terms such as $\cos(y_i\,\pi\,/\,2)$ can evaluate to zero, thereby allowing flexibility in the selection of other parameter values. Technically speaking, these objectives are non-separable, as attempting to optimize them one parameter at a time (in only one pass) will not identify all global optima. As this is a minor point, one can classify the objectives of DTLZ as being separable irrespective, as attempting to optimize them one parameter at a time will identify at least one global optima. Incidentally, there being multiple global optima is why many of the DTLZ problems are Pareto many-to-one.

DTZL1

$$\begin{gathered}
f_1 = (1+g)\,0.5 \prod_{i=1}^{M-1} y_i \\
f_{m=2:M-1} = (1+g)\,0.5 \left(\prod_{i=1}^{M-m} y_i \right) (1 - y_{M-m+1}) \\
f_M = (1+g)\,0.5\,(1 - y_1) \\
g = 100 \left[k + \sum_{i=1}^{k} \Big((z_i - 0.2)^2 - \cos(20\,\pi\,(z_i - 0.5)) \Big) \right]
\end{gathered} \tag{7}$$

DTZL2

$$\begin{gathered}
f_1 = (1+g) \prod_{i=1}^{M-1} \cos(y_i\,\pi/2) \\
f_{m=2:M-1} = (1+g) \left(\prod_{i=1}^{M=m} \cos(y_i\,\pi/2) \right) \sin(y_{M-m+1}\,\pi/2) \\
f_M = (1+g)\,\sin(y_1\,\pi/2) \\
g = \sum_{i=1}^{k} (z_i - 0.5)^2
\end{gathered} \tag{8}$$

DTZL3

$$\begin{gathered}
f_1 = (1+g) \prod_{i=1}^{M-1} \cos(y_i\,\pi/2) \\
f_{m=2:M-1} = (1+g) \left(\prod_{i=1}^{M=m} \cos(y_i\,\pi/2) \right) \sin(y_{M-m+1}\,\pi/2) \\
f_M = (1+g)\,\sin(y_1\,\pi/2) \\
g = 100 \left[k + \sum_{i=1}^{k} \Big((z_i - 0.2)^2 - \cos(20\,\pi\,(z_i - 0.5)) \Big) \right]
\end{gathered} \tag{9}$$

DTZL4

$$\begin{aligned}
f_1 &= (1+g)\,\prod_{i=1}^{M-1}\cos(y_i\alpha\,\pi/2)\\
f_{m=2:M-1} &= (1+g)\left(\prod_{i=1}^{M=m}\cos(y_i^{\alpha}\,\pi/2)\right)\,\sin(y_{M-m+1}\,\pi/2)\\
f_M &= (1+g)\,\sin(y_1^{\alpha}\,\pi/2)\\
g &= \sum_{i=1}^{k}(z_i-0.5)^2
\end{aligned}\tag{10}$$

where $\alpha > 0$

DTZL5

$$\begin{aligned}
f_1 &= (1+g)\,\prod_{i=1}^{M-1}\cos(\theta_1\,\pi/2)\\
f_{m=2:M-1} &= (1+g)\left(\prod_{i=1}^{M=m}\cos(\theta_1\,\pi/2)\right)\,\sin(\theta_2\,\pi/2)\\
f_M &= (1+g)\,\sin(\theta_1\,\pi/2)\\
g &= \sum_{i=1}^{k}(z_i-0.5)^2
\end{aligned}\tag{11}$$

where $\theta_i = \frac{\pi}{4(1+g(r))}\,(1+2\,g(r)\,x_i)$

DTZL6

$$\begin{aligned}
f_1 &= (1+g)\,\prod_{i=1}^{M-1}\cos(\theta_1\,\pi/2)\\
f_{m=2:M-1} &= (1+g)\left(\prod_{i=1}^{M=m}\cos(\theta_1\,\pi/2)\right)\,\sin(\theta_2\,\pi/2)\\
f_M &= (1+g)\,\sin(\theta_1\,\pi/2)\\
g &= \sum_{i=1}^{k} z_i^{0.1}
\end{aligned}\tag{12}$$

where $\theta_i = \frac{\pi}{4(1+g(r))}\,(1+2\,g(r)\,x_i)$

4 Results and Analysis

The evaluation of the developed technique was grounded on the operation simulation. The estimation of the running time and the minimum fitness values were composed of five mathematical test functions.

4.1 Experiment Setup

In order to test the effectiveness of IGA, AIS, and GA when solving timetabling problems, a comparison with the PSO algorithm was performed to investigate trends of performance. All coding was written in MATLAB, and the test case focused on the four above algorithms. All tests were executed on a 3.30 Ghz Intel core i5 processor with 16 GB of ram. The convergence graph for IGA, AIS, GA, and PSO below shows progress until a valid solution for each of the algorithms were discovered. The specifications of the problem are shown in Table 1.

Table 1: General parameters used by the algorithm

No	Operator	Quantity/Type
1	Number of individuals	1000
2	Crossover probability	0.9
3	Mutation probability	0.02
4	Number of generations	1000
5	Selection mechanism	Tournament selection
6	Crossover type	Two point crossover

Table 2: Comparison of run time

Generation	GA (sec)	AIS (sec)	IGA (sec)	PSO (sec)
100	11	10	**8**	9
200	23	22	**20**	23
400	46	39	**37**	39
800	95	80	**78**	82
1000	121	92	**85**	93

In the PSO calculation, the practice set size of population for GA, AIS, and IGA with $c_1 = 2, c_2 = 2$, was $w = 1/(2x \log 2)$. In the GA and IGA calculation, mutation rate = 0.02 and crossover = 0.7. By observing the simulation results and the graphs, it is inferred that, for most of the populations, there was no great difference between the execution times and fitness values for the GA, AIS, IGA and PSO. Until the population reached 500, the IGA had an increase of time whereas PSO remained linear. However, the GA and IGA fitness values rose more than PSO when the population increased. In AIS, it still has less fitness values at the beginning but then rose higher when the population increased. This might stem from the fact that the parameters of GA, AIS and IGA were different from PSO such as velocity, global and local search space, and so on. Over all, when comparing between IGA and GA, it is cleared to see that IGA had less fitness than GA but IGA consumed less time to reach the optimal solution than GA. From Table 2, it is clear to see that IGA greatly reduced the running time. Off course, the search space could be less and also the generation of run time more effective. The results illustrated that the highest fitness value occurred when crossover probability was in the range of 0.9.

4.2 Single Objective Mathematical Result

In order to compare and evaluate different algorithms, various benchmark functions with various properties were used. Five test functions were used in this research as follows, Ackley, Bohachevsky, Sphere, Rastrigin, and the fifth function of De Jong were compared between GA, AIS, IGA, and PSO. The following are the five functions as presented in section 3.

Figure 8 to 9 shows the results between the hybrid algorithms compared to other algorithms with various mathematical functions as presented in section 3. It is cleared to see that the hybrid algorithm performed better than other techniques. Surprisingly, the Ackley function performed closely to the hybrid algorithm. Figures 10 and 12 shows that PSO reached the optimal result very quickly because this algorithm works as a local search which makes a narrow space for the search of a solution, rather than other algorithms which work as a global search space.

4.3 Multi Objectives Mathematical Results

Experiments were conducted to compare the differences between the DTLZ1 to DTLZ6 problems. This is to show the advantage of a modified algorithm over the DTLZ suite when all problems use the same configuration as shown in

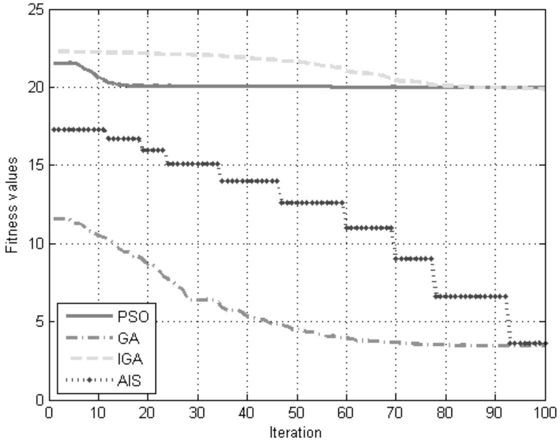


Fig. 8: The Comparison between fitness Ackley function

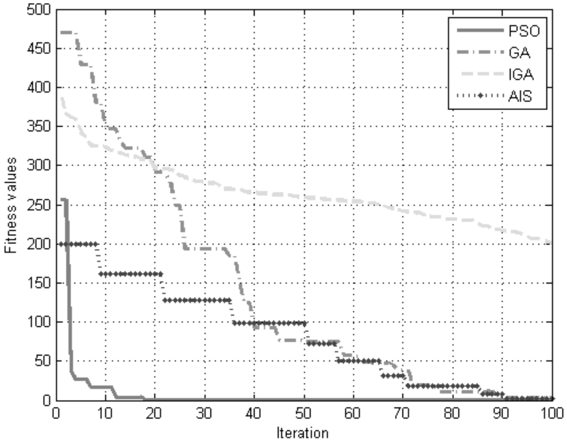


Fig. 9: The Comparison between fitness Bohachevsky functions

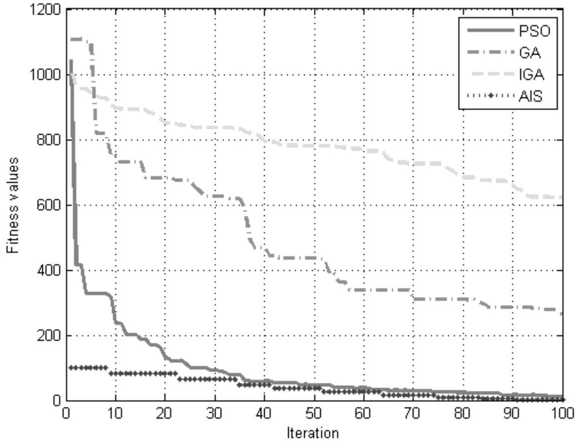


Fig. 10: The Comparison between fitness Sphere function

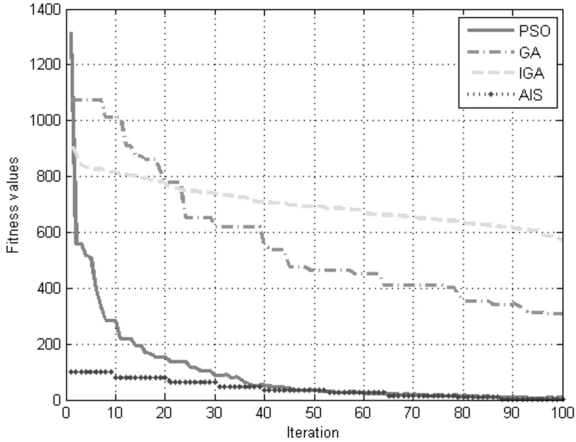


Fig. 11: The Comparison between fitness Rastrigin functions

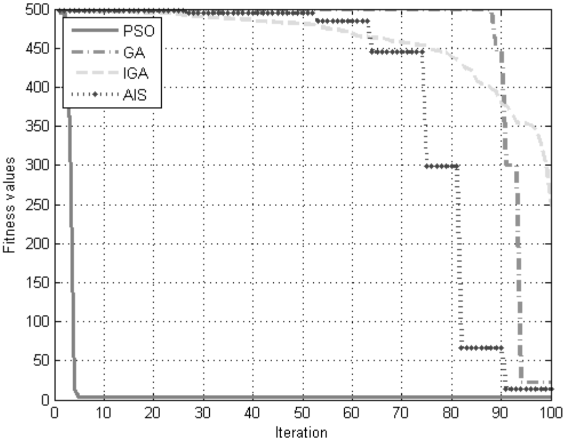


Fig. 12: The comparison between fitness fifth function of De Jong

table 1 which also compares their best and average cost obtained over the ten runs for each of the problems. The results are compared in Figures 13 to 17 at 20–100 generation times.

Results in Figure 13 shows that the standard PSO algorithm gives a better alternative to solve the problem of DTLZ1 when the number of iteration are small and quickly convergences a local optimal solution. Based on the results, this indicates that the modified algorithm IGA takes more computational fitness values for a feasible solution compared to other techniques. The AIS algorithm performed well but when the iteration reached 100, it decreased under IGA while GA was stable.

Both the IGA and AIS were able to produce feasible solutions to the DTLZ2 problem. Figure 14, focuses on the iterations of 100, and shows the best and average fitness values. Note that the fitness value is the summation from (8). Moving to the test of DTLZ3 shown in Figure 15, at the iterations of 100, average value of IGA and AIS were not different. These results support the claim that the IGA can greatly improve the performance of GA and PSO when solving problems.

In Figure 16, the statistical results of the four algorithm applied to the DTLZ4 problem are shown. Although they all obtain the optimal solution, it is apparent

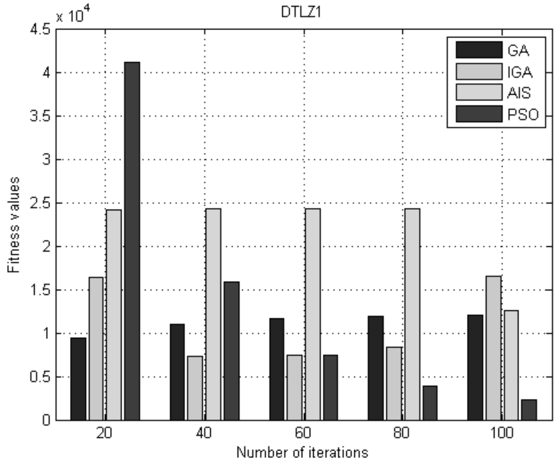


Fig. 13: The DTLZ1 testing function

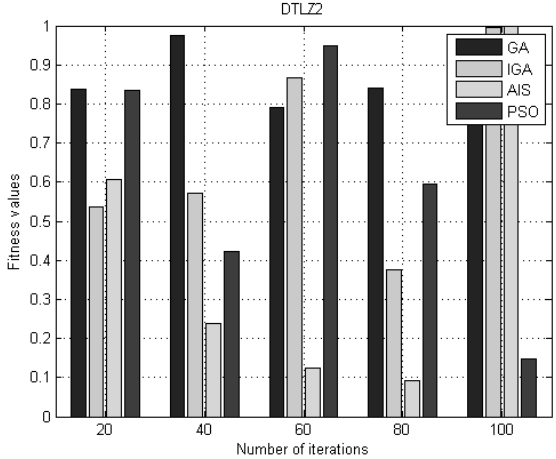


Fig. 14: The DTLZ2 testing function

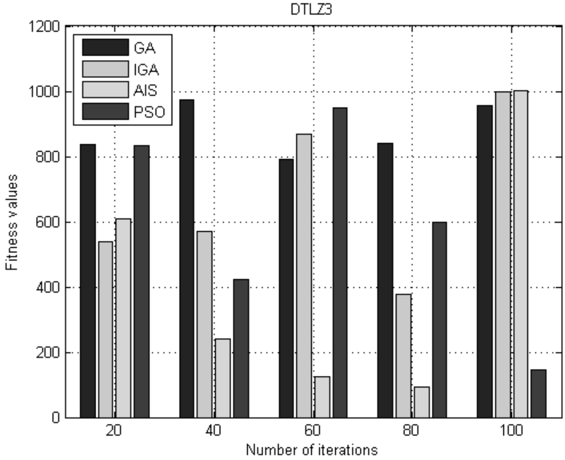


Fig. 15: The DTLZ3 testing function

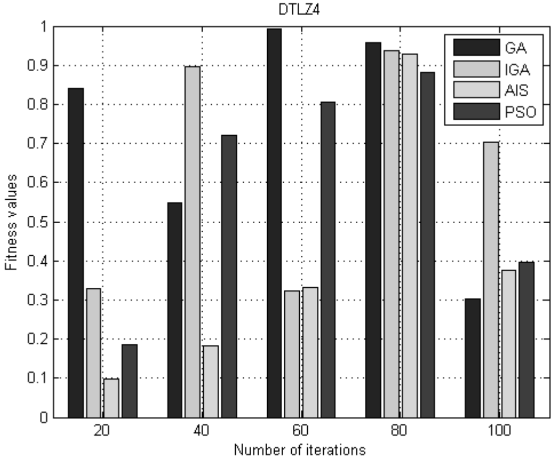


Fig. 16: The DTLZ4 testing function

that IGA is superior. However, in this problem of 20 times, it behaves worse and fails to ensure the performance. Through the further experiments, it was found that IGA could achieve better performance after several runs of time compared with others. Apparently, in the aspect of average fitness, IGA is better than other standard algorithms.

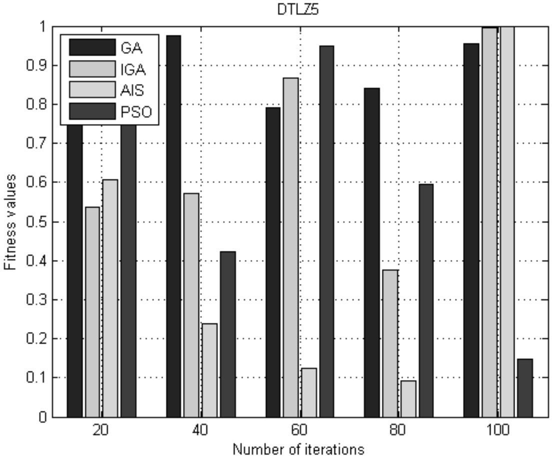


Fig. 17: The DTLZ5 testing function

Next investigation shows the statistical information results of the DTLZ5 problem. In Figure 17, it can be seen that the IGA improves after 80 times and the average results are not much different from the AIS results due the number of chromosomes for the initial solution.

From Figure 18 it is evident that the IGA performs better than the PSO and GA, but not AIS at 100 times which means that both of the algorithms failed to find a feasible solution at the beginning of the corresponding DTLZ6 problem.

4.4 Applied to Course Timetable Application Results

The timetabling problem consists of fixing a sequence of meetings between the teacher and students in a prefixed period of time, satisfying a set of constraints of various types [13]. Timetabling problems can arise in many different settings,

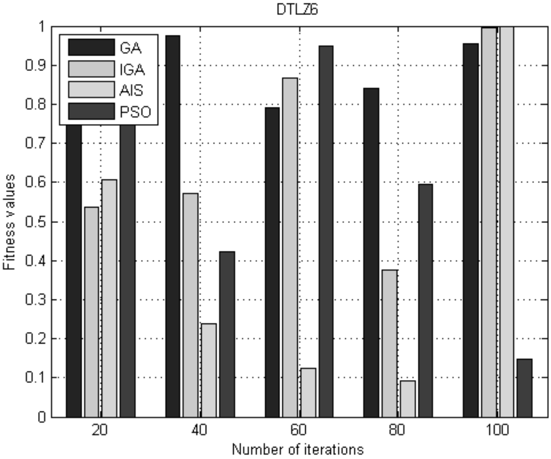


Fig. 18: The DTLZ6 testing function

but generally it refers to the timetabling at educational institutions. The significance of this problem is mainly due to the difficulty of constructing a feasible timetable that satisfies the preferences of the administration, the instructors, and the students. In certain cases, it may be extremely difficult even to find a single feasible solution. The main timetabling class problems can be divided into three classes: School, Course, and Examination timetabling. The school timetabling is a weekly scheduling for all the classes of a school avoiding teachers attending two classes at the same time and vice versa. The course is a weekly scheduling for all lectures of a set of university courses, minimizing the overlap of lectures of courses having common students. Finally, the examination timetabling is a schedule for the exams of a set of university courses, avoiding overlap of exams of courses having common students, and spreading the exams for the students as much as possible.

Regardless of the differences among problem types, similar solution approaches can be used for all. This paper focuses on the course timetabling problem. The problem can be defined as assigning a set of courses $E = e_1, e_2, \ldots, e_e$ into a limited number of ordered timeslots $T = t_1, t_2, \ldots, t_t$ and rooms of certain capacity in each timeslot $C = C_1, C_2, \ldots, C_t$, subject to a set of constraints. The complexities and the challenge presented by timetabling problems arise from the fact that a large variety of constraints, some of which contradict each other, need to be satisfied in different institutions [14]. Two types of constraints can be defined in every timetabling problem. First, the constraints which are basic for the feasibility of the timetable obtained are normally called hard constraints:

1. The class meetings of two courses with the same student enrollment cannot be assigned to the same time slot.

2. An instructor cannot teach more than one class meeting at the same time slot.

3. The assigned number of class meetings that are scheduled for the same time slot cannot exceed the number of available classrooms.

4. All class meetings should be assigned to a time slot.

Second, the constraints which do not affect the feasibility of the solution found, but their fulfillment makes it more appropriate in terms of some defined criteria. These constraints are usually called soft constraints:

1. The professor prefers to take classes in the morning or in the afternoon.

2. The professor prefers classes in week days.

3. The professor wants gaps between his lectures.

4. The student may not have two consecutive classes.

The objective of this problem is to satisfy the hard constraints and to minimize the violation of the soft constraints. Although, this model of the problem lacks many of the constraints and resource issues found in real world problems.

In the original formulation [15], there are E exams to be scheduled in P periods with S exam seats available for each period. There are three periods per weekday and two periods on Saturday. No exam is held on Sundays. It is assumed that the exam period starts on a Monday.

The problem can be formally specified by first defining the following

a_{ip} is on if exam i is allocated to period p, zero otherwise.
c_{ij} is the number of students registered for exams i and j.
s_i is the number of students registered for exam i.

The corresponding mathematical formulation is as follows:

$$\text{Maximum} \sum_{i=1}^{E-1} \sum_{j=i+1}^{E} \sum_{p=1}^{P} a_{ip}\, a_{j(p+1)}\, c_{ij} \tag{13}$$

and

$$P \tag{14}$$

Subject to

$$\sum_{i=1}^{E-1} \sum_{j=i+1}^{E} \sum_{p=1}^{P} a_{ip}\, a_{jp}\, c_{ij} = 0 \tag{15}$$

$$\sum_{i=1}^{E} a_{ip}\, s_i \leq S, \forall\, p \in \{1, \ldots, P\} \tag{16}$$

$$\sum_{p=1}^{P} a_{ip} = 1\, \forall\, i \in \{i, \ldots, E\} \tag{17}$$

Equation (13) is the objective of minimizing the number of clashes in a timetable, which is the solitary objective of the original formulation [15]. In order to prevent excessively long timetables in the process of achieving (13), the multi-objective formulation studied in this paper considers the minimization of the number of periods used in a timetable as the second objective (14). Equation (15) is the constraint that no student is to be scheduled to take more than one exam at any one time, while (16) states a capacity constraint that for each period, there must be sufficient seats for all the exams that are scheduled for that period. These two hard constraints define a feasible timetable. Equation (17) indicates that every exam can only be scheduled once in any timetable.

Timetable Evaluation

In order to test the effectiveness of IGA, AIS, and GA when solving timetabling problems, a comparison with the PSO algorithm was performed to investigate trends of performance. All coding was written in JAVA, and the test case focused on the four above algorithms. All tests were executed on a 3.30 Ghz Intel core i5 processor with 16 GB of ram. The convergence graph for IGA, AIS, GA, and PSO below shows progress until a valid solution for each of the algorithms were discovered. The specifications of the problem are shown in Table 1.

Timetable Results

This section shows the results of the timetable application coded with Java programming. The sample data test of the problem was divided into two problems, small and large. The small problems were acquired from the Faculty of Information Technology which included 23 classes, 33 subjects, 20 lecturers, 10 rooms, and 18 timeslots. And the large problems were acquired from the Faculty of Applied Science which included 48 classes, 278 subjects, 167 lecturers, 101 rooms, and 18 timeslots. Both datasets were from faculty members of King Mongkut's University of Technology North Bangkok.

When performing the simulations, it was found that the convergence of the algorithm depends on a number of timetabling initials including the number of subjects, professors, and hours per week of each subject. The graph in Figure 8 provides a comparison of the proposed algorithm with the conventional population operator based algorithm.

In the PSO calculation, the practice set size of population for GA, AIS, and IGA with $c_1 = 2$, $c_2 = 2$, was $w = 1/(2\log 2)$. In the GA and IGA calculation, *mutation rate* = 0.02 and *crossover* = 0.9. By observing the simulation results and the graphs, it is inferred that, for most of the populations, there was no great difference between the execution times and fitness values for the GA, AIS, IGA and PSO. Until the population reached 500, the IGA had an increase of time whereas PSO remained linear. However, the GA and IGA fitness values rose more than PSO when the population increased. In AIS, it still has less fitness values at the beginning but then rose higher when the population increased. This might stem from the fact that the parameters of GA, AIS and IGA were different from PSO such as velocity, global and local search space, and so on. Over all, when comparing between IGA and GA, it is cleared to see that IGA had less fitness than GA but IGA consumed less time to reach the optimal solution than GA. This is mainly because IGA detected and reduced the number of individuals during the initialize process. Off course, the search space could be less and also the generation of run time more effective. For the next investigation, the fitness value was studied focusing on the effectiveness of crossover probability varying from $0.0 - 1.0$ increasing in 0.1 increments at each step, as shown in Figure 19. The size of problem was then reduced to a medium size, the number of individuals and generations were 500. In this evaluation, the fitness values were determined when it was stable at 1.0. The results illustrated that the highest fitness value occurred when crossover probability was in the range of 0.9. In the other hand, the AIS and PSO could not be evaluated in Tables 3 to 5 because these two algorithms contained different parameters.

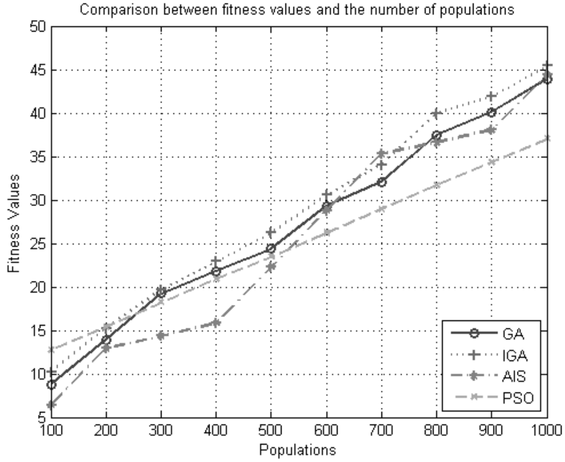


Fig. 19: The algorithm comparison between fitness values and the number of populations

For the next investigation, the fitness value was studied focusing on the effectiveness of crossover probability varying from 0.0 – 1.0 increasing in 0.1 increments at each step, as shown in Table 3. In this evaluation, the fitness values were determined when it was stable at 1.0. The results illustrated that the highest fitness value occurred when crossover probability was in the range of 0.9.

Table 4 illustrates the effects of mutation rates with two different crossovers with probability at 0.9, which is the best result of fitness value as found in Table 4. The result shows that the highest fitness value occurred when mutation rates were in the range of 0.02 with double point crossover. This is the optimal value to solve the problem because of the soft constrain violations from the fitness function described in the previous section. Finally, Table 5 illustrates the effects of mutation rates with two types of selection methods. In this evaluation, the double point crossover was used and increased 0.02 at each step of the mutations. The result shows that the highest fitness value occurred when mutation rate was in the range of 0.02.

Table 3: Effect of crossover effects probability to fitness value

Crossover	Number of Generations		Fitness Value	
	GA	IGA	GA	IGA
0.0	195	454	22471	23344
0.1	399	104	23493	23720
0.2	24	208	23860	24028
0.4	75	178	22954	24101
0.7	59	16	18532	24954
0.9	75	20	**24413**	**25596**
1.0	75	22	24285	25306

Table 4: Effects of mutation rate with different types of crossover probability to fitness value

Mutation	Single Point Crossover		Double Point Crossover	
Rate	GA	IGA	GA	IGA
0.02	**18989**	**25282**	**18184**	**25646**
0.04	18603	25219	18085	25439
0.06	18520	24593	17343	25212
0.08	17741	24543	17666	24067
0.10	17898	24758	17138	24065
0.12	16988	24532	17286	24377
0.14	17293	24426	17815	23988
0.16	17434	24199	17397	24251
0.18	17459	24066	24109	16820
0.20	17098	23989	24026	16389

Table 5: Effects of mutation rate with different selection methods and double point crossover probability to fitness value

Mutation	Roulette-Wheel Selection		Tournament Selection	
Rate	GA	IGA	GA	IGA
0.02	**17723**	**24019**	18084	**25646**
0.04	17465	24003	**18185**	24439
0.06	17077	23941	17343	25212
0.08	16986	23890	17666	24067
0.10	16045	23591	17138	24065
0.12	16249	23421	17286	24377
0.14	16514	23322	17815	23988
0.16	16136	23363	17397	24251
0.18	16122	23450	16820	24109
0.20	16155	23597	16389	24026

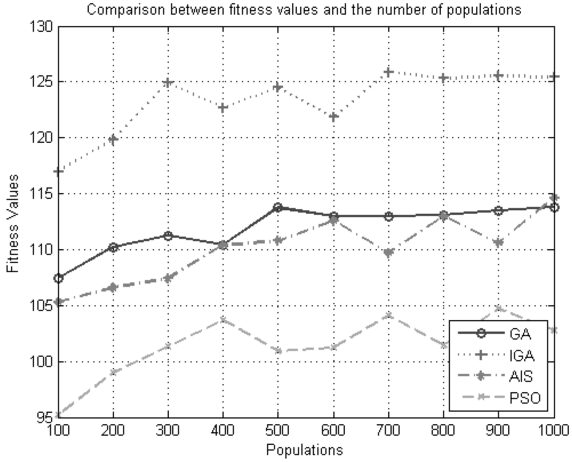


Fig. 20: Comparison between fitness values and the number of populations with large size problems

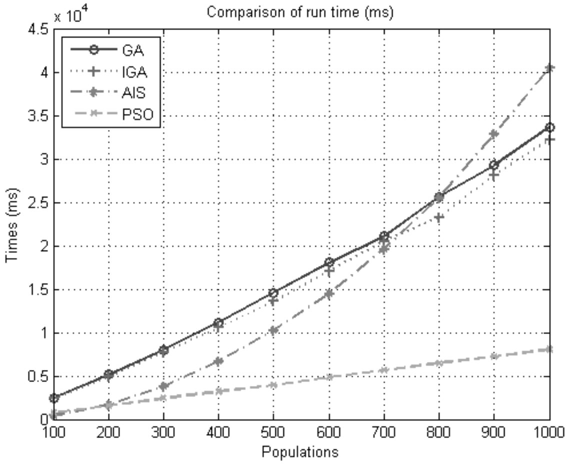


Fig. 21: Comparison of run time (ms) with large size problems

Next was the investigation on large problem sizes, Figure 20 illustrates the result comparison of population sizes. The graph shows the modified algorithm performed better than others and also reduced running time from the original algorithms, apart from the PSO which achieved by local search results faster than others (shown in Figure 21) but fitness was still poor. Figure 22 to 25 illustrates the steps of comparison from effects of crossover and mutation. The results show that the modified algorithm can reach optimal fitness values better than GA in each step and takes less time to find the optimal values.

Figure 22 shows the results of effects from crossover ranged 0.1 to 1. One can see that IGA performed better than GA as it served the highest fitness values. The optimal values were discovered when crossover reached 0.7, the fitness values were 128. Moreover, Figure 23 shows that IGA saved time when searching for the optimal solution, the result from Figure 23 is 2,636 ms. Thus, IGA is faster than GA at around 3,083 ms.

Figure 24 shows the results of effects from mutation ranged 0.2 to 2. The IGA performed better than GA as it served the highest fitness values. The optimal values are observed when mutation is 0.2 where the fitness values are 128. Moreover, Figure 25 shows that IGA also saved time when searching for the optimal solution, the results from Figure 25 is 2,636 ms.

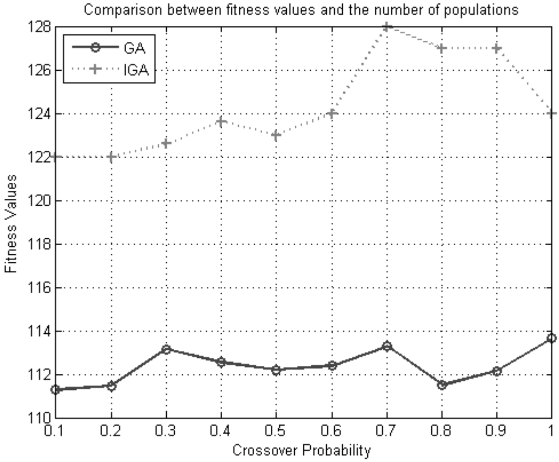


Fig. 22: Effects of crossover probability to fitness value with large size problems

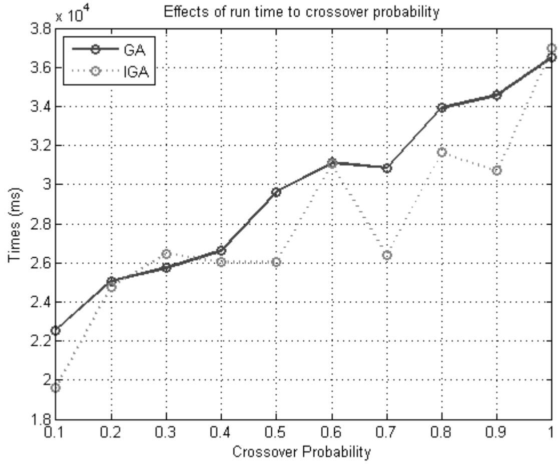


Fig. 23: Effects of run time to crossover probability with large size problems

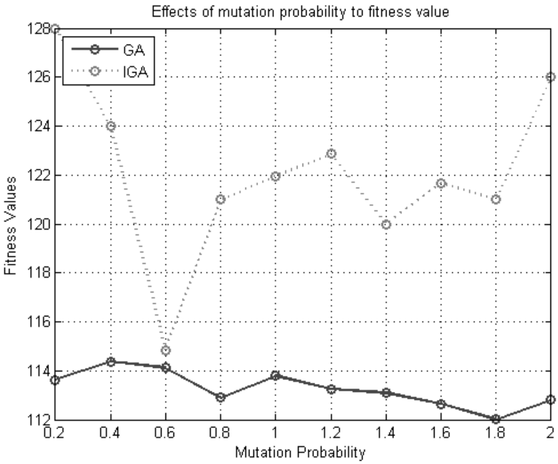


Fig. 24: Effects of mutation probability to fitness value with large size problems

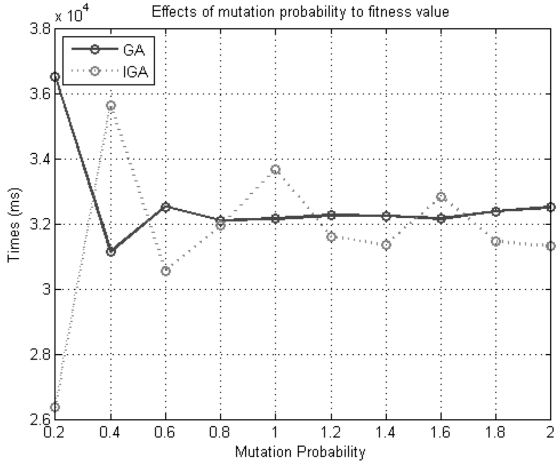


Fig. 25: Effects of run time to mutation probability with large size problems

5 Conclusion

This paper presented the evaluated measurement of a modified method from a Genetic Algorithm (GA) and the Artificial Immune System (AIS) called the Immune Genetic Algorithm (IGA). Although Genetic Algorithms can rapidly locate the region in which the global optimum exists, they take a relatively long time to locate the optimum in the region of convergence. In practice, the population size is finite, which influences the sampling ability of a genetic algorithm and as a result affects its performance. Incorporating a negative selection method within a genetic algorithm can help to overcome most of the obstacles that arise as a result of finite population sizes. Due to the GA limited population size, a Genetic Algorithm may also sample bad representatives of good search regions and good representatives of bad regions. A negative selection method can ensure fair representation of the different search areas by sampling their self and nonself antigen which in turn can reduce the possibility of population size. The new modified algorithm could be used to improve the quality of initial solutions which are generated randomly.

Negative selection which is one of techniques in the Artificial Immune System, was employed to determine the input variables (populations) of the system. Basic concepts of negative selection theory, especially, the concept of attribute reduction were also used to define the chromosome populations. In order to illustrate the effectiveness of the Immune Genetic Algorithm, the comparison with a steady-state genetic algorithm, artificial immune system, and particle swarm optimization were also investigated.

The testing of the performance was conducted in two parts. First, the mathematical function testing. In the mathematical testing, problems were divided into single and multiple objectives. The five single objectives were then used to test the modified algorithm, the results showed that IGA performed better than all of the other methods. The DTLZ multi-objective testing functions were then used. The result also illustrated that the modified approach still had the best performance.

The second part of testing was the implementation of the education timetabling problem. The implementation was coded with Java programming. The size of the chromosomes encoding was separated into two problems from schedules at King Mongkut's University of Technology North Bangkok, Thailand. The results revealed that the proposed modified approach was able to find a feasible solution for problem instances and provided superior performance for small instances by generating the best solution with no violation of constraints existing in the solution. The small problem fitness values discovered illustrated that the IGA fitness value was 25,596 whereas the GA fitness value was 24,413. From another test focusing on large problems, the IGA fitness value was 128 whereas GA was 113.28. It is clear to see that the proposed modified approach was able to find a feasible solution for problem instances and provided superior performance for small instances by generating the best solution with no violation of constraints existing in the solution.

Thus, suggested optimal attribute crossover is 0.9 and mutation rate is 0.02. The selection mechanism would be tournament selection with a value set of 0.02, this would find the optimal fitness values efficiently.

References

[1] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[2] T. A. El-Mihoub, A. Hopgood, and I. A. Aref, ”Self-adaptive hybrid genetic algorithm using an ant-based algorithm,” in: *2014 IEEE International Symposium on Robotics and Manufacturing Automation (ROMA)*, 2014, pp. 166–171.

[3] J. Sharma and R. S. Singhal, ”Comparative research on genetic algorithm, particle swarm optimization and hybrid GA-PSO,” in: *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2015, pp. 110–114.

[4] J. Zhao, H. Xu, and W. Li, ”A hybrid genetic algorithm for Bayesian network optimization,” in: *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)*, 2014, pp. 906–910.

[5] J. Wu and Z. Lu, ”A novel hybrid Genetic Algorithm and Simulated Annealing for feature selection and kernel optimization in support vector regression,” in: *2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI)*, 2012, pp. 999–1003.

[6] L. N. D. Castro, J. Timmis, ”Artificial Immune Systems: A Novel Paradigm to Pattern Recognition,” in: *University of Paisley*, 2002, pp. 67–84.

[7] Y. Zhu and X. Qiu, ”A Hybrid AIS-based Algorithm for Solving Job Shop Scheduling Problem,” in: *2012 International Conference on Communication Systems and Network Technologies*, 2012, pp. 498–502.

[8] F. Barani, ″A hybrid approach for dynamic intrusion detection in ad hoc networks using genetic algorithm and artificial immune system,″ in: *2014 Iranian Conference on Intelligent Systems (ICIS)*, 2014, pp. 1–6.

[9] M. O. Ali, S. P. Koh, K. H. Chong, and D. F. W. Yap, ″Hybrid Artificial Immune System-Genetic Algorithm optimization based on mathematical test functions,″ in: *2010 IEEE Student Conference on Research and Development (SCOReD)*, 2010, pp. 256–261.

[10] Z. Ji and D. Dasgupta, ″Revisiting Negative Selection Algorithms″, *Evol. Comput.*, vol. 15, no. 2, pp. 223–251, Jun. 2007.

[11] T. Stibor, K. M. Bayarou, and C. Eckert, ″An Investigation of R-Chunk Detector Generation on Higher Alphabets″, in: *Genetic and Evolutionary Computation - GECCO 2004*, 2004, pp. 299–307.

[12] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, ″*Scalable test problems for evolutionary multi-objective optimization*,″ 2001.

[13] A. Schaerf, ″A Survey of Automated Timetabling,″ *Artif. Intelligence Rev.*, vol. 13, no. 2, pp. 87–127, 1999.

[14] N. Najdpour and M. R. Feizi-Derakhshi, ″A two-phase evolutionary algorithm for the university course timetabling problem,″ in: *2010 2nd International Conference on Software Technology and Engineering*, 2010, vol. 2, pp. V2-266–V2-271.

[15] C. Y. Cheong, K. C. Tan, and B. Veeravalli, ″Solving the Exam Timetabling Problem via a Multi-Objective Evolutionary Algorithm – A More General Approach,″ in: *2007 IEEE Symposium on Computational Intelligence in Scheduling*, 2007, pp. 165–172.

Stock Selection by Using an improved quick Artificial Bee Colony Algorithm

Dit Suthiwong[1], Maleerat Sodanil[1], Gerald Quirchmayr[2], and Herwig Unger[3]

[1]Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand

[2]Faculty of Computer Science, University of Vienna, Austria

[3]Chair of Communication Networks, FernUniversitat in Hagen, Germany

Abstract: Computation Intelligence have inspired many researchers to develop the capability of computer to learn and solve the complex task in real-world problems. In this work, we proposed a Artificial Bee Colony (ABC) to deal with the Stock Selection problem. We apply a Sigmoid-based Discrete-Continuous with ABC to select appropriate features for stock scoring. The empirical study tests the performance of ABC compare with Genetic Algorithm (GA) and Differential Evolution (DE) algorithm by using data from the Stock Exchange Thailand. The empirical results show that the novel model stock selection significantly outperforms in terms of investment return, diversity and model robustness.

1 Introduction

The Artificial Bee Colony (ABC) algorithm was first proposed by Karaboga [1] in 2005, ABC had shown competitive performance [2] on many real world problems [3]. It has an advantage in fewer control parameters, simple structure, easy to implement and good in exploration. ABC can efficiently deal with multimodal and multidimension problems. And It also has been successfully extended to solve multi-objective optimization problems [4]. However, The major drawback of ABC is a slow convergence speed. This is mainly caused by its solution search equation. Its search equation focus on exploration but weak on exploitation [5]. Many researchers proposed solution to improve the performance of ABC [6, 7].

A Stock Selection model is a challenge problem in finance. The investors have to make a decision based on their investment experience. Normally, The smart investors use data in the financial statement to evaluate the stock by interpreting performance and the competitiveness of each stocks compare to their competitors in same business. The example of financial data is profitability, asset ratio, net profit growth, and price ratio. In General, A Stock Selection model is comprised of two key steps, i. e, stock scoring and stock ranking. Related to the study in a Stock Selection model,It can be mainly divided into two categories: Traditional statistical regression approaches and Computational Intelligence (CI) [14] approaches. Traditional statistical regression approaches are easy to understand and implement. Example include the testing of the forecasting of stock market returns using the dividend yield, the earnings growth, and the price earnings ratio growth [8].

However, CI approaches have shown more efficient than the Traditional statistical regression approaches [9]. Many CI models have been applied to stock evaluation, such as Genetic Algorithm (GA), Artificial Neural Networks (ANNs), ABC and Differential Evolution (DE). For GA, Huang et al. [10] utilized GA to optimize feature selection, Soam et al. [11] used GA with local search to select stock into portfolio, Chen et al. [12] applied ABC create cardinality-constrained portfolio, Tsai et al. [13] combined multiple selection methods ANN, GA and Decision Tree. Yu et al. [14] used DE with sigmoid-based to solve stock scoring problem. All these example studies demonstrated that the CI approaches outperformed the traditional regression approaches.

Our proposed model is an improved quick Artificial Bee Colony algorithm by using sigmoid-based conversion. This modification enhanced ABC to solve a mixed discrete-continuous decision variable. This paper also proposed a stock selection model by evaluate stock based on various fundamental financial data features. The detailed description of the stock selection model provides is in Section 2, The methodology of improved quick ABC is in Section 3, The empirical design, training, testing and results are in Section 4, and Conclusion is in Section 5.

2 Problem Formulation

A Stock Selection is a concept of to select best stocks in to portfolio with expectation to get maximum investment return. Generally Stock Selection has two main steps, i.e, stock scoring and stock ranking. In the first step, a stock scoring can be calculated from their stock information such as profit return, amount of sale growth, etc. In the second step, a stock ranking ranks stocks according to their scores.

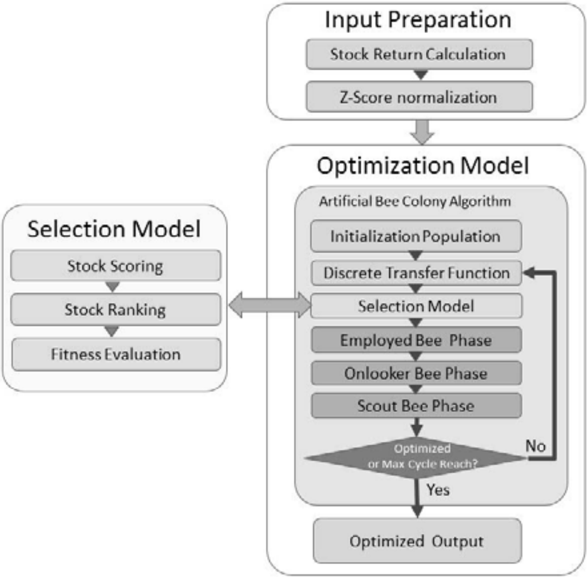


Fig. 1: Stock Selection Framework

The stock selection model in this study extended from the model proposed by Yu et al. [14] which using Differential Evolution (DE) algorithm. The proposed model uses Artificial Bee Colony algorithm and Sigmoid conversion to handle both discrete and continuous decision variables. As shown in Fig. 1, The proposed stock selection model is composed of three steps. First step is Input Preparation, Second step is Selection Model and Third step is Optimization model.

2.1 Input Preparation

In this step, The stock returns are calculated in terms of the natural logarithm of price ratio (i.e.,$R_{i,t} = ln(P_{i,t}/P_{i,t-1})$, where $R_{i,t}$ denotes return of stock i at current time t, $P_{i,t}$ denote price of stock i at current time t, $P_{i,t-1}$ denote previous price of stock.

Then The Z-score normalization of features j denote by $Y_{i,j,t}$ are calculated. The score $Y_{i,j,t}$ is assumed to follow a normal distribution with mean zero and deviation one. If a larger feature value implies that the direction will be up then $Y_{i,j,t}$

can be calculated according to the following form.

$$Y_{i,j,t} = \frac{V_{i,j,t} - \overline{V_{j,t}}}{D_{j,t}} \tag{1}$$

$$\overline{V_{j,t}} = \frac{1}{N} \sum_{i=1}^{N} V_{i,j,t} \tag{2}$$

$$D_{j,t} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (V_{i,j,t} - \overline{V_{j,t}})} \tag{3}$$

where $V_{i,j,t}$ is the actual score of feature j and $\overline{V_{j,t}}$ is the average value of feature j across all N stocks at time t, and $D_{j,t}$ is the standard deviation of feature j at time t. If a smaller value implies that the direction will be up then $Y_{i,j,t}$ can be calculated according to the following form.

$$Y_{i,j,t} = \frac{\overline{V_{j,t}} - V_{i,j,t}}{D_{j,t}} \tag{4}$$

2.2 Selection Model

Stock Scoring

Feature Selection had been proposed to solve complexity of stock scoring. Various stock features including profitability, Price ratio, Growth, Efficiency can be used to calculate score. The model proposed by Yu et al.[14] applied the wrapper method and optimized solution by using Differential Evolution (DE). The model combines feature selection with the corresponding weight optimization. First, binary variable $F_j = \{0, 1\}$ is utilized represent whether feature j is used in stock evaluation $(F_j = 1)$ or not $(F_j = 0)$, Let W_j denote the weight on the jth feature. The stock scoring $S_{i,t}$ of stock i at time t can be formulated as follows:

$$S_{i,t} = \sum_{i=1}^{N} F_j W_j Y_{i,j,t} \tag{5}$$

Stock Ranking

A Stock Ranking sorts stock according to their scores. Let $r_{i,t=\{1,2,..,N\}}$ denote the ranking of stock i at time t, i.e., $r_{i,t} \leq r_{k,t}$ if $S_{i,t} \geq S_{k,t}$ where $i, k \in \{1, 2, ..., N\}$ represent any two different stocks. The higher rank means a high potential of price increase. Then the model creates an equal-weighted portfolio for the next period by selecting the stock with m rankings $r_{i,t} = \{1, 2, ..., m\}$ at the end of each period. Accordingly, The average return of portfolio in next period is calculated and used for evaluate the performance of all selected stocks.

$$R^p_{t+1} = \frac{1}{m} \sum_{r_{i,t}=1}^{m} R_{t+1}(r_{i,t}), \tag{6}$$

where $R_{t+1}(r_{i,t})$ is the next period return of stock with current ranking $r_{i,t}$ at time t, and R^p_{t+1} is the next period return of the portfolio constructed by the proposed model.

Fitness Evaluation

To capture effectiveness of the selection model, The Information Coefficient (IC) between features and future returns has been selected as fitness function.

$$minF = -\frac{1}{T} \sum_{t=1}^{T} IC_t, \tag{7}$$

$$IC_t = \frac{cov(r_{i,t}, r'_{i,t+1})}{\sqrt{var(r_{i,t})var(r'_{i,t+1})}} \tag{8}$$

where $r_{i,t}$ is the score ranking of stock i by the proposed model at time t,$r'_{i,t+1}$ is the actual return ranking in the next period, and T is the total number of training periods. The function cov() and var() are the covariance and variance estimations. IC_t is the Spearman correlation between the currently predicted ranking of stock $r_{i,t}$ and their actual return ranking $r'_{i,t+1}$ in the next period.

2.3 Optimization Model

To Optimize solution in term of feature selections F_j and the corresponding weights W_j, The Artificial Bee Colony (ABC) algorithm has been proposed to solve problem. Since the original ABC focus on continuous values, This study improved ABC algorithm by using the sigmoid-based mixed discrete-continuous.

The Original ABC Algorithm

In nature, Honey bees live in colonies and have a social communication. Artificial bee colony (ABC) algorithm[1] is a population-based evolutionary algorithm which mimic communication of Honey bees. There are three groups of bees in the colony who responsible to find and collect food to nectar. They are employed bees, onlookers bees and scouts bees.The pseudo code is shown as below.

```
Initialization phase
REPEAT
   Employed bee phase
   Onlooker bee phase
   Scout bee phase
UNTIL (cycle=maximum cycle number or Optimized)
```

ABC algorithm start with initialization phase which create population by random create food sources represent to a possible solution to the problem space. The nectar amount of a food source denotes the quality of the associated solution.

Then ABC algorithm repeat execute all three phases until max cycle number reach or found optimized solution. Employed bees phase takes charge of exploring the solution space in which nearby to existing food sources. Onlooker bees phase chooses a number of food sources to exploit based on the perform waggle dance of Employed bees. When there are no improvements to the food sources after several times Scout bees phase will sent scout bee out randomly to find new brand food sources. If the new food source has better nectar they will memorize the new position and forget the previous one.

Proposed improved quick ABC

The proposed model modified original ABC algorithm by using Sigmoid conversion to solve mixed discrete-continuous variable. And the proposed model also including Selection model in previous step to evaluate effectiveness of optimized solution. The pseudo code is shown as follow.

```
Initialization phase
   Discrete Transfer Function
   Selection Model
```

```
REPEAT
   Employed bee phase
     Discrete Transfer Function
     Selection Model
   Onlooker bee phase
     Discrete Transfer Function
     Selection Model
   Scout bee phase
     Discrete Transfer Function
     Selection Model
UNTIL (cycle=maximum cycle number or Optimized)
```

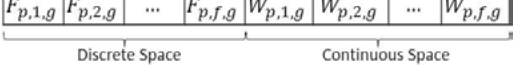
Initialization Phase

In this phase, food sources are randomly initialize with Equation (9) in a given range.

$$x_{p,d} = l_d + rand(0,1) * (u_d - l_d) \tag{9}$$

where $x_{p,d}$ is the value of the d dimension of the p solution. l_d represents the lower bound and u_d represents the upper bound of the parameter $x_{p,d}$.

In proposed model, p represents population of food sources and d represent dimension of search space. Number of dimensions are equal to number of features and number of their corresponding weights. The dimension of search space is shown in Fig. 2.



$F_{p,1,g}$	$F_{p,2,g}$	...	$F_{p,f,g}$	$W_{p,1,g}$	$W_{p,2,g}$	...	$W_{p,f,g}$

Fig. 2: Encoding Feature Selection and Weight

Discrete Transfer Function

The discrete term $F_{p,j,g} = \{0,1\}$ represents the selection decision of feature j in pth solution at iteration g, and the continuous term $W_{p,j,g} \in [0,1]$ is the corresponding weight. If $F_{p,j,g} = 0$ then $W_{p,j,g}$ is accordingly set to 0.

Sigmoid Conversion method is used to identify a candidate feature as a key factor for stock scoring ($F_{p,d,g} = 1$) or a poorly information one ($F_{p,d,g} = 0$). The conversion from the continuous variable $x_{p,d,g}$ to the binary form $F_{p,d,g}$ is conducted according to the probability $P(x_{p,d,g})$ which follows a logistic distribution formula as follow.

$$F_{p,d,g} = \begin{cases} 1 & ,if\ r_{p,d,g} \leq P(x_{p,d,g}) \\ 0 & ,otherwise, \end{cases} \tag{10}$$

$$P(x_{p,d,g}) = \frac{1}{1 + e^{-x_{p,d,g}}}, \tag{11}$$

where $r_{p,d,g}$ is a random term following a uniform distribution on the range of (0,1). ABC algorithm can be extended to the sigmoid-based ABC for mixed discrete-continuous problems.

Selection Model

By using food sources from initialization phase calculates with Z-score normalization of features j. The proposed evaluates fitness function by using stock scoring and stock ranking as shown in previous topic. The fitness value from Selection model is used in ABC algorithm to evaluate food sources quality by following formula.

$$fit(x_m) = \begin{cases} 1/(1 + f(x_m)) & ,if\ f(x_m) \geq 0 \\ 1 + abs(f(x_m)) & ,if\ f(x_m) < 0 \end{cases} \tag{12}$$

where x_m represents fitness value of the stock selection model.

Employed Bees Phase

Each food source will be further exploited by one and only one employed bee. Then it search for the neighborhood of a target food source. This study focuses on the change of the associated v_m values for each asset selected by equation as follows:

$$v_{m,i} = x_{m,i} + \phi_{m,i}(x_{m,i} - x_{k,i}) \tag{13}$$

In equation (13) x_k is a food source selected from neighborhood randomly and i is also a randomly weight parameter. $\phi_{m,i}$ denotes a random number generated from a uniform distribution with the range of [-1,1]. $v_{m,i}$ is the new candidate food source which will be converted with Discrete Transfer Function and then evaluate fitness value with Selection Model. Then, A greedy selection is applied between v_m and x_m. If a food source is not be improved employed bee will increase a certain number of iteration limit by one.

Onlooker Bees Phase

The onlooker bees determine the food sources to search using the probability based on the quality of each food source. The probability of selection p_m can be calculated as follows.

$$p_m = \frac{fit(x_m)}{\sum\limits_{m=1}^{SN} fit(x_m)} \tag{14}$$

According to the probability, Onlooker bees choose a food source v_m to exploit by using equation (13) similar to employed bees, its fitness value is computed by using Discrete Transfer Function and Selection model. Then, A greedy selection is used to determine between v_m and x_m.

Scout Bees Phase

If some of food sources cannot be improved through a certain number of iterations limit. Scout bees will be dispatched to explore new brand food sources. This make new solutions randomly generated by using Equation (9). Discrete Transfer Function and Selection executed to calculate fitness value. Then, A greedy selection is used to determine between the old solution and the new food source. If new solution is better than the old solution, then the old food source will be replaced with new food source.

3 Proposed improved quick Artificial Bee Colony Algorithm

It is known that ABC algorithm is powerful. But it still has some drawbacks such as slow convergence. Because the search equation of ABC is good in exploration but badly in exploitation. The proposed model improve ABC algorithm by apply Gbest concept in Employed Bee phase and apply quick ABC concept in Onlooker Bee phase.

3.1 Modification in Employed Bee Phase

Inspired by the Particle Swarm Optimization algorithm, We proposed to use global best(gbest) apply as follows.

$$v_{m,i} = x_{m,i} + \phi_{m,i}(x_{m,i} - x_{k,i}) + \varphi_{i,j} * (gbest_i - x_{k,i}) \tag{15}$$

In Employed bee phase, We replace equation (13) with (15). Where φ represents a uniformly distributed random number in [0,1.5], $gbest$ is the current global best solution in the whole swarm, and $gbest_i$ represents the i th variable of $gbest$.

3.2 Modification in Onlooker Bee Phase

By Comparing to nature of honey bees, Employed bees and Onlooker bees exploit foods in different ways. Employed bees exploit the food source that they visit before. Onlooker bees exploit food source based on communication from employed bee dancing (we called "wagged dance") which will be interpreted for which food sources will be selected. In Original ABC Employed bees and Onlooker bees use the same search equation (13). Karaboga [6] introduced new equation of onlooker bees phase modified. The equation for onlooker bee had been modified as follow.

$$v_{N_m,i}^{best} = x_{N_m,i}^{best} + \phi_{m,i}(x_{N_m,i}^{best} - x_{k,i}) \tag{16}$$

From equation (16), $x_{N_m,i}^{best}$ represents the best solution between the neighbors of x_m and itself N_m. The neighborhood of individual m is determined by the Euclidean distance between X_{N_m} and the other food sources. The mean Euclidean distance between x_m and the rest food sources is calculated and then compare it with A new parameter r which refers to the "'neighborhood radius"' is added into the parameters of standard ABC algorithm. If a solution which Euclidean distance from x_m is less than the mean Euclidean distance md_m then this food sources could be accepted as a neighbor of x_m as equation (17).

$$md_m = \frac{\sum_{j=1}^{SN} d(m,j)}{SN-1} \tag{17}$$

This solution is similar to nature that onlooker bees selects the region which is centered by the food source x_m. The pseudo-code for determine a neighbor of x_m is given as follow.

```
if $d(m,j) \leq r \times md_m$ then $x_j$ is a neighbor of $x_m$, else not
```

4 Computational Analysis

Our proposed model improved quick ABC algorithm. It is compared with - GA, DE, and original ABC. The proposed model is coded in Matlab R2013a and run on a notebook computer with Intel Core i7-4510U CPU 2.00GHz and 4.0 GB memory.

4.1 Test Instance

We brought trading data from Stock Exchange Thailand from Quater1,2012 to Quater1,2017 and then calculate return from about 600 stocks. Closing price from the last trading day of the quarter are used to compare with closing price in previous quarter to calculate stock return. The testing period is formulated by 4-quarters, 8-quarters, 12-quarters and 16-quarters.

In this study, Fifteen Candidate Features has been used to test for Stock Selection. Example Stock features are Price per Earning ratio, Price per Book Value ratio, Dividend Yield, Total Assets, Revenue, Net Profit, Earning Per Share and others.

4.2 Parameter Setup

ABC and iqABC parameter setup are the same by using Population(P)=30, Limit=10, Generation=50. GA uses Generation=2000, Population=30. And DE uses Generation=1000, Population=30, beta=0.6, Cr=0.5.

4.3 Performance Measures

To Evaluate performance of proposed model, This study evaluates in terms of the average return of the formulated portfolios over all testing periods:

$$MR = \frac{1}{T'} \sum_{t=1}^{T'} R^{p}_{t+1}, \tag{18}$$

where R^{p}_{t+1} is the next period return of the portfolio formulated at time t, and T' is the total number of testing periods.

4.4 Empirical Results

All algorithms are executed 10 times on each instance with a different random seeds. The test result as show in table 1. It is obvious seen that iqABC given high return. Especially for 8-quarters testing period return from portfolio is 0.07634 or equal to 7.634% compare to market average return at 4.368%. The original ABC is second high portfolio return for 8-quarters testing data.

Table 1: Comparison Portfolio for different periods

	Periods			
Algorithms	4	8	12	16
GA	0.04312	0.06199	0.06728	**0.05922**
DE	0.01224	0.01919	0.02754	**0.03339**
ABC	0.03025	**0.07362**	0.05386	0.07265
iqABC	0.01136	**0.07634**	0.03218	0.05421

5 Conclusion

The Artificial Bee Colony Algorithm had shown its powerful to solve the Stock Selection problem. For qABC, It also shows improvement for accuracy and convergence. And We have proposed iqABC which developed based on qABC. By using Gbest direction concept in employed bee phase, we got the more accuracy result and more non-dominated points compare to others.

For future works, In algorithm development topic about convergence should be developed search equation and detail work in algorithm to enhance result and reduce time of execution.

References

[1] Karaboga D., *An idea based on honey bee swarm for numerical optimization[R]*. Technical Report, Computer Engineering Department, Engineering Faculty, Erciyes University, 2005.

[2] D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, *Journal of Global Optimization*, 2007.

[3] Y. C. Liang, A. H. L. Chen, Y. H. Nien, Artificial Bee Colony for workflow scheduling, in: *2014 IEEE Congress on Evolutionary Computation (CEC)*, 2014, pp. 558–564.

[4] Z. Wang, M. Li, L. Dou, Y. Li, Q. Zhao, J. Li, A novel multi-objective artificial bee colony algorithm for multi-robot path planning, in: *2015 IEEE International Conference on Information and Automation*, 2015.

[5] L. Cui et al., A novel artificial bee colony algorithm with depth-first search framework and elite-guided search equation, *Information Sciences*, Vol. 367, pp. 1012–1044, 2016.

[6] D. Karaboga, B. Gorkemli, A quick artificial bee colony -qABC- algorithm for optimization problems, in: *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 2012.

[7] Y. Shi, C.-M. Pun, H. Huà, H. Gao, An improved artificial bee colony and its application, *Knowledge-Based Systems*, 107, pp. 14–31, 2016.

[8] M.A. Ferreira, and P. Santa-Clara, Forecasting stock market returns: The sum of the parts is more than the whole, *Journal of Financial Economics*, Vol. 100, No. 3, pp. 514–537, 2011.

[9] R. C. Cavalcante, R. C. Brasileiro, V. L. F. Souza, J. P. Nobrega, A. L. I. Oliveira, Computational Intelligence and Financial Markets: A Survey and Future Directions, *Expert Systems with Applications*, pp. 194–211, 2016.

[10] C. F. Huang, T. N. Hsieh, B. R. Chang, C. H. Chang, A Comparative Study of Regression and Evolution-Based Stock Selection Models for Investor Sentiment, in: *2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications*, 2012, pp. 73–78.

[11] V. Soam, L. Palafox,H. Iba, Multi-objective portfolio optimization and rebalancing using genetic algorithms with local search, in: *2012 IEEE Congress on Evolutionary Computation (CEC)*, 2012.

[12] A. H. L. Chen, Y.-C. Liang, C.-C. Liu, Portfolio optimization using improved artificial bee colony approach, in: *2013 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr)*, 2013.

[13] C.-F. Tsai, Y.-C. Hsiao, Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches, *Decision Support Systems*, 50, 1, pp. 258–269, 2010.

[14] L. Yu, L. Huà, L. Tang, Stock Selection with a Novel Sigmoid-Based Mixed Discrete-Continuous Differential Evolution Algorithm, in: *IEEE Transactions on Knowledge and Data Engineering*, 28, 7, pp. 891–1904, 2016.

Named Entity Recognition in Thai Messages to Detect Natural Disasters

Maleerat Sodanil[1], Yuttana Lungkatoong[1] and Gerald Quirchmayr[2]

[1]Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand

[2]Faculty of Computer Science, University of Vienna, Austria

Abstract: Social media platforms like Twitter have become extremely popular forms of blogging, especially on the mobile web. It is much more convenient to communicate with people compared to other interaction norm. To create a way that easier and faster to communicate with people online and keep them informed about relevant to disaster management, shareable information at the same time. These information are beneficial for people in order to help and support just in time or in advance. Normally, data type can be either structured or unstructured which is difficult to extract only useful information. This research aims to develop a method that can be detected and recognized the entity name which is suitable for using in disaster management. A microblog corpus created from sampled tweets of 100,000 words between January 28th and September 14th, 2016. An algorithm for Named Entity Recognition (NER) which applied in this research as a semi-supervised learning is Conditional Random Fields (CRFs) which naturally represent rich domain knowledge with features. The performance was evaluated by comparing with other related algorithms. The results given such a good score which can be applied to use in the real world.

1 Introduction

Disaster management is one of the most interesting research. As a result of the natural disaster caused serious damage both to human life and property. Once the scene of natural disasters, people tend to communicate by using different channels. The most popular channels to communicate is via social media. The nature of information dissemination through social media are much faster than traditional methods. Users can either publish or subscribe at the same time.

Twitter is a website that allow users to send short message with a length not exceeding 140 letters. As one of the social media is used to publish information about the disaster was. These information can be used in various fields of planning and managing about helping victims. However, the problem of the information is that caused by the amount of information from the user to publish through Twitter of enormous. Many researches using a collection data from in order to generate and test the model, for example Kongthon et al. [1] analyses the distribution of information during the flooding in Thailand. During the month of October to December 2011, the results shown that people use Twitter to disseminate information about flood was increased exponentially during the flood in a year. Maleerat and Yuttana [2] studied about Thai text classification from microblog twitter which related to the earthquake disaster. There are four categories of the message:

1. About the disaster situation,
2. Surveillance of the earthquake,
3. the message about the request and providing assistance to disaster victims and
4. other text that is irrelevant.

In which each text in addition to separate categories and also other information, both beneficial to assemble together in one message, such as the type of disaster. Organizations, people, places, and other irrelevant or useful. This research aims to extract and classify the message by means of proper name recognition (Named Entity Recognition: NER) from the tweet messages of Thai language which related to natural disaster in order to apply for disaster management and planning.

2 Related Works

2.1 Social Media and Disaster Management

In the present, social media has becomes a very popular platform that allowed people to share and distribute information in various type of digital such as text images and video. Not only people who familiar with internet technology, but also the elderly people and novice user who necessary to use social media when some situation happened and they need some helps from people around the world. Twitter is one of the social media site that users prefer to tweet the text message with the size not exceeding 140 characters. They can use tag in order to identify the subject of tweet and grouped them together.

Recently, there were many disaster happened around the world which leads to financial, environmental or human losses. However, natural disasters come without warning and they take lives of tens, hundreds and thousands of people. While natural disasters occurring, there is an increased communication via social media because people needs to contact family and friends in the disasters zone, and searching for information regarding food and helps. So that why social media has played a significant role in disseminating information about these disasters by allowing people to share information and ask for help.

In Thailand recently, natural disaster has occurred several times, such as the big flood in the 2011. At that time, the tweet message were disseminated and increased rapidly. In other country like China, social media is the main choice of information distribution as well. In 2013, Yanquan [3] proposed a classification of microblogs for support emergency responses. The case study used a microblogs of Yushu earthquake in China. The results of their proposed model for the case achieved over 85% recall rate in average. The dissemination of information is purpose to request or provide assistance to the victims surveillance monitoring. These data are passed and published online via social media quickly that a classical media.

2.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) also known as entity identification is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Therefore, the main task of Named Entity Recognition and Classification can be described as the identification of named entities in computer readable text via annotation with categorization tags for information extraction. NER has been well studied and its solutions can be divided into three categories:

1. rule-based,
2. machine learning based and
3. hybrid methods [4] most studies focus on identifying the types of words from the text.

In [5] presented a Twitterbot for named entity extraction (NEE) and disambiguation (NED) for Tweets. The novel framework was proposed to copes with the challenges of disambiguates for tweets. If the machine can’t separate that between proper names and words, it will caused of mistakes in the translation system. For example in Thai language, the phrase “วันดีเดินทางไปต่างประเทศ” (Wandee (person name) travel abroad) the translation system translates into ”Good day to travel abroad” which is different meaning. Liu and others [4] studied the problem of named entity normalization (NEN) for tweets and proposed jointly conducting NER and NEN for multiple tweets using a factor graph, to address the rich variations challenges. Some of the research instance name ”Lady Gaga” and ”Lady Gaaaga” are the individual category and style of writing is used as the benchmark. The results showed that the method yields better F1 for NER and Accuracy for NEN than the baseline system. Also Küçük and Steinberger [6] proposed the method of named entity recognition using text from Turkish tweets with a rule-based recognition system and gradually extended it in two directions: adapting the rules/resources of the system and introducing a tweet normalization scheme into the recognition procedure. The experimental results do not increase the efficiency of named entity recognition, but when used with a specific name, several types of named entity recognition efficiency decreased. In the second experiment using the corpus with the patterns of the normal form of proper name. The experimental results performed that named entity recognition increased slightly. Che and others [7] proposed the method that formulates the problem of exploring such signals on unannotated bilingual text as a simple Integer Linear Program, which encourages entity tags to agree via bilingual constraints. The results of Bilingual NER experiments on the large Chinese-English corpus showed that the proposed method can improve strong baselines for both Chinese and English. They also suggest to apply the bilingual constraint based method to other tasks, such as part-of-speech tagging and relation extraction.

Social media is used as an online channel of communication and dissemination of information. The appearance of the information published a variety of formats and languages. Twitter is one of the social media was used a lot. Text from Twitter is sometimes added letters or adapted to different dictionaries from causing problems in cutting words. It caused the problem of the pattern of language to specific the named entity recognition like Turkish language [6], problems similar to the specific named entity recognition from the archive twitter Thai language. There are a variety of the users, such as the use of accents, some text is in used, but in some text, not in use. The specific name to the recognition is different. The repeated letters into words make a word that have no meaning and difficult to perform specific name recognition. However, the efficiency of named entity recognition from twitter Thai corpus is necessary to use it as a tool to extract useful information messages from twitter when disaster strikes.

3 Proposed Framework

The proposed framework as shown in Fig. 1 which started from the data collection process that gathering tweets message from twitter website with tag that related to natural disaster using twitterAPI. Then a set of data will be transferred to prepare for model creation and evaluation in the next step. The tokenization and labeling is done in the step of data preparation. The details for each process will be described as following:

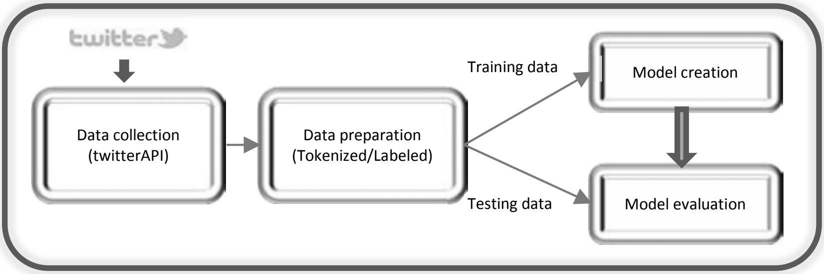

twitter
Data collection (twitterAPI)
Data preparation (Tokenized/Labeled)
Training data
Testing data
Model creation
Model evaluation

Fig. 1: Proposed Framework

3.1 Data Collection

Data collection is the first part that needs to be collected and used throughout the experiment for modeling of named entity recognition. In this research, we used word segmentation program LexTo (`http://www.sansarn.com/lexto/`) as a tool for Thai segmentation and modify some source code add a special characters (Pipe: ”|”) as a reference letter in order to increase the performance of word segmentation. The information about earthquake disasters during the 1st April 2015 to date 30th April 2015 and flooding disaster since the day 28th January 2015 to date 14th September 2015 were collected in total of 100,000 tweets. The data is divided into two sets for learning 70 percent and test data 30 percent.

The program we used in the experiments is Stanford Named Entity Recognizer (StanfordNER) to create a model for named entity recognition which given output of named entity recognition of six groups:

1. Type of disaster (TYPE),
2. Place (LOC),
3. Organization (ORG),
4. Person (PER),
5. Damage level and impact (LEV) and
6. Other words that are irrelevant (OTR).

3.2 Data Preparation

In this process, the important feature (terms) will be extracted using term-document matrix in order to prepare a training set, it needs to label according to the class as shown in Table 1. In the process of documents step, TF-IDF weighting is set as a parameter for weight vector. Then tokenizing, filtering token and attribute selection of top 3,000 attributes for the next process.

3.3 Model Creation

In this process, the classification model for Thai language will be based on natural disasters occurred in Thailand. The corpus of Thai tweets is created in order to use as training and testing model. The output of classifier is the type of six proper names:

1. words related to the type of natural disasters
2. words associated with the person
3. words about the place
4. words about agencies or organizations
5. words on the level, or the impact of disasters, and
6. the other relevant

respectively by using Conditional Random Fields (CRFs) with a specific name recognition program from Stanford University. CRFs are undirected graphical models trained to maximize the conditional probability of a sequence of labels given the corresponding input sequence. The resulting classifier is called the naive Bayes classifier. It is based on a joint probability.

3.4 Model Evaluation

The model evaluation is done along with model creation in order to find the best model. For experiment, the corpus is divided randomly into 2 sets of 70:30 for training and testing respectively. The performance of system were measured using precision recall and f-measure. The optimized model will be used to classify the tweets message with unknown type of entity via web application with the result of visualization.

4 Experimental Results

For the evaluation, Thai twitter corpus was split into training and test sets as 70:30 percent respectively. The tool we used for CRFs is Stanford NER (also known as CRFClassifier) from Stanford University.

Table 1 illustrates the precision and recall for the six types of NEs of the extracted data using CRFs. The average F-score is 87.98% and each score for six types of NEs also shown in Figure 2. Due to our corpus, the precision and recall are high enough for the most of named entity and overall for precision.

Table 1: Precision and Recall of the extracted training data

Named Entity	Precision	Recall
TYPE	0.9941	**0.9849**
LOC	0.9384	0.8294
ORG	0.8947	0.6623
PER	0.9439	0.8211
LEV	**1.0000**	0.8462
OTR	0.8625	0.7995
Total	**0.9158**	**0.8465**

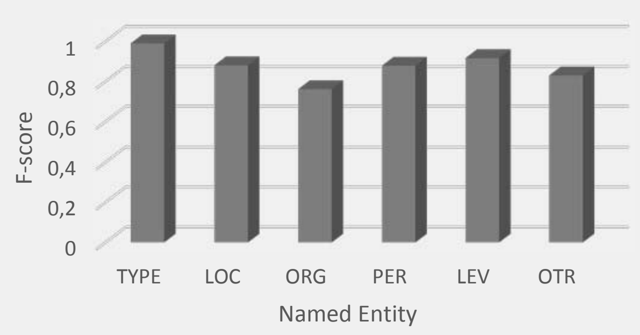


Fig. 2: Overall F-Score performance

Fig. 3 and Fig. 4 shown the results of visualization of the NER class TYPE of flood and earthquake respectively. The size of bubble indicates the number of tweets during disaster occurred in the period of data collection.

5 Conclusion

This paper presented the method for Named entity Recognition (NER) using conditional random fields (CRFs). The experiments were conducted using Thai twitter corpus which collected from twitter with tagged related to natural disaster. The results given in Table 1 show that modeling named entity recognition for natural disasters messages from microblog twitter is good enough to perform further development for information extraction from Thai language and also a part of disaster management.

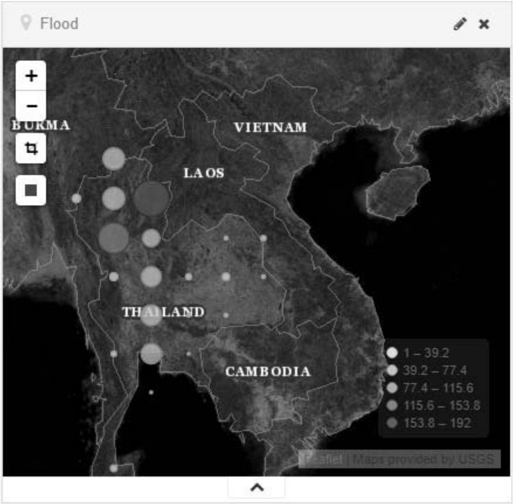


Fig. 3: Visualization with TYPE of Flood

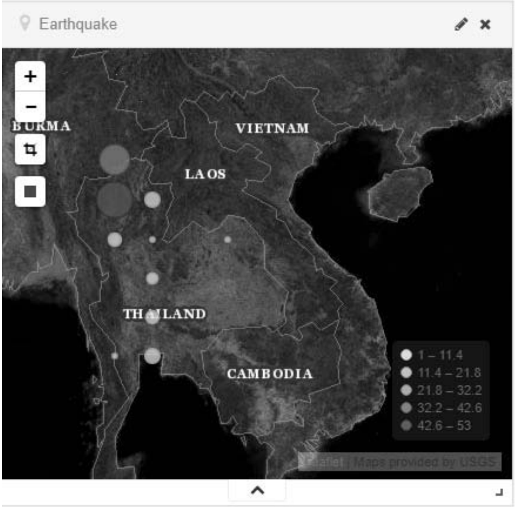


Fig. 4: Visualization with TYPE of Earthquake

References

[1] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, *"The role of Twitter during a natural disaster: Case study of 2011 Thai Flood"*, pp. 2227–2232.

[2] M. S. Y. Lungkatoong, "Classification of open source earthquake disaster information in Thai language", in: *11th National Conference on Computing and Information Technology*, 2015.

[3] Z. Yanquan, Y. Lili, B. Van de Walle, and C. Han, *"Classification of Microblogs for Support Emergency Responses: Case Study Yushu Earthquake in China"*, pp. 1553–1562.

[4] X. Liu, M. Zhou, F. Wei, Z. Fu, and X. Zhou, *"Joint inference of named entity recognition and normalization for tweets"*, pp. 526–535.

[5] M. B. Habib, and M. van Keulen, *"NEED4Tweet: a Twitterbot for tweets named entity extraction and disambiguation"*, 2015.

[6] D. Küçük, and R. Steinberger, *"Experiments to improve named entity recognition on Turkish tweets"*, arXiv preprint arXiv:1410.8668, 2014.

[7] W. Che, M. Wang, C. D. Manning, and T. Liu, *"Named Entity Recognition with Bilingual Constraints"*, pp. 52–62.

Metrology and Machine Learning

Gerhard Sartorius

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: Interest in connectionism has sharply increased in recent years. One of the main reasons for this is that this approach to problem solving has been very successful for a lot of applications in everyday life. In this approach artificial information carriers are connected to a networked system in such a way that their arrangement is able to generalize new and untrained information sensibly, after a learning or training phase. Due to their structural similarity to the neurobiological organization of the human brain, these connectionist systems are also known as artificial neural networks. They are able to process nonlinear multivariate data quickly and effectively. The rapid development in metrology in recent years has led to the fact that not only electronic and digital technologies but also processing methods are playing an increasingly important role in the representation of measurement results. For advanced measuring systems of the type presented here, the measured values are needed in digital form. It is not necessary for the multivariate processing of measured values that the measured values have a linear relationship. Nonlinear measurement series can be expediently processed using artificial neural networks. The MAE-method (Multivariate Adaptive Embedding) can be used for evaluation. Unwanted noise effects were reduced by the averaging properties of the algorithms used. Each data record structure, which represents the status of the measurement object to be processed, will be separately and optimally prepared for the subsequent processing steps by the MAE-method.

1 Introduction

This paper is a comprehensive treatment of typical circumstances and applicable definitions which are useful in the field of automated measuring and is a supplement or preface to *Data Pre Processing and Outlier Detection in Multivariate Data* [12] by the author of this paper.

Rapid developments in the field of metrology in recent years had led to the fact that not only electronic and digital technology but also methods for the processing and presentation of measurement results of data, obtained through the use of measuring devices are playing an increasingly important role. Multivariate measurement data, for example spectra, must meet additional criteria in contrast to univariate values in order to be usable as reference data. In practical applications the available data is not ordinarily free of errors. Before using such data, they must be checked and adapted accordingly. The data of the measured values are more or less noise afflicted, unfavorably scaled and/or unsuitably formatted. These inadequacies must be recognized in the data pre.preprocessing stage and taken into account so that one can design a appropriate model with the recorded data set for the task at hand. Classical metrology provides measured values in the form of display, analog or digital values as well as numerical values, which represent a spectrum. These data have to be converted into a machine-readable form, and it must be possible to store them on a data-storage medium. For modern measuring systems of the form presented here, measured values in digital format are required. In the multivariate processing of measured values, it is not necessary for the values to have a linear relationship. Nonlinear series of measured values can be readily processed using artificial neural networks (here, the MAE-method is used)[1]. Noise influences will be reduced by the algorithms used. Each data-set structure representing the status of an object to be classified will be isolated and optimally prepared for processing with the MAE-method [11].

Processing: Learning and measuring in machine learning consist of two steps:

1. Detection of a real status or object in the form of data which is usable as a reference for later processing steps: training phase, creation of a reference and design a model for the detected object or status.

2. Assessment of one or more important status variables in relation to trained comparative values: working phase in machine learning, classification of measured values in relation to comparative values of the model, e. g. generalization of new measured enquiry values into the vicinity of the comparative values.

[1] Based on the knowledge gained in recent years and after the addition of important process related to parameter-free design for work flow, a stable procedure for the calculations, the associated mathematical continuity of the manifold in the multidimensional space, the new method, multivariate adaptive embedding (MAE-method), was developed and with it came easy interpretation of results. The MAE-process is patented [8].

The structure of the chain of working steps in the processing unit of the MAE-method is shown in Figure 1 in [11] and [12].

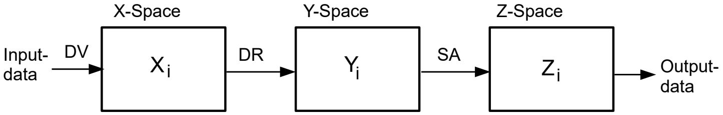


Fig. 1: Structure of the working unit VE: feature space (X-Space) for multidimensional input signals, association space (Y-Space) for links and scalings, depiction space (Z-Space) for the embedment and presentation of the results, data pre-processing (DV), dimensionality reduction, wavelet transform, normalization (DR), scaling and adaption and generalization (SA)

2 Measuring Errors

In machine learning it is possible to avoid some error sources which occur in classical metrology. Still, data measured for practical applications are not error free. The measuring equipment is specified by tolerance. Measuring signals are more or less noise afflicted; the measured values are not immune to interference effects; the measuring instruments have drifts and differences in scaling, and the status to be detected is often not free of superpositions through other physical values. For example, flow-sensor-readings can be slightly affected by pressure and/or temperature. First a few comments with respect to the tolerance of measuring equipment and the required precision for the measuring process will be given. The common understanding for the tolerance and precision of measuring equipment can be summarized by the golden rule of metrology.

The Golden Rule of Metrology: For a measuring result in the acceptable range, one assumes that the measurement result is within the tolerance limits. If the value is very near to a tolerance limit, it may be that the real value exceeds the limit. To avoid making bad decisions following the golden rule of metrology, the measurement uncertainty (u) should be smaller than the tolerance (T) and not greater than one-tenth of the tolerance of the measuring task. The usual value is the relation u/T = 1/10. In exceptional cases u/T can be changed to 7/5. Due to increasing quality and safety awareness, the trend is towards stricter to u/T relations [3].

In the following the main causes responsible for measuring uncertainties are listed. Uncertainties of type **A** (see below) are based on repeated measurements or random samples and can be calculated with statistical methods. Uncertainties of type **B** are only detectable by repeated measurement. To take them into account, all information for possible deviations must be used. This, however includes experience and general know-how related to the measurement task being carried out. One advantage to machine learning is that the measurements are always carried in a uniform manner. In the learning process, the training phase for creating a model and for the characteristics of the data set, many influences can be *calibrated in* or *taught in*. Through this the effects of type **A** can be minimized. Moreover, as mentioned above, monitoring criteria can be added to offer the possibility of switching over to a reserve signal in the event of error. The effort which is necessary is dependent on the safety requirements of the task. These requirements will generally be found via risk analysis or FMEA.

2.1 Error Causes

1. **Subjective measuring errors, type B:** incorrect reading of the measurement scale (misreading), wrong or erroneous entering onto the test-record list (typing errors), entering of inaccurate estimates and measurement values. Subjective measuring errors can be excluded by machine learning.

2. **Objective measuring errors, type B:** objective measuring errors arise when the measuring equipment unintentionally alters the measurement readings through interaction with the measured object (feedback); shortcomings in the measuring system and unfavourable choice of the measuring location or measurement time can lead to feedback effects. The magnitude of this error is dependent on the deviation caused by energy exchange, modification of mass, modification of surface and so on.

3. **Systematic measuring errors, type B:** systematic measuring errors cause imprecise measurement results. They arise, when the measuring arrangement or the measuring equipment is inappropriate. These errors are not detectable by repeated measurement. They can be only be found by use of suitable equipment or methods.

4. **Random or stochastic errors, type A:** they are caused by undetectable changes in measurement equipment and changes in the ambient conditions surrounding the measuring object. Numerical calculations carried out with a computer are not inherently precise. The data format provides the precision. In addition, mathematical rounding causes random errors. Stochastic errors cannot be detected with a single measurement. This applies to the magnitude as well as the sign (+/-) of these values; they are uncorrectable and cause inaccuracies in the result.

5. **Error propagation during the evaluation:** before the measured values are transferred from the measurement location to the processing unit, the values were quantized, but the measurement values were not error free. So, the transformation into digital values is not error free either. In addition, a fuzziness arises, caused by the k-NN-method in the generalization process. The estimation of error propagation can be carried out with an *error calculation*.

6. **Error calculation:** To investigate the error status in each processing step, see Figure 2, one can calculate/record the propagation of the measurement errors through the complete measurement chain with the *error calculation* [7].

2.2 Measurement System with Interference Sources

In modern measuring instruments the measurement signal is transformed into a digital measurement value using an analog-to-digital converter (ADU). Before the signal is processed by the subsequent units, the digital value has to be brought into an adequate data format for the task. Figure 2 shows a measurement arrangement for machine-supported measurement with all components for acquiring the data. The measurement steps prior to the ADU are not free of external influences. The impact is different for each measurement task and has to be taken into account for the measurement uncertainty. After the measuring variable is transformed into a digital value, there is practically no influence to the value. When a digital value is disturbed on the path (S) to the processing unit (VE), it can be repaired by high performance and functional data transmissions (e. g. CAN-bus, Hamming distance (HD) = 6) up to a certain bit-error rate (CAN-bus = 2). When this is not possible, the data transfer can be repeated. After this, the correct measured value is present in the processing unit.

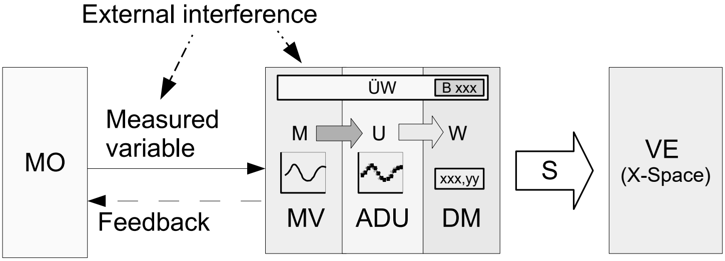


Fig. 2: Measuring system, MO: measuring object, MV: measuring arrangement, ÜF: transfer function, ADU: analog-digital-converter, DM: digital measuring data, W: value xxx,yy in machine readable form, ÜW: monitoring unit for the measuring system, B: readiness code xxx, S: communication path, VE: processing unit for the measured values

The measurement variable of the measuring object *MO*, shown in figure 2, will be detected using arrangement *MV*. The transfer function (ÜF) of the detector delivers the value U. This analog value is transformed by the analog-to-digitally converter *ADU* into a digital measured value *W*. *W* is stored after the transformation into a data logger *DM* and then transferred via the transmission path *S* to the processing unit *VE*. In VE in figure 2, the digital measuring value will be brought in line with other measured values to act as multivariate input the form of an input vector in the feature space, the so called *X-Space* in figure 1.

Monitoring: the whole measuring system will be continuously monitored with the monitoring unit (ÜW), and the result is transferred in sensible time intervals as a readiness code BC to the processing unit (VE). By using the readiness code (BC), different operating states can be transferred. With this it is possible to accept data when the code BC signals readiness. Additionally, it can handle pre-alarms, e. g. *low battery status*. With such messages, the system can be controlled to use values from a reserve system. We speak of multivariate data, when several features are detected. The measured features are, in most cases, independent of one another. Each dimension of the X-Space represents one measuring value. The vector or data point DP represent the multivariate measuring value.

2.3 Interfering Effects

1. Effects: Unintended interaction of the measuring device to the measuring object (feedback). The sensor used has an impact to the measuring object. Example: The sensor of a temperature-measuring device cools down or heats up the measured object after contact. Within such an arrangement a new temperature will result lead to an altered actual reading.

2. Effects: Disturbances overlap the sensor signal (accumulated). Examples: electromagnetic fields disturb the measurement signal, alter the reading (disturbing voltages, currents, field intensities, bad long-term stability all have a offset and gain impact on scaling.)

3. Effects: Disturbances having impact on the characteristic of the transfer function of the measuring unit. Example: DMS-bridge is not assembled tension free manner, drifts caused by aging effects, signal transmission channel is disturbed and so on.

4. Internal effects: Measuring deviations caused by physical impact on the measured object (friction, clearance) or dependency of the measured variable direct at the measuring object from additional effects (insufficient pressure or temperature stability, hysteresis, elastic aftereffect . . .).

5. Noise: Noise signals arise in the sensor and in the amplifier [13]. Noise is a stochastic measuring deviation and is not reproduceable. Even when the measurement is repeated with the highest degree of precision, the measured value will deviate from the true value.

6. Quantization errors: They arise during the transformation of analog values into digital values by using analog-to-digital converters. Different types of converters exist which are designed for different purposes to get fast, precise and reliable results [7].

7. Representation errors: In the course of measurements, the status of objects are detected and recorded/indicated in the form of a value/figure. Firstly, the question arises: how does the value represent the status of the measured object? The measured value of the object is only valid at one location, namely the location where measurement was carried out. At all other locations a deviation can exist, can be possible and the representation error has be taken into account when the measurement uncertainty is determined. Often the knowledge of the the measurment operations is incomplete, and this results in false information, false interpretation and incorrect application or incorrect use [7].

A typical example for the representation error is the measurement of the room temperature.

3 Quantitative Detection of Measuring Errors

Measurement values in a measurement series are interpretable as results of a random sample in the statistic. The probability distribution of this random sample is defined by the parameter of the expected value μ and the standard deviation s. The true value X_w and the expectation value μ match when no measurement deviations, systematic or otherwise, are present. The expected value μ and the standard deviation σ must be calculated out of a measurement series (outlier free). It should be noted, that a minimum of 20 measured values has to be detected per parameter of a model (the exact number has to be calculated with formula 9). The distribution of these values must be suitable for the determination of μ and σ [6]. With the Gaussian probability distribution, measured values are distributed symmetrically around the expected value μ and at $X = \mu$ the function reaches the maximum value.

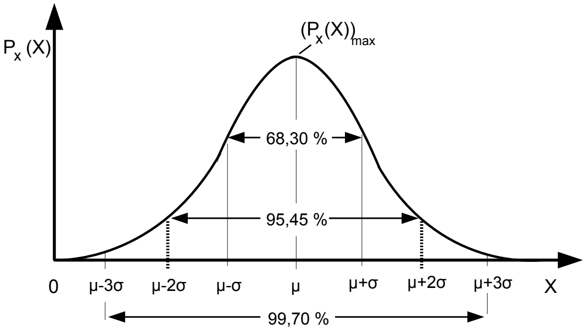


Fig. 3: Normal distribution: Gaussian distribution density function $P_x(X)$, shifted for the expected value μ in relation to zero as an error model for the expected value μ and standard deviation σ

Assumption: the data being used conforms to Gaussian distribution. With this, the data are in position 1 for the single, in position 2 for the dual and in position 3 for the triple standard deviation (3σ) with the values of the listed probabilities in table 1.

Table 1: Confidence intervals and significance levels

Pos.	Standard deviation s	Probability P	z-value λ Error bound $\Delta = \lambda * \sigma$
1	σ	0.683	1
2	1.96σ	0.95	1.96
3	2σ	0.954	2.00
4	2.58σ	0.990	2.58
5	3σ	0.997	3

3.1 Probability and Variance

Normal distribution:

$$P_x(X) = \frac{1}{\sigma\sqrt{2}\pi} \exp\left(-\frac{X^2}{2\sigma_x^2}\right) \tag{1}$$

Probability $P_x(X)_{max}$:

$$P_x(X)_{max} = \frac{1}{\sigma\sqrt{2}\pi} \tag{2}$$

Variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - X_w)^2 \tag{3}$$

Standard deviation:

$$\sigma = \sqrt{\sigma^2} \tag{4}$$

Empirical variance:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_w)^2 \tag{5}$$

Empirical standard deviation:

$$\sigma = \sqrt{\sigma^2} \tag{6}$$

Expected value μ:

$$\mu = \sum_{i=1}^{n}(X_i)/n \tag{7}$$

Half-width FWHM:

$$FWHM = 2\sqrt{2ln2}\sigma \approx 2.3348\sigma \tag{8}$$

The term FWHM ist an abbreviation for *full with half maximum* and describes the characteristic of a function with the width of this distribution at $f_{max/2}$. When the data points of the normal distribution to the left side of μ are known, the variance σ^2 can be determined using formula 8. FHWM is then $2 * (\mu - 3\sigma)$ for a significance level of 3σ (99,7 %). This results in $\sigma \approx FWHM/2,3348$ for the normal distribution. With this, it is possible to estimate the standard deviation, without knowing the upper half of the data. More on this subject in [5] and chapter 4.4 in [12].

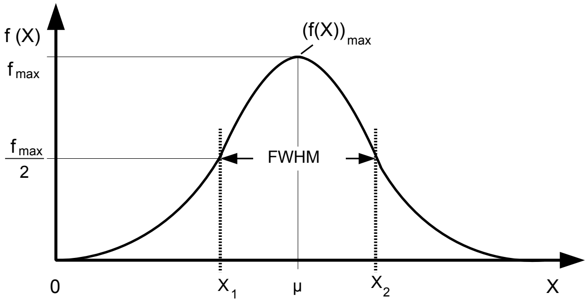


Fig. 4: Half-width FWHM of the normal distribution

The size of the variance σ^2 corresponds with the planar moment of inertia, the distribution density and the standard deviation corresponds with the radius of gyration. An estimation of the standard deviation is given by the measure of dispersion S [7]. With a small number of random samples or measured values, a correction factor, obtained out of the t-distribution, can be used [2]. The higher the desired precision, the greater the number of measured values or the sample size must be.

3.2 Sample Size und Distribution

As described in chapter 2.2 and shown in figure 2, the measurement is carried by machine. During the procedure the measurement equipment is monitored to avoid incorrect measurement data. Thus, only event-triggered *external influences* (EMV-pulse, shocks to the measuring object etc.) remain as isolated disturbances. Influences with a *feedback* to the measured object and changes in the measurement equipment during the measurement procedure will also be present in the working phase and have to be included in the TDS. The measurement itself is affected by statistical variations (additive or intrinsic noise). These variations determine the variance for the data set that was collected to characterize the object status. The curve of the relative frequency for a data point X_i in reference to all other possible data points for such a measurment procedure results in a normal distribution. The measuring of the characteristics of a measured object is similar to a random sample for the determination of characteristic values of the population. For this reason, the statistical rules were applied for the measurement with the MAE-method.

Possible z-values to determine the significance level or confidence limits are listed in table 1.
The following applies for the data set during a measurement:[2]

The central limit theorem from of normal distribution means that the sum of independent, identically-distributed random variables results in a normal distribution, when n approaches infinity. During practical implementation it has been shown or it is common in practice that a random sample of more than 30 stochasticly independent random values, the normal distribution can apply for all intents and purposes. This also means, that the sample size n must be greater than 30 for the creation of a TDS (formula 9).

Sample size of a finite population n:

$$n \geq \frac{Nz^2s^2}{z^2s^2 + (N-1)e^2} \tag{9}$$

Sample size of a infinite population n:

$$n \geq \frac{z^2s^2}{e^2} \tag{10}$$

[2] The determination of the sample size is geared toward the statements in [1].

Table 2: Determination the sample size

Pos.	Steps	Formula	Comments
1.	Required accuracy	X_i: Sample X_{max} X_{min}	The absolute accuracy e, maximum deviation of samples referenced to the true value
2.	Standard deviation s		s not yet known and can be estimated with: $s \approx \frac{X_max - X_min}{3{,}5}$
3.	Confidence limits or error limit	z-value λ, error limit: $\Delta = \lambda * \sigma$	Determination of required significance level or of significance level. For different significance levels z-values are listed in table 1
4.	Resulting sample n, in a finite population	N: number of elements in the population	Formula 9 $n \geq \frac{Nz^2s^2}{z^2s^2+(N-1)e^2}$

4 Measuring Errors with Different Tasks

The application possibilities of machine-learning techniques are many. Classifications monitoring of states and other tasks like pattern recognition can be realized. The effect of measurement errors is similar for all these tasks fields and will be explained in the following using an example of the monitoring of factory activities.

In industrial areas, public facilities, technical equipment or vehicles etc., one is obligated to continuously check the readiness and monitor the condition of equipment. This requires automatic measurement methods and clear presentations of the results in order to be able to quickly and efficiently locate errors. Big machines designed for special purposes and factories or other objects to be monitored are complex and designed in a specific way so that for each object to be monitored a special adaptation is necessary. Furthermore complex factories are difficult to mange even for experts working in the respective field of the factory. Errors and messages must be assessedin order to recognize wear and malfunctions. For this kind of application, there is no standardized solution for problems, which allow the development of an efficient and reliable monitoring system. In one situation information is transferred too late, so that the limits are exceeded. This leads to a belt stop. Hence, a belt stop is initiated at the other end of the spectrum, monitoring is oversensitive, which results in false alarms. This could even erode the confidence in the monitoring system. When setting up monitoring system, it must be started with an all-clear status, which is trained in by machine learning (training phase). This condition will later be compared with the actual-status (working phase). In most cases, the so-called *off line learning* is not sufficient to monitor precisely and reliably. A fine tuning to the system (model M for the trained object) is needed, which follows predefined rules in order to adapt model in according with the rules for training and adaption [9] so as to prevent an overtraining of the model. Thousands of features must be simultaneously monitored, and it must be possible to present the results in an understandable form, for example by using a polar diagram, which indicates a circle when everything is in order. The circle is distorted when a failure occurs. This can be immediately seen by the operators to initialize a fast and efficient location of the indicated errors and identify their causes.

4.1 Limited Online-adjustment

The need to continually adapt and improve the model can be met with an adequate, limited online-adjustability of the training process. This is basically motivated by an application specific principly motivation, to match the fine structure of the model to the monitored object so as to be able to update the model to a limited extent. The manner how this intrinsic motivation can be realized is discussed in in detail in [4]. For this task it is necessary to match the model to each data structure or to its data distribution. The adjustment is linked to the NN area and ensures that the group being monitored is isolated from other groups. The overlapping of group data has to be avoided to allow clear decisions.

5 Monitoring the Measurement Signals

To order to comply with reliability requirements for the results in the depiction space (Z-Space), a monitor for each measuring signal should be implemented to indicate whether the measuring signal is ready for processing or not. The rules for the signaling of readiness must be complied with in the respective application because it is there at the signal source that the special procedures for measuring management have to take place. Realization at a higher level will be ineffective; it has to take place on site at the measuring unit. This makes it possible to deliver standardized information to the feature space (X-space) in the processing unit and in the event of malfunctions of the measuring equipment, reserve signals can be used.

Summary: In this article the technical requirements for the machine measuring of characteristics and influences and typical errors which appear during the acquisition were discussed and explained. The measurement effects and errors were reduced with the arrangement in chapter 2.2 and with the proposed structure in figure 2 of the machined acquisition of measuring except for the effects caused by *feedback* to the measuring object, systematic errors in the sensor (drift in the measuring equipment during measurement), statistical variations (noise) and event triggered *external influences* (EMV-pulse, flash, shocks to the measuring object etc.) to the measured object or measured status. The latter results in outliers in the measured data in the data set, representing a measured object or the status of a measured object. The recognition and the elimination of the outliers is a subject discussed in chapter 4.8 in [12]. The curve of the relative frequency for a data point X_i in reference to all other possible data points over the variability for such a measuring procedure, results in a normal distribution. Feedback to the measuring object and systematic errors of the sensors arise not only during the training phase, but also during the working phase of the automated measuring process. For this reason they have to be accepted within the TDS in order to adapt the receptive area of the working phase of the MAE-method to the conditions of the real-life measurement situations.

References

[1] Gesamtredaktion und fachliche Beratung: Bundesverwaltungsamt, Bundesministerium des Innern, Handbuch für Organisationsuntersuchungen und Personalbedarfsermittlung, Kapitel 5.1.4.4 Ermittlung des Stichprobenumfangs, Herausgeber: Bundesministerium des Innern, Alt Moabit 1, Internet 2017 (`http://www.orghandbuch.de/OHB/DE/node.html`) 40 10557 Berlin

[2] Diehl, Joerg M.; Kohr, Heinz: *Deskriptive Statistik*. 13. Auflage, ISBN 3-88074-110-7, Verlag Dietmar Klotz, Eschborn, 2004

[3] Hernla, Michael.: *Messunsicherheit und Fähigkeit.* Internet 2017 (`http://dr-hernla.de/Hernla%20QZ%201996%20Messunsicherheit.pdf`)

[4] Hofmann, Martin; Neukart, Florian; Bäck, Thomas: Künstliche Intelligenz und Data Science in der Automobilindustrie. In: *Ar-*

tificial Intelligence, Use Cases /by Volkswagen, Germany: Internet(2017) (https://data-science-blog.com/blog/2017/03/23/kunstliche-intelligenz-und-data-science-in-der-automobilindustrie/)

[5] Leo, W. R.: STATISTICS AND THE TREATMENT OF EXPERIMENTAL DATA, The Gaussian or Normal Distribution. In: *Techniques for Nuclear and Particle Physics Experiments*, Internet 2017 (https://ned.ipac.caltech.edu/level5/Leo/Stats2_3.html)

[6] Patzelt, Rupert: *Elektrische Meßtechnik*. Springer-Verlag Wien, New York, 1996

[7] Pfeifer, Tilo; Profos, Paul: *Handbuch der industriellen Messtechnik*. Oldenbourg Industrieverlag, 2008

[8] Halang, Wolfgang; Sartorius, Gerhard; Talbot, Steven: *Verfahren und Vorrichtung zur Detektion von Kampfstoffen*. Patent Nr. 10 2008 054 345, München 2013

[9] Sartorius, Gerhard: *Multivariate Adaption mit künstlichen neuronalen Netzen*. VDI-Verlag, 2009

[10] Sartorius, Gerhard; Talbot, Steven, R.: *Multivariate adaption and classification method for characteristic sample properties*. Engineering Life Science 2011, No. 5, 1–5, WILEY-VCH Verlag GmbH & Co. KCaA, Weinheim, 2011

[11] Sartorius, Gerhard: Multivariate Adaptive Embedding, MAE-Process. In: *Autonomous Systems: Developments and Trends, studies in Computational Intelligence*. Springer-Verlag Berlin Heidelberg, 2011

[12] Sartorius, Gerhard: Data Pre Processing and Outlier Detection in Multivariate Data. In: *Autonomous Systems 2017, Proceedings of the 10th Conference*, VDI-Verlag GmbH, 2017

[13] Wupper, Horst; Niemeyer, Ulf: *Elektronische Schaltungen 1: Grundlagen, Analyse, Aufbau (Springer-Lehrbuch)*. 2. Aufl., ISBN-13: 978-3540606246, Heidelberg: Springer Verlag, 1996

Used Abbreviations

ADU	- Analog to digital converter
DR	- Dimensionality reduction
DP(s)	- Data point (s)
DS	- Data set
DV	- Data pre-processing
FMEA	- Failure Mode and Effects Analysis
KNN	- Artificial neural net
M	- Model, contains the Characteristics of the TDS
MAE	- Multivariate adaption und embedding
Dist2NN	- Distance to the nearest Neighbor (NN)
MaxDist2NN	- Maximum distance to the NN
DiffDPx	- Difference to data point DPx, x = 1...n
MD	- Manhattan-Distance
MF	- Manifold
NN	- Nearest neighbors, nearest neighbor
k-NN	- Number of k nearest neighbors
SA	- Scaling and adaption
SD	- Spectral data input
Skala-i	- Scale-i for the TPs in X-, Y-, Z-space (abscissae)
TDS	- Training data set
TP(s)	- Training point(s)
UGR	- Lower bound
WT	- Wavelet transformation
z-value	- Error bound, significance level), fixes UGR and OGR

Table of Symbols

μ	- Expected value
σ	- Standard deviation of normal distribution
s	- Standard deviation of random samples
σ^2	- Variance of normal distribution
s^2	- Variance of random samples
X-Space	- Feature space for input vectors
i	- Index for training vectors $\vec{x}_i$ in $\underline{X}$
m	- Index for dimensions in X-, Y-, Z-space
N	- Number N of training vectors, in general number of ...
$\underline{X}$	- NxD-matrix of training data points in X-Space)
X_i	- Training data point in $\underline{X}$, in feature space
$\vec{x}_i$	- Training vector of matrix $\underline{X}$
$\vec{x}_a$	- Query vector in X-Space
D	- Vector component in D-dimensional space dimension related variables X_i^m, Y_1^m, Z_i^m, F_i^m ...
$\vec{x_{iD}}$	- Vector component in general
m	- Vector component in D-dimensional space
$\vec{X}_i^m$	- Vector component, training vector of matrix $\underline{X}$
$\vec{x}_a{}^m$	- Query vector in component m in D-dimensional X-Space
X_a	- Query data point in X-Space
$\vec{\eta}_j$	- Nearest-Neighbor-vector of $\vec{x}_i$ in X-Space
K	- Number of NN
j	- Index for the NN X_j
ϵ	- Reconstruction error
Y-Space	- Association space
$\underline{V}$	- Transformation matrix
Z-Space	- Presentation space, space for an allocated X_a
$\underline{Z}$	- Nxd-matrix of Training points in Z-Space)
d	- Dimension of the Z-pace
f_g	- Cutoff frequency
O	- O-Notation, Compare of algorithms without implementation details of program language and hardware
X_w	- True value

Data Preprocessing and Outlier Detection in Multivariate Data

Gerhard Sartorius

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: The n-dimensional input structure in an artificial neural network for the classification of objects consists of different signal sources, such as signals from temperature sensors, position encoders, microphones, cameras, spectrometers or data from other measurement instruments. The raw data, delivered from these measurement data sources do not have the same structure; rather, they have a different number of dimensions and they can be presented in a parameterized form as linked values, as data packets or as discrete values at the input. The data distribution of these values will be analyzed in the data pre-preparation procedures; the statistical values were identified, non-applicable values will be eliminated and the characteristic values will be saved in a model for the training data set (TDS) for the underlying object. In order to match these values to the input structure of the artificial network, different methods were used. Sensor signals were quantized, filtered and standardized. Signal trends presented in a parameterized manner can be transformed via wavelet transform into spectral components, filtered and standardized. To create the input vectors, single measuring values, packets of spectral measuring data and other data sources of information can be combined in order to get reliable characteristic properties for the trained object. To obtain a pre-defined exactness for the mapping to a low dimensional space, the training vectors should have a limited maximum distance to their nearest neighbors (NN). This and a continuous manifold without gaps will be assured by the data pre-preparation procedure so that the data can be processed as an entirety in the following work steps.

1 Introduction

This paper uses definitions and knowledge of metrology and machine learning introduced in *Metrology and Machine Learning* [15] by this author to prepare incoming multivariate data sets to the requirements of machined processing for classification purposes.

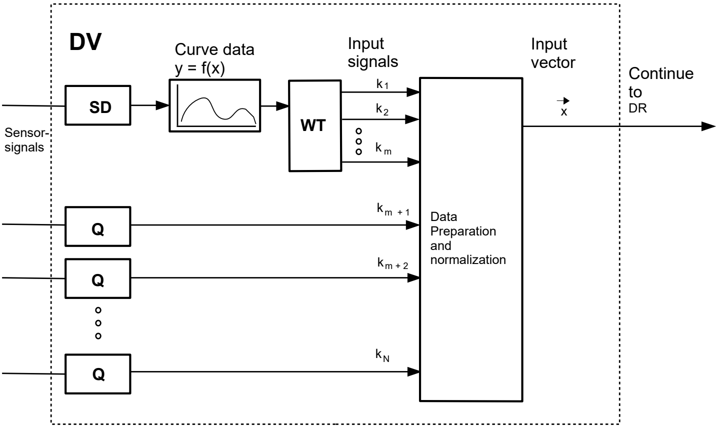


Fig. 1: Detailed characteristics of DAV module in Figure 2

In practice, sensor signals are present as signal values which are limited in frequency to f_g and adapted to the measurement range of an analog-to-digital converter, and which must be sampled with a frequency of at least $2 \cdot f_g$, in order to be digitally processed[1].

It is intended that quantized signals be made directly available as input signals and continuously compiled measurement curves, e. g. spectra, after a wavelet transform (WT) in the form of spectral components, at the end of the DV, as a vector, which is standardized at the entrance to the connectionist system. The quality of data saving and/or the access accuracy in the multidimensional space depend on the kind of distance function used[2]. The use of a standard does not change anything about the position of the DPs in the multidimensional space; it only provides a scale for the distance measures. In order to exploit the advantages of the WT, the Euclidean distance has to be used in the vicinity of the corresponding data. An orthonormal WT has the effect of maintaining distance

[1] A band limited-signal limited by f_g is completely described by individual signal values which are extracted at intervals of $T = \frac{1}{2f_g}$ Shannon Sampling Theory [10, S. 82].

[2] Various metrics can be used as distance functions, e. g. the L^1-Norm (Manhattan distance) or the L^2-Norm (Euclidean distance) [4, S. 602].

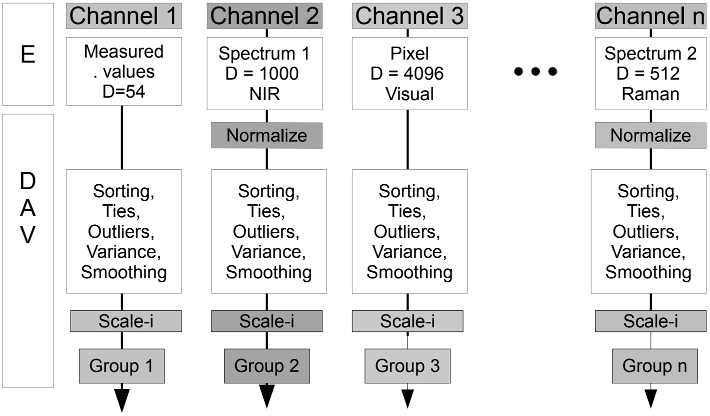


Fig. 2: Organization of the input data with measurement channels

if the scaler product is used for standardization [7]. After determining the coefficients a different standard can be used. Maintenance of distance is important for the formation of the input vector of the network so that wavelet-transformed input data and input signals can occur mixed together in the input vector as is depicted in Figure 1. The figure shows the structure of the data pre-preparation (DV). The spectral data input module (SD) records measurement points, such as the data of a spectrometer, which could originate from the measurement of special substances, gasses, noises etc. When dealing with such measurements, it is probable that neighboring measurement values are similar. The subsequent wavelet transform (WT) provides the wavelet coefficients $k_1, ..., k_m$ of these signals. The resolution, or more precisely the required degree of differentiation, determines the minimal number of coefficients in the spectral range. The quantization (Q) converts analog signals into discrete values $k_{m+1}, ..., k_{m+N}$, whereby N indicates the total number of measurement values of the input-value bundle which determines the dimensionality that is required for this measurement channel. The signals $k_1, ..., k_N$ are prepared in the *Data pre-preparation and Standardization* module in Figure 1, so that they can be fed to the the next processing stage for dimension reduction (see also Figure 2 in [15]).

The characteristic of a measurement object can be recorded in various ways. The way of recording is assigned to a channel and a model (M), in which all essential parameters and statistical characteristics of the measurement values of each channel are stored. Figure 2 depicts this organization. The distribution of the

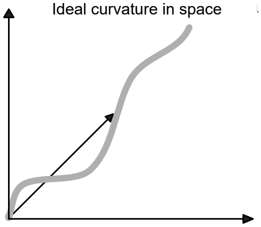


Fig. 3: Smooth structure in a low dimensional space

data depends on the influences listed in chapter 2.3 in [15]. If random noise is the only disturbance present, then the data are normally distributed (Gauss). In the event of other influences, the distribution can more or less deviate from the normal distribution. As a rule, however, the measurement points of a measurement source, when conducting practical measurements, are distributed around an expected value. However, the following is important when cleansing the data: real data such as measured values are often incomplete and afflicted with flaws such as outliers. One outlier is enough to lead to false results in the next processing stage. As such, the data set would be unusable as a TDS. When cleansing, the following points are important:

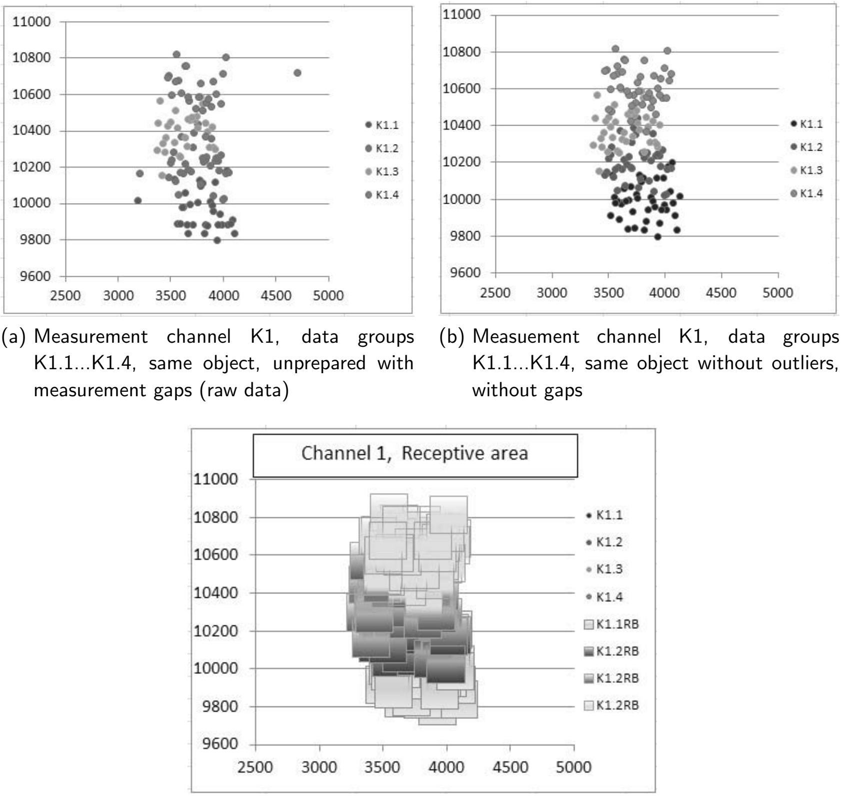
1. If possible, no new information is to be added.
2. Inserted values are to be information neutral in order to not skew or falsify the distribution of the measured values on hand. The eigenvalues of the manifold are not to be changed or changed as little as possible.
3. With the MAE-method there is a described way of dealing with missing values in section 4.7 and in [14].

1.1 Generalization about an Ensemble of Expected Values

When searching for association, however, not only are the individual measured values around an expected value stored, but often also many similar values for similar situations, which, as an ensemble of expected values, make up a trail of juxtaposed expected values in space, as is depicted in Figure 3. For example, when a face is rotated about an angle, this has the effect of producing such a construction in the multidimensional space. An ensemble of measurement values for a certain measurement range of the mass-flow measurement in the following example serves as another instance of this. The data distribution around each expected value for each face position or for a measurement situation is labeled as Kn.1 ... Kn.m. The channel number is labeled with n, and the number of the distribution of the measurement situations is labeled with m. A prerequisite for later recognition of association with objects is the inclusion of the multivariate measurement data for all occurring measurement systems and assignment of these training data points to the corresponding display values in the result or depiction space during the training phase of the network. The data distribution of the individual measurement situations must be free of outliers. With respect to this, they are to be checked and prepared using the methods in chapter 4 so that an allocation of these data in the association space (Y-Space) is possible after the training phase. When doing this, the associated model M for this measurement channel and the characteristic of the respective object is created in the depiction space (Z-Space). The measurement values presented in the working phase of the network can then generally be displayed as results. The MAE method is based upon this principle.

Organization of Measurement Data: With Figures 4a to 4c, a basic overview of the data pre-preparation with a small, two-dimensional data set has been given in order to provide an overview of the measures and processing steps for determining the statistical characteristics of the measurement data with respect to the organization of the MAE-method that will be described below. The model of an object can become randomly complex and represent various characteristics. The complexity of the model is optimized in connection with the wavelet transform and adjusted in such a way that the generalization ability is ensured. Furthermore, an object can be identified through the use of various measurement channels. There can be diverse training sets for one and the same object. For example, certain properties of a product or a raw material can be dependent on boundry conditions which shift the characteristic measured values, which are, however, used for identification. Such combinations and other combinations are possible. They were processed as an ensemble of expected values, as shown in Figure 4 and Figure 18.

Pre-preparation of Measurement Data: In Figure 4a such a case for 4 measurement events which apply to the same object is depicted. Here, there are twenty measurement values each, which were measured at different times, at different locations or under differing conditions, which represent the measurement object and have not been prepared. Outliers are present and the data sets contain gaps. These conditions reoccur in an identical manner in the working phase. These data thus stand in their entirety for the characteristic and the boundry conditions of the object to be classified. Each data group K1.1 ... K1.4 which was measured in the same way for the same object is to be prepared for the require-



(a) Measurement channel K1, data groups K1.1...K1.4, same object, unprepared with measurement gaps (raw data)

(b) Measuement channel K1, data groups K1.1...K1.4, same object without outliers, without gaps

(c) Measurement channel K1, data groups K1.1...K1.4, same object with receptive range

Fig. 4: Outlook: procedures result in a receptive range related to the training data set (TDS)

ments of the MAE-method using the operating steps described in this chapter. In Figure 4b data prepared in this manner are presented. After the characteristic values of the distributions of the sets K1.1 ... K1.4 are known, the receptive range can be indicated. As a result of the data pre-preparation, the receptive range for the training data set (TDS), which is defined by the characteristic values of each measurement event K1.1 ;...; K1.4, depicted in Figure 4c are present for the working phase. All necessary parameters and data for an object are stored in the model M of the TDS. After generalization in the depiction space (Z-Space), it becomes recognizable to which extent the measured values tend towards K1.1,

Table 1: Example for operating conditions and tolerances

Dependency	Tolerance-specification of the manufacturer	Compen-sation, correction-calculation	Empiri-cal-value	Commentary
X =f (flow)	yes			Flow range manufacturer's specifications
X = f(temperature)		yes		Effect of ambient temperature manufact-urer's specifications
X = f(oscillation frequency)		yes		Density of the medium in the measuring tube changes the oscillation frequency
X = f (age of the measurement device)			yes	Replacement after expiration of maximum service life
X = f (counter pressure at the output	yes	yes	yes	Check operating conditions and manufacturer information

K1.2, K1.3 or K1.4 through the position of the depicted enquiry points X_a relative to the training-data points.

1.2 Determination of measurement uncertainty

In the case of individual measurement values, the measurement uncertainty can be determined using the key figures of the measurement equipment, while taking the factors listed in chapter 2.3 in [15] in connection with the error calculation in chapter 2.1 in [15] into consideration [11, 12].

Example Mass Flowmeter: The measured value here is dependent on:

1. the torsion which is generated by the flow in the oscillating measuring tube. The measured bending of the measurement tube is the measurand;

2. the aging or fatigue of the material that the measuring tube is made of;

3. the operating temperature which the modulus of elasticity of the material is subjected to;

4. the pressure at the outlet end of the measuring tube, which also has an effect on the modulus of elasticity and thus on the torsion.

As this example shows, there are three independent measurands (temperature, pressure, torsion) and one dependent measurand (flowrate, measured by a sensor), which are to be determined from the four measurands.

Cause-effect Lists and Cross Tables: The boundry conditions listed in the table 1 for the operation of the mass flowmeters used as an example show that a priori information can be used as well in order to indicate tolerances with regard to the individual measurements. This is nothing new but rather a common practice for indicating measuring accuracy for different measuring purposes. Also in the case of machine learning with multivariate data, which often encompasses a large number of dimensions in oder to include all of the measurement values for the purpose of the data collection, an indication of the reliability of the results is required or extremely important during the training phase and the generalization in the working phase. After all, the exactness, or measurement uncertainty, of the final results in the result or depiction space (Z-Space) should be known. As a rule, characteristic curves and characteristic diagrams are used for indicating tolerances in the case of flowmeters. These data can be presented in a clear form by using cause-effect lists or cross tables so that the the effects of the tolerances on the final result can be readily recognizable. For instance, with a cylindrical piston meter, the tolerance characteristic curve is in the negative range for small flow rates and in the positive range for large flow rates. The next table shows the dependent and independent variables which specify the use of a measuring device, using a mass flow meter as an example.

2 One-dimensional Data Sets

One-dimensional or univariate data sets are examined using the methods of descriptive statistics. The following identifies essential information that is relevant for machine learning.

Excess Kurtosis: The excess kurtosis indicates the the curvature of the distribution curve and describes whether the distribution is pointed or flat in comparison to the normal distribution.

Skewness: Skewness indicates whether or not the distribution is symmetric. Positive skewness (left steep, right skewed): there are many small values and fewer large values in the data. Negative skewness (left skewed, right steep): there are many large values und fewer small values in the data.

Normal Distribution: Excess kurtosis = 0, Skewness = 0. The larger the excess kurtosis or skewness, the less a distribution corresponds to the normal distribution. With a normal distribution the data are symmetrically arranged around μ, as in Figure 3 in [15]. With machine learning data distribution corresponds to the normal distribution in very many cases, as mainly only noise effects from the error sources listed in chapter 2 in [15]. In order for the processing of the data to lead to reliable results, the following factors can be relevant:

1. **What kind of distribution applies?**
2. **Are there outliers?**
3. **Are the measured values independent or dependent?**
4. **Are we dealing with a sampling measurement (e.g. for quality assurance)?**
5. **What is the variance of the data being processed?**
6. **What is the level of statistical significance?**

To 1: Using a test procedure, it is determined whether or not deviations are random and thus correspond to the normal distribution or a different distribution, a t-, F-or χ^2-distribution. For each kind of distribution, there are corresponding tests. When the measured values are symmetrically distributed around a mean value μ as in Figure 3 in [15], then a Gaussian or normal distribution is present. If the measured values are not symmetrically distributed around the expected value μ , then certain tests must be used to determine the kind of distribution. In order to check if a normal distribution is present, the following tests, among others, can be used: D'Agostino-Test, Lillifors-Test, Shapiro-Wilk-Test.

To 2: Various outlier tests for one-dimensional data sets: Grubbs, Nalimov, David, Hartley and Pearson, Dixon, Hampel, Barda, Pope, Walsh.

To 3: Moreover, it is usually determined whether the data are independent or connected with other measured values. In the case of independent data, there is no relation to other data identified in the data set. Dependent data are present if there is an internal connection for the change in several measurement characteristics.

Chi-square-test: It serves to test the independence of the variables being examined in connection with all variation possibilities of the inputs and thus the analysis of the statistical significance of a variable. Relationships to other distributions than the Gaussian distribution can be found in technical literature on statistics [6]. Such dependencies occur frequently within the framework of technical measurements. This is, however, hardly relevant for the MAE method as multivariate measurands are being processed anyway, and only the internal relationships between the measurands are used for the generalization of the results in the depiction space (Z-Space).

To 4: The entirety of a statistic examination consists of measurements taken under the same conditions. The characteristic corresponds to a certain distribution in the population, e.g. the normal distribution. A random sample is a subset with n elements , which is formed through random selection from within the population. In a random selection each element has the same probability of being included in the selection. Through the use of random sampling, the statistical characteristics of a distribution can be estimated. They minimize the measurement work needed and thus save costs [8].

To 5: When no more outliers are present the variance can be calculated using the formula 5 in [15].

To 6: This depends on how high the demands for exactness are and how many data are available. For the standard deviation or the variance, at least two values are needed (as division is by N-1) for the computation.

An Example: The mean value (MV) is 150, and the standard deviation is 30. It is to be ensured to a certainty of 95 % that the MV of the random sample does not deviate by more than 10 from the MV of the population. The *confidence interval* would then be 140 ... 160, and thus 20 wide. The number would then have to be: $N \geq (\frac{2*1{,}96*30}{20})^2$. This results in $N \geq 35$. Here, the following holds true $1\sigma \rightarrow 1{,}64 \rightarrow 68{,}3\%$, $2\sigma \rightarrow 1{,}96 \rightarrow 95\%$ and $3\sigma \rightarrow 2{,}58 \rightarrow 99\%$. See Figure 3 and table 1 in [15]. At 99 % it would be 2.58 and at 95 % 1.64. This 95 %-confidence interval means that from 100 possible random samples with a size of 35 elements taken from a population with a mean value of 150 and a standard deviation s of 30, then 95 random samples should be present with a mean value between 140 and 160. Five random samples would have a mean value outside of the confidence interval. For a higher degree of certainty (99 %), a larger random sample with $N \geq (\frac{2*2{,}58*30}{20})^2$ would be required. This results in $N \geq 60$. For a smaller random sample $N \geq (\frac{2*1{,}64*30}{20})^2$ means a lower degree of certainty with a random sample size of 25.

Variance of the Data: First of all, the variance is the measured value of interest so that the level of significance can be established. For every measured-value data set, the variance must be determined. With the MAE method the variance of the measured data can be determined using the differences in distances between the the individual data points. The arrangement of the measured values and their sequence to a defined fixed point, to which the distances in the multidimensional space can be related, is depicted in a one-dimensional manner on Scale-i in Figure 20. This is how a multidimensional case is reconverted to a one-dimensional structure within the MAE method. The variance of the data and the assessment of the errors is geared solely toward the difference in distances between the the data points. These differences represent, on average, the variance of measured values through all the dimensions. They are used in chapter 4 for data cleansing. This method thus allows a single-factor variance analysis with formula 5 in [15]. Because automated measurement and classification are based on an assumption of a data set that represents the actual relationships with all their deficiencies that the respective measurement delivers, and hence only the distribution of the difference values between the data points from DP to DP have to be checked, a Gaussian distribution almost always applies in practice. The data are arranged in a normal distribution around the centroid of these measured values. How the centroid can be determined is explained in chapter 1.2.

3 Multidimensional Data Sets

The measured values to be processed with the MAE method which are being discussed here are multivariate data sets. A correct statistical evaluation of the measured data is equally decisive for the reliability of the results in the result or depiction space (Z-Space). In order that no errors can occur, the statistical characteristics must be determined using the same rules as for one-dimensional data sets. Due to the multivariate, high-dimensional structure of the data. it is simply possible to determine the variance and hence the significance level of these measured values in oder to exclude double or nearly identical values (ties) and outliers. It is unclear in which dimension the largest proportion of the variance is to be found. There are various methods for estimating the variance of multidimensional distributions: table 2 in [16], [2] shows five iterative methods. These, however, are very specially designed and unsuitable for the MAE-method, due to their iterative approaches.

Reducing to a Univariate Distribution: With the MAE method a different approach is used: with this method, in order to effectively give consideration to all the variances of the multivariate data set, the difference to NN (Dist2NN) is used as a measure for determining dispersion, and not the measured value in each dimension. The multidimensional structure is reduced to a univariate distribution through the measurements of the distances from DP to DP. Furthermore, the Manhattan distance (L1-norm) is used for the determination of the distances from DP to DP. This ensures that no inaccuracies arise in the determination of distances when larger dimensions (¿ 100) are present. The corresponding relationships are described in [3][3].

Table 2: Various methods for estimating variances

Criterion	ML	GLS	ULS	SLS	ADF
Assumption of Multivariate normal distribution	yes	yes	no	no	no
Invariant from level to level	yes yes	yes yes	no no	yes yes	yes yes
Size of random sample	>100 >100	>100 >100	>100 >100	>100 >100	1.5p(p+1) 1.5p(p+1)
Inference- (χ^2) distribution	yes	yes	no	no	yes

Effects of Errors in Machine Learning: As described in chapter 2.2 in [15], the boundry conditions during the measurements are constant apart from effects prior to the quantization (external interference, Figure 2 in [15]). When forming the training or reference data sets, it must be ensured that the measured values are free of effects of the latter. Effects that are caused by the measurement object itself must be included in the training set as they belong to the characteristics of the measurement object and also occur in the working phase, after the training phase. This kind of usage can also be used for generalization of an ensemble of expected values described in chapter 1.1 and in Figure 4.

Required Structure of the Manifold (MF): The data sets to be processes within the course of the MAE-method must be continuous and without gaps. The characteristic of the MF in the multidimensional space can have any random form.

[3] Resulting error of distance as a function of the dimensionality for different metrics.

The outlier tests based on the measured distances from data point to data point or short from DP to DP can be used [9], [14], [3]. The MF required for further processing must be numerically stable, smooth and without gaps or jumps (discontinuities). Otherwise, there are no restrictions with regard to the point arrangement of the MF. How these data can be made available, will now be described in what follows.

Ties and Outliers: If multiple dimensions are present, then the situation with respect to the outlier tests is more complicated than with univariate distributions. In [1] it is suggested that each multivariate observation be represented by a curve diagram (the so-called Andrews curves) in order to visually identify outliers [1]. For machine learning this method is too uncertain and therefore unsuitable. In [9] a method is suggested that works with the distances to the k-NN of the DPs in the data set. In [14] a procedure is explained that works with the NN range of the k-NN of the data sets[4]. There it is discussed how many outliers a data set can sustain and which measures must be taken to ensure error-free data sets. The last method mentioned is used for the MAE process. Double or nearly identical values do not contribute any information, and outliers falsify the the training set and the expected value[5]. These values must be eliminated from the data structure of the given measured-value set. One possible approach would be to reject all values whose difference to the NN is smaller than the lower significance level (ties) as well as all values whose difference exceed the upper significance level (outliers). Only values within the confidence range in Figure 3 in [15] are to be permitted as training points[6].

3.1 Manhattan Distance

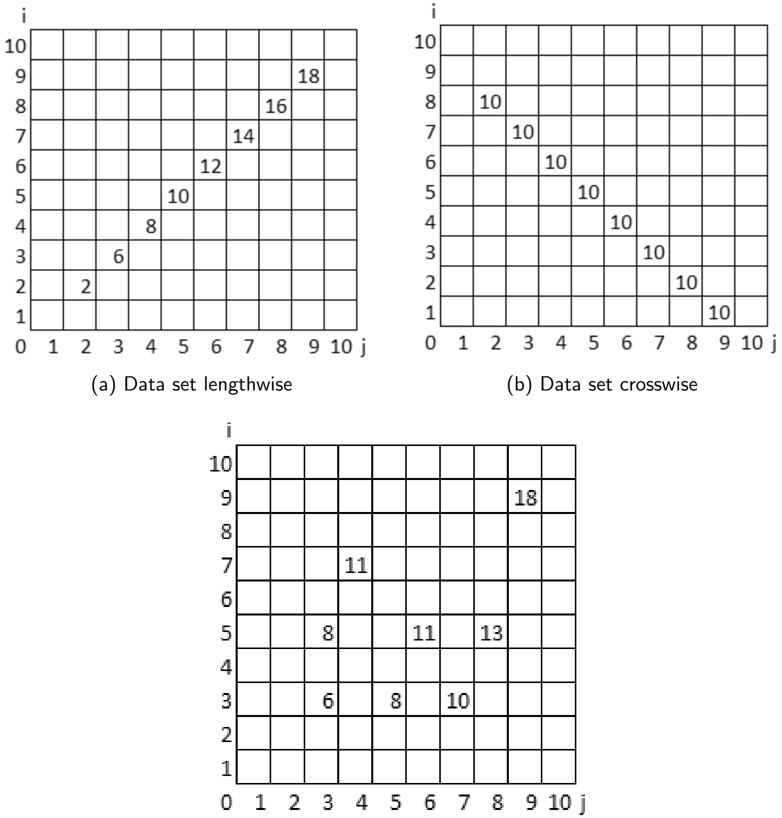
The advantages of determining distances using the Manhattan distance lie in the exactness when using high-dimensional data sets. The relationships are described in [3]. In Figures 5a and 5b, the values of the Manhattan distances from the origin are entered into the position boxes of the corresponding DPs. The diagrams used for explaining the work steps for processing the data described below are two dimensional, each with an abscissa axis j and an ordinate axis i, so as to be able to depict the principle processes in a clear and easily-understandable

[4] The nearest neighbor range of k nearest neighbors (k-NN) of the training data set (TDS).

[5] An overview to the fatal impacts of outliers on the calculation results are given in [5]. There a so called breaking point tells us, wheather the number of outliers in a population is acceptable and which steps have to be taken to ensure outlier-free data sets.

[6] For calculation of the variance of a random sample or calculations of parts of the population s is used instead of σ.

manner. First, the prerequisites and conditions are specified in order to obtain a data set free of outliers for forming the TDS or MF, and subsequently the processing steps for this will be presented.



(a) Data set lengthwise

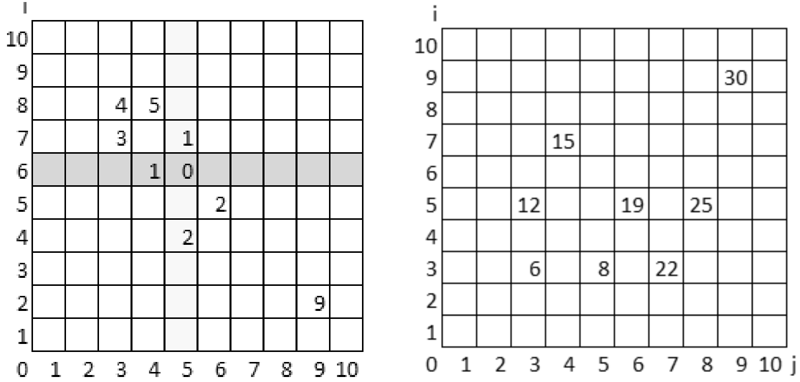
(b) Data set crosswise

(c) Data set round

Fig. 5: Different DP-arrangements in space

Identifiability and Distance Measures: When determining distance, the position of the DPs to the origin is to be taken into consideration. The DP arrangement running diagonally lengthwise to the origin in Figure 5a has the effect of generating continuously increasing distance values. In comparison, the DP arrangement running crosswise to the origin in Figure 5b produces the same distance values. Hence, such points are indistinguishable using distances and cannot be ordered on a list. Also in Figure 5c this case occurs for the Manhattan distance (MD) eight and eleven.

Determining the Distance Relative to the Center of the Data Set (TDS): Due to the identical values, which cannot provide any meaningful information with respect to outlier recognition, the center of the point set , the median, shall be used as a reference point for distance measurement for the outlier recognition procedure described in the following. The outlier MD = 9 in Figure 7c has a large distance to the the center of the data set, the position of the multidimensional median. This is the intersection of the blue/grey and yellow/light-grey lines in this figure. The determination of this multidimensional median is described in chapter 4.2.



(a) Manhatan distance relative to the median

(b) Manhattan distance to the previous point

Fig. 6: Distances relative to the centre of the data set

Arangement in Ascending Relation to the Origin: For the arrangement of the data relative to the origin, which is also described below in section 4, the first step is to find the point with the smallest distance to the origin. Starting from this position, the distance to NN is respectively determined and added to the last distance determined. Using this approach, no identical distances occur and sortability is guaranteed. The corresponding results are shown in Figure 6b. The results without using this measure can be seen in Figure 5c.

3.2 Arranging and Smoothing the Data of the Manifold (MF)

The measured values, bundled together as data groups without outliers, are initially entered into the table in an unordered manner. In the first step a fixed point (e.g. the origin) in the multidimensional space is specified so that the data being used can be entered into the table, from top to bottom, in a rank order, from minimum to maximum distance to this fixed point. After this, the variance of this *Dist2NN*, using formula 5 in [15] and the standard deviation, using formula 4 in [15], are calculated. After ordering, the transitions from DP to DP and for each DP are clearly defined, and the fill-value method described in [14] can be applied. With this method it is ensured that there are no gaps or jumps in the MF which could decrease accuracy during the dimension reduction and generalization in the depiction space (Z-Space). Due to the rank order of the data, the arrangement of which runs from minimum to maximum distance to a fixed point in the space, the smoothing can be carried out[7]. With this approach, the arrangement of the DPs in ascending order of distance requires only a minimal number of fill values. The eigenvalues are not altered by the arrangement of the data since the position of the DPs in the multidimensional space remain unchanged. After smoothing, the MF is continuous and differentiable. The individual DPs are depicted on a one-dimensional scale, the Scale-i, referenced 1st point 2nd point, as in Figure 20.

4 Establishing the Outlier Threshold

In measurement technology the normal distribution is assumed to apply to the dispersion of the measurement errors. Random sample and population are terms from statistics. Ideally, a random sample contains all the information needed for determining the characteristics of the population. With a random sample, the characteristics of the distribution of the population can thus be estimated. In order to suitably formulate the prerequisites for the estimation of

[7] The cubic splines, used for smoothing the MF in [14], are much more suitable for keeping the topology as polynomials.

the characteristics, the following must be entered as a priori knowledge when determining the characteristics:

1. The dispersion of the measurement errors for a certain measurement event corresponds to a Gaussian distribution.

2. The shifting of the data points around a constant value or the formation of the sum or the difference of normally-distributed data points does not change the structure of the distribution.

3. A listing of the data in ascending order of the distances does not change the distribution of the data.

4. Conscious selection of measured values for determining the standard deviation using formula 8 in [15]: distance values from zero to the median of the population between 0 % and 50 % of the list of distances arranged in ascending order forms the set of these measured values. Outliers lie above 50 % and hence are not part of the sample.

5. The volume of the measured values is geared toward the accuracy demands or the specification of the level of significance.

The variance and expected value of the multidimensional distribution of of each measurement event are needed in order to determine the receptive range of the object to be classified. For determining the characteristics, for determining the receptive range, for a measurement channel and for the arrangement of the one-dimensional Scale-i, the following steps are required:

4.1 Outlier Test

1. Determination of the median 4.2

2. Values below the multidimensional median (lower 50 % of the sorted list, half-width) for determining the variance using formula 8 in [15]

3. Determine the outlier threshold value and remove outliers

4. Determine characteristics μ and σ of the TDS

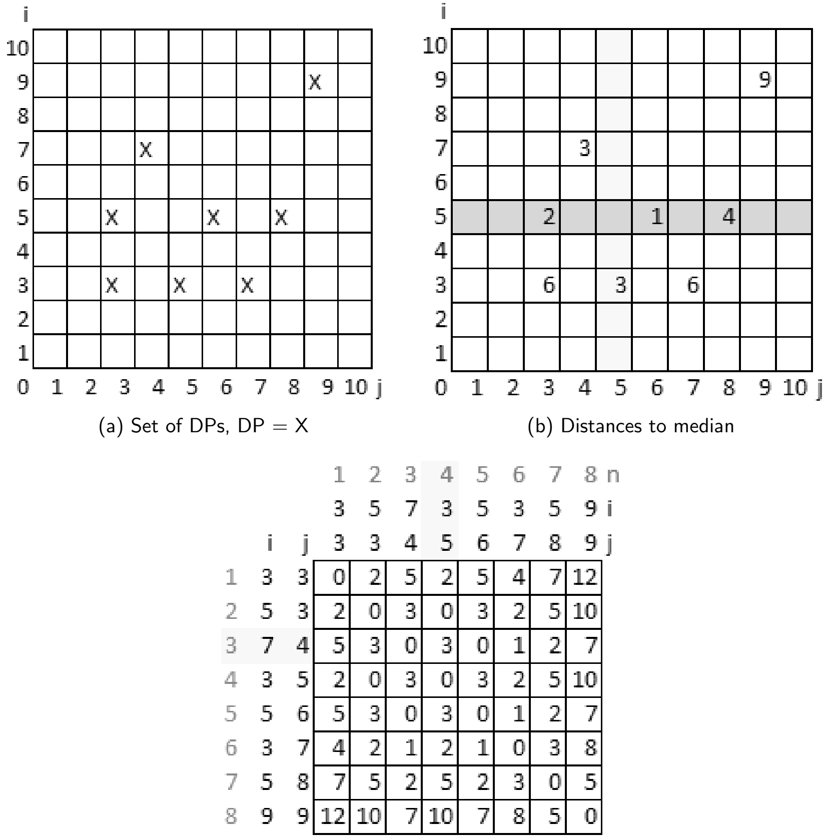
5. Remove double or nearly identical values (ties)

6. determine maximum distance between the data points

7. Fill in gaps and jumps in the list

8. Define the receptive range

9. Define Scale-i for the group or for the kind of measurement of the object



(a) Set of DPs, DP = X

(b) Distances to median

			1	2	3	4	5	6	7	8	n
			3	5	7	3	5	3	5	9	i
	i	j	3	3	4	5	6	7	8	9	j
1	3	3	0	2	5	2	5	4	7	12	
2	5	3	2	0	3	0	3	2	5	10	
3	7	4	5	3	0	3	0	1	2	7	
4	3	5	2	0	3	0	3	2	5	10	
5	5	6	5	3	0	3	0	1	2	7	
6	3	7	4	2	1	2	1	0	3	8	
7	5	8	7	5	2	5	2	3	0	5	
8	9	9	12	10	7	10	7	8	5	0	

(c) Hodges-Lehmann distance matrix

Fig. 7: Hodges-Lehmann matrix for the set of DPs

In order for the characteristics of the distribution and the parameters for the model M of the data of a channel to be able to be correctly determined, the reference data to be used in the training phase must not have any outliers. Robust estimation methods are independent of the values of the data so that the burden on the estimator caused by outliers in the data remains small. The usual calculation of the expected value via formation of mean values becomes unusable as soon as a measured value is grossly incorrect. An overview of robust estimators and burdening of the estimators by outliers is provided by [5]. The median is insensitive to large outliers. When employing the MAE-method, the **Hodges-Lehmann estimator** is used for determining the median[8]. As the multidimensional structure has been reduced to a univariate distribution through the measurement of the distances from DP to DP, this estimator can be used. The Manhatten distance is used for distance measuring.

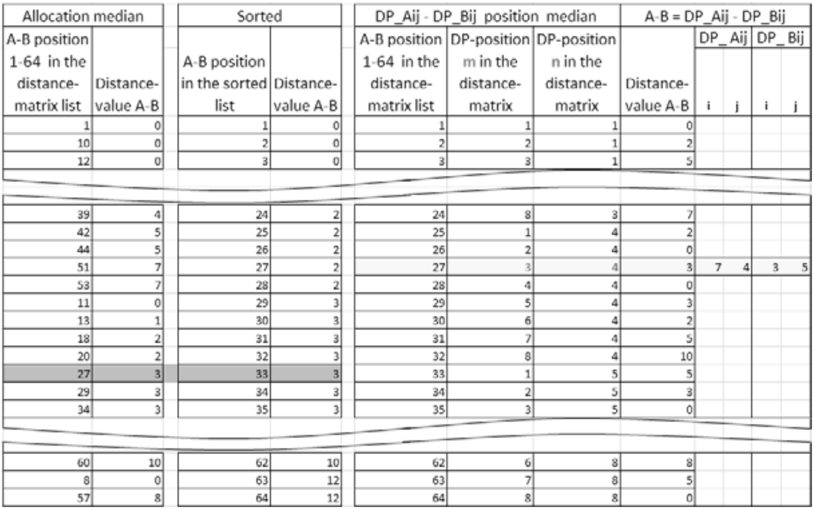
4.2 Determination of the Median

1. Starting point: two dimensional list with 8 DPs in Figure 7a. Distance measurement is carried out from DP to DP with the Manhattan distance. The position of the DPs are marked there with an X.
2. Note the distance from each DP to each other DP in a distance matrix (display of the principle for 8 DPs in Figure 7c).
3. Arrange the distances in ascending order.
4. With n entries on the list, the median $= n/2$.

Notes to the Figures: Figure 7a shows a two-dimensional coordinate system with the origin equal to 0 and 10 units for the first dimension (i) and the second dimension (j). The positions of the DPs are marked with an X. The distance matrix depicted in Figure 7c contains the Manhattan distances (A-B) of each DP (A) to each other DP (B) in Figure 7a. The tuples of the DPs are to the left of and above the distance matrix. These tuples are labeled with l (blue) and m (red) from 1 ... 8 . In Figure 7b the distances of the DPs to the median (i = 5, j = 5) are entered as a value in the corresponding box of the DP.

Table of Values to Determine the Median: The 64 distance values A-B of the Hodges-Lehmann distance matrix arranged in ascending order are entered in column 4 in the *Sorted* area. The position number of the ordered list is in

[8] The Hodges-Lehmann estimator is a robust non parametric estimator for one-dimensional symmetric distributions. Estimation of the median with m x n differences of the data points in a distribution [13].

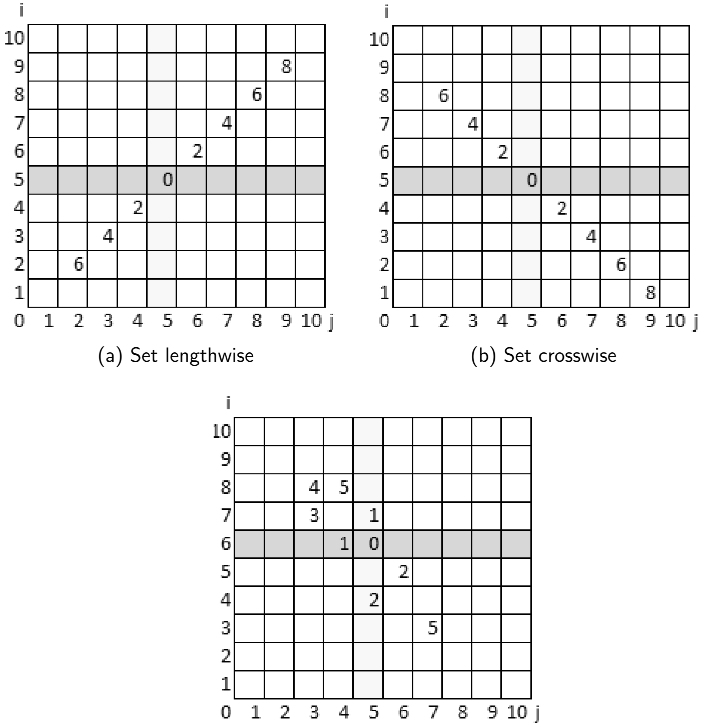


Allocation median		Sorted		DP_Aij - DP_Bij position median			A-B = DP_Aij - DP_Bij				
								DP_ Aij		DP_ Bij	
A-B position 1-64 in the distance-matrix list	Distance-value A-B	A-B position in the sorted list	Distance-value A-B	A-B position 1-64 in the distance-matrix list	DP-position m in the distance-matrix	DP-position n in the distance-matrix	Distance-value A-B	i	j	i	j
1	0	1	0	1	1	1	0				
10	0	2	0	2	2	1	2				
12	0	3	0	3	3	1	5				
39	4	24	2	24	8	3	7				
42	5	25	2	25	1	4	2				
44	5	26	2	26	2	4	0				
51	7	27	2	27	3	4	3	7	4	3	5
53	7	28	2	28	4	4	0				
11	0	29	3	29	5	4	3				
13	1	30	3	30	6	4	2				
18	2	31	3	31	7	4	5				
20	2	32	3	32	8	4	10				
27	3	33	3	33	1	5	5				
29	3	34	3	34	2	5	3				
34	3	35	3	35	3	5	0				
60	10	62	10	62	6	8	8				
8	0	63	12	63	7	8	5				
57	8	64	12	64	8	8	0				

Fig. 8: Value table for determining the median

column 3. With 64 list entries, the median has a value of 32.5. Position 33 (green/dark-grey) has been rounded up. This distance A-B corresponds to position 27 of the distance matrix in column 1 and the A-B value in column 2 (green/dark-grey). The 64 positions of the distance matrix (DP A_{ij}, DP B_{ij}) are located in column 5, in column 8 the corresponding A-B values, in column 6 the mth DP of the distance matrix (blue/far left-column) and in column 7 the nth DP of the distance matrix (red/upper row above matrix). The corresponding tuple of the DP for A is in column 9 and for B in column 10 (yellow/light-grey).

Multidimensional Median: The median is found in each dimension in the center between DP (7.4) for A and DP (3.5) for B, as their distance A-B is found exactly in the middle of the list arranged in ascending order (median) of the Hodges-Lehmann distance matrix. In Figure 7b the median for the i-axis between 7 and 3 at 5 (blue/dark-grey) and between 4 and 5 for the j-axis at 5 (yellow/light-grey) is shown. The multidimensional median of the point set is at the intersection of row 5 (blue/dark-grey) and column 5 (yellow/light-grey) in Figure 7b. The positions of the medians for the sets in Figures 9a to 9c are the intersections of the blue/grey and yellow/light-grey marked row/column, at which the Manhattan distances to the median are also entered in the position boxes of the DPs.



(a) Set lengthwise

(b) Set crosswise

(c) Set oval

Fig. 9: Different arrangement of DPs around the median

4.3 Measuring Values and Multidimensional Median

To establish the characteristic values of a distribution, estimations can be used. The data is sorted in ascending order on a list. In the following example 50 % of the data are involved for the estimate. This data set is selected from zero up to the median. Possible outliers are placed at the upper end of the sorted list.

1. The reference point for distance measuring is the DP on or nearest to the median.

2. The data from zero up to the median definitely include no outliers. They were used to estimate $f_{max}/2$. This data is located between the blue/dark-grey line at position 0 and the green/— line at position 5 on the abscissa axis in Figure 10a.

3. Outliers have a high difference value and are positioned on the right side above the 50 % boundry on the list of values sorted in ascending order in Figure 10a.

   Variance s^2 and threshold for the outlier detection: the significance level is part of the model M (parameter xy) and is fixed for this example with the data list in Figure 10c to 3σ for 99.7 % of the values.

4. Calculate the standard deviation s of the distribution with formula 8 in [15][9].

5. The distance values of the measured values for the TDS are positioned to approximately 99 % between the expected value $\mu + 3\sigma$ in Figure 10b, below the green/-dotted line, which represents the threshold for the outlier detection.

4.4 Removing of Outliers

Even though in this example 20 % of the data of the training data set (TDS) would be outliers, they can be safely removed out of the TDS. The upper end of the range for the determination of the variance is at 50 % of the number of values [10].

1. The threshold for outliers is the distance value at half-width $\frac{f_{max}}{2}$ of the sample plus three times of the standard deviation, estimated by s with formula 8 in [15] for the data set (significance level 3 σ).

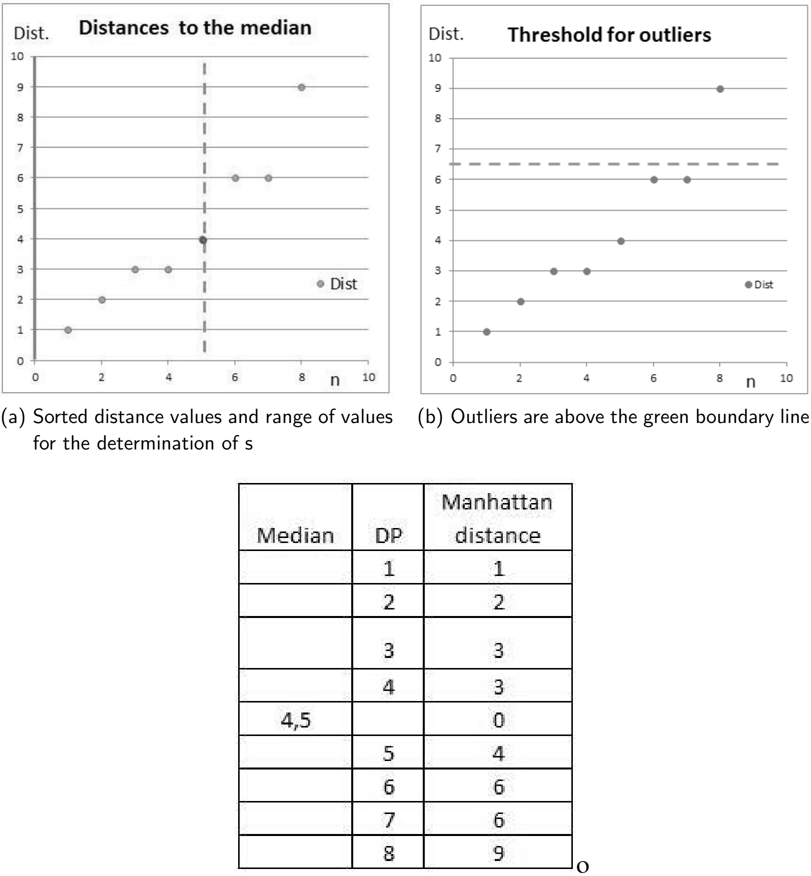
2. Distance values of the TDS, greater than the threshold for outliers, must be removed out of the TDS, to make it free of outliers.

3. All other values are part of the TDS. With this data set, the characteristic values μ and σ of the outlier-free population can be determined.

Table of Values of Manhattan-Distance to the Median: The 8 distance values to the origin of the data set in Figure 7a are located in the list of values in Figure8 in ascending order and and have been established as described in chapter 4. The range for the measured values, 0 % is marked with a blue/black line, 50 %, the median, is in Figure 10a marked with a green/dotted line. The threshold for outlier detection in Figure 10b is marked with a green/dotted line.

[9] For calculation of the variance of a random sample or calculations of parts of the population s used instead of σ.

[10] *As shown in: Der Bruchpunkt von Schätzern-Universität der Bundeswehr* [5].



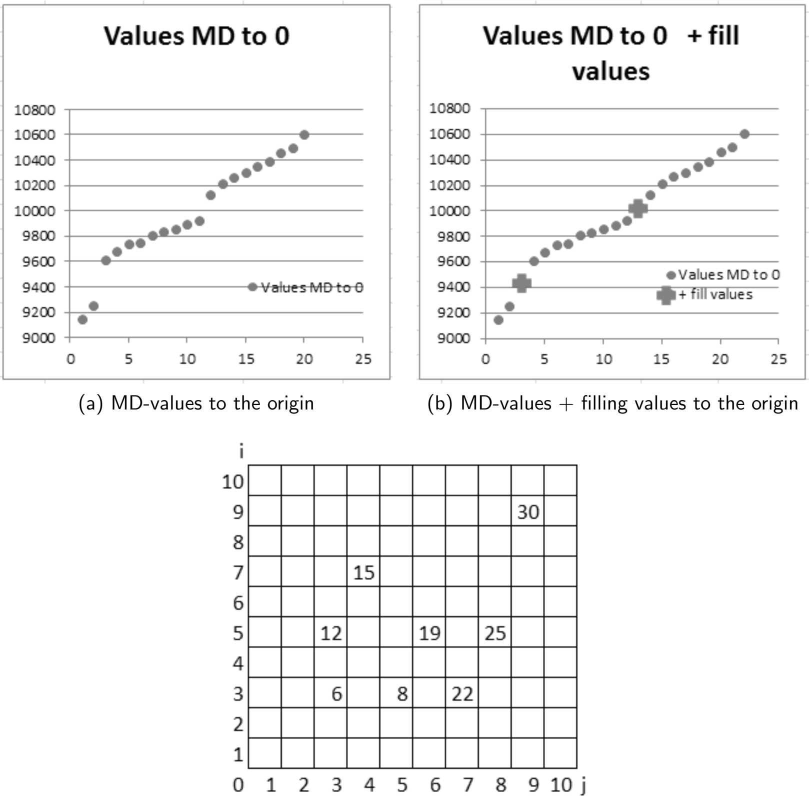
(a) Sorted distance values and range of values for the determination of s

(b) Outliers are above the green boundary line

Median	DP	Manhattan distance
	1	1
	2	2
	3	3
	4	3
4,5		0
	5	4
	6	6
	7	6
	8	9

(c) Table of values: Manhattan distance to median

Fig. 10: Distances to the median and outlier threshold



(a) MD-values to the origin

(b) MD-values + filling values to the origin

i \ j	1	2	3	4	5	6	7	8	9	10
10										
9									30	
8										
7				15						
6										
5			12			19		25		
4										
3			6		8		22			
2										
1										

(c) MD-values to the last value

Fig. 11: DPs arranged in ascending order of distance to the origin.

4.5 Removing Ties

1. Distance values smaller than $\mu - 2\sigma$ will be removed from the TDS data list. These values carry no information, increase the calculation time and need memory space.

2. Renumber the remaining data points (DPs) of the distance list. This complies with column *Scale-i* in table 12.

4.6 Maximum Distance between the Data Points

A set of data in multidimensional space, measured with a suitable measuring method, represents as a reference data set after the training phase a certain characteristic. For measuring the similarity to this data set, different metrics exist. Compared to other distance metrics, the Manhattan-Distance is precise, even with high dimensionality. In the MAE-method, the Manhattan-Distance, described in section 3.1, is used. With this measure (mathematical norm), the distances Dist2NN of the data points in the reference data set to their nearest neighbors (NN) were measured. The distribution of these distances must be checked and the significance level, in accordance with table 1 and figure 3 in [15] must be fixed, e. g. to 2σ. The significance level will be converted into the absolute Distance MaxDist2NN and is used to eliminate outliers above the limit $\mu + 2\sigma$ and values below the limit $\mu - 2\sigma$ in chapter 3 in [15]. The maximum distance *MaxDist2NN* is set to equal with the significance level $2\sigma = \mu$+2σ. *MaxDist2NN* is the first parameter of the model M for the examined TDS.

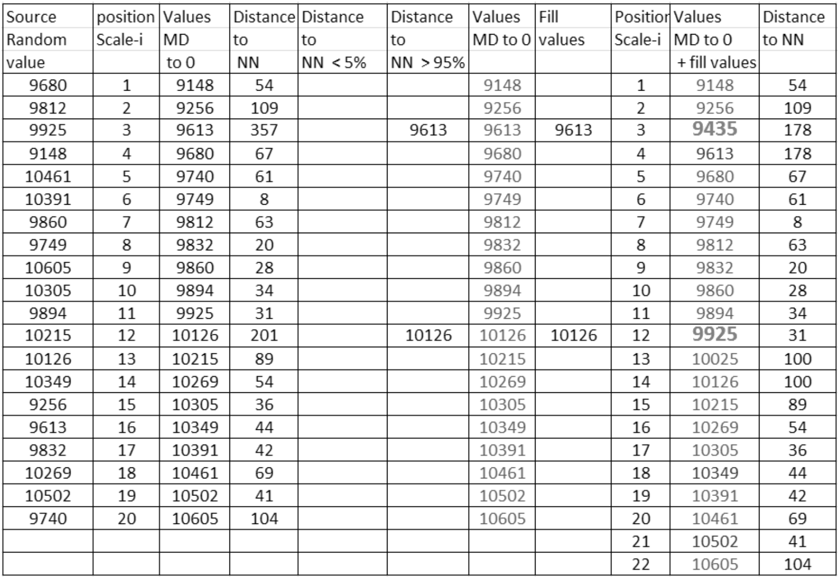
4.7 Supplementation of Gaps and Jumps (Discontinuities) in the TDS

1. Distances greater than $\mu + 2\sigma$ on the distance list will be reduced with fill values. A smooth manifold (MF) is a prerequisite for the application of NN-methods.

2. Gaps and jumps can be removed with the method, described in [14] in chapter 4.5.3. The fill-value method does not vary the eigenvalues of the TDS or varies them very little.

3. The remaining DPs in the distance list will be renumbered. This complies with column *Values MD to 0 + fill values* (Without gaps and jumps) in table 12.

Smoothing the MF: Ties will be removed and gaps and jumps (discontinuities) will be filled in. To do this, 20 random values around the mean value of 10000 (10,000) were generated. The distance values were arranged in ascending order and are listed in column 3. The distances between these values are listed in column 4. This DPs are shown in Figure 11a. Values with a smaller distance



Source	position	Values	Distance	Distance	Distance	Values	Fill	Positior	Values	Distance
Random	Scale-i	MD	to	to	to	MD to 0	values	Scale-i	MD to 0	to NN
value		to 0	NN	NN < 5%	NN > 95%				+ fill values	
9680	1	9148	54			9148		1	9148	54
9812	2	9256	109			9256		2	9256	109
9925	3	9613	357		9613	9613	9613	3	**9435**	178
9148	4	9680	67			9680		4	9613	178
10461	5	9740	61			9740		5	9680	67
10391	6	9749	8			9749		6	9740	61
9860	7	9812	63			9812		7	9749	8
9749	8	9832	20			9832		8	9812	63
10605	9	9860	28			9860		9	9832	20
10305	10	9894	34			9894		10	9860	28
9894	11	9925	31			9925		11	9894	34
10215	12	10126	201		10126	10126	10126	12	**9925**	31
10126	13	10215	89			10215		13	10025	100
10349	14	10269	54			10269		14	10126	100
9256	15	10305	36			10305		15	10215	89
9613	16	10349	44			10349		16	10269	54
9832	17	10391	42			10391		17	10305	36
10269	18	10461	69			10461		18	10349	44
10502	19	10502	41			10502		19	10391	42
9740	20	10605	104			10605		20	10461	69
								21	10502	41
								22	10605	104

Fig. 12: Table of values: Smoothing the MF

than 5 % to the nearest neighbors (NN) are ties. They are listed in column 5 and have been removed out of the TDS. Values with a distance greater than 95 % to the NN are defined as gaps and jumps. Fill-values were inserted with the fill value method, described in [14] in chapter 4.5.3. These DPs are listed in column 10 and are marked red/bold. All other values of the TDS are marked blue. The resulting TDS is listed in column 10, the associated position numbers are in column 9, the new difference values are in column 11 (check-values). In Figure 11b the smooth profile of the DPs of column 10 is shown.

4.8 Determination of the Characteristic Values μ and σ of the TDS

The reference point for determining the distances is now the origin. The resulting MF is now a TDS without outliers.

1. Calculate the difference values of the DPs of the TDS referenced to the origin and enter them onto a list.
2. Search for the lowest difference value DiffDP1 of the TDS.
3. Start the search from this DP to the NN and calculate the difference.

4. Enter the value DiffDP2 = DiffDP2 + DiffDP1 into the list as DiffDP2 (recursive).

5. Complete this list for all DPs of the TDS in the same manner.

   Determination of the expected value μ:

6. Arrange the DPs on the distance list in ascending order.

   Determine the variance and and the significance level:

7. Calculate the variance σ^2 with formula 5 in [15] and the standard deviation σ of the TDS with formula 4 in [15].

8. The significance level is part of the model M (parameter 2) and will be fixed here to 2 σ.

9. With this, the confidence interval equals with the range $\mu - 2\sigma; ...; \mu + 2\sigma$. These are the values UGR in column 3 and OGR in column 4 in table 13.

5 Definition of the Receptive Area

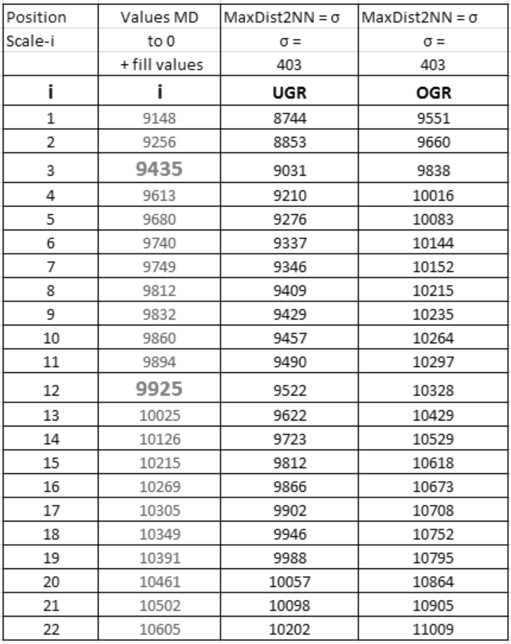
The multidimensional expected value μ is the centroid of the data set.

Comments on the Table of Values in Figure 13: The data values in the table in Figure 12 were edited with the instructions in the sections 4.2 to 4.4 and the results were listed in the table 13. The values in column 2 (blue) are original values, the red/bold values are fill values. The lower boundry UGR of the receptive area is listed in column 3 and the upper boundry in column 4. Figure 16a shows the range for the dimension i and Figure 16b for the dimension j.

Characteristic Values of the TDS: The characteristic values μ and σ of the TDS were established with the instructions in section 4.8 and are: expected value μ = 10008 (10,008), the variance σ^2 = 1054, the standard deviation σ is 32, the confidence interval is between the boundries μ - 2σ = 9943 and μ + 2σ = 10073.

Ties: With the instruction in section 4.5, all values lower than μ - 2σ were removed. The other values are listed in column 10 in Fig. 12.

Fill Values: With the instructions in section 4.7, all distance values between the DPs, which are greater than 4 σ were filled in with fill values (red/bold).



Position	Values MD	MaxDist2NN = σ	MaxDist2NN = σ
Scale-i	to 0	σ =	σ =
	+ fill values	403	403
i	**i**	**UGR**	**OGR**
1	9148	8744	9551
2	9256	8853	9660
3	**9435**	9031	9838
4	9613	9210	10016
5	9680	9276	10083
6	9740	9337	10144
7	9749	9346	10152
8	9812	9409	10215
9	9832	9429	10235
10	9860	9457	10264
11	9894	9490	10297
12	**9925**	9522	10328
13	10025	9622	10429
14	10126	9723	10529
15	10215	9812	10618
16	10269	9866	10673
17	10305	9902	10708
18	10349	9946	10752
19	10391	9988	10795
20	10461	10057	10864
21	10502	10098	10905
22	10605	10202	11009

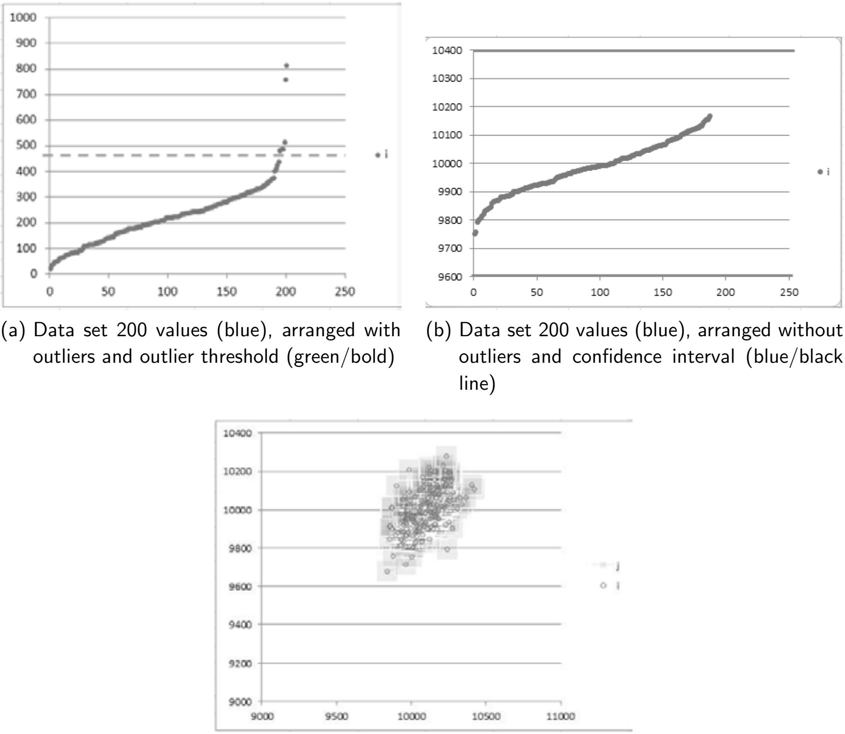
Fig. 13: Characteristic values of the training data set (TDS)

5.1 Applying the Method to a Larger Data Set

To demonstrate the method with more data, a data set with 200 values with outliers has been processed. The data values were processed with the instructions in sections 4.2 to 4.8. The outliers shown in Fig. 14a were then eliminated. In Figure 14a the sorted data values are shown in an ascending order. In Figure 14b the unsorted data set with the threshold for outlier detection is shown. In Figure 14c the resulting two dimensional TDS with the associated receptive area is shown.

5.2 Receptive Area

Receptive Area for One Expected Value in One Channel: The organization of the different types of measured values with channels were shown in Figure 2. The measured values of one channel, in the following Figures, the extension and the course of the receptive area are depicted. The receptive area represents the



(a) Data set 200 values (blue), arranged with outliers and outlier threshold (green/bold)

(b) Data set 200 values (blue), arranged without outliers and confidence interval (blue/black line)

(c) Data set 200 values (blue/dark), unsorted, with confidence interval (grey)

Fig. 14: Data set with 200 DPs and the corresponding receptive area

hull, into which all data points are placed (training) and which carries characteristic information for the object to be identified and includes the NN-points for the generalization of new measured enquiry points X_a to be classified. It is part of the model M for an object and fulfills the requirements with respect to model complexity in [14]. The corresponding characteristics of the TDS determine the size of this area. The regions around each point are mutually overlapping and together form the hull for all data points which represent the object. For this reason, the measure MaxDist2NN applies essentially to the hull. The overlapping of the regions around each data point will be achieved by smoothing of the MF with suitable fill values[11]. This situation is shown in Figure 15 for a two

[11] In chapter 4.5.3. in [14] the fill method for smoothing the MF is explained.

dimensional MF from the list. Its data points, arranged in ascending order and referenced to the origin were displayed in in Figure 15.

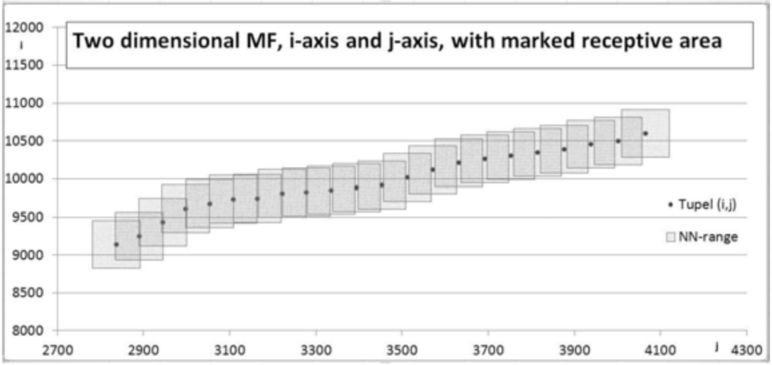
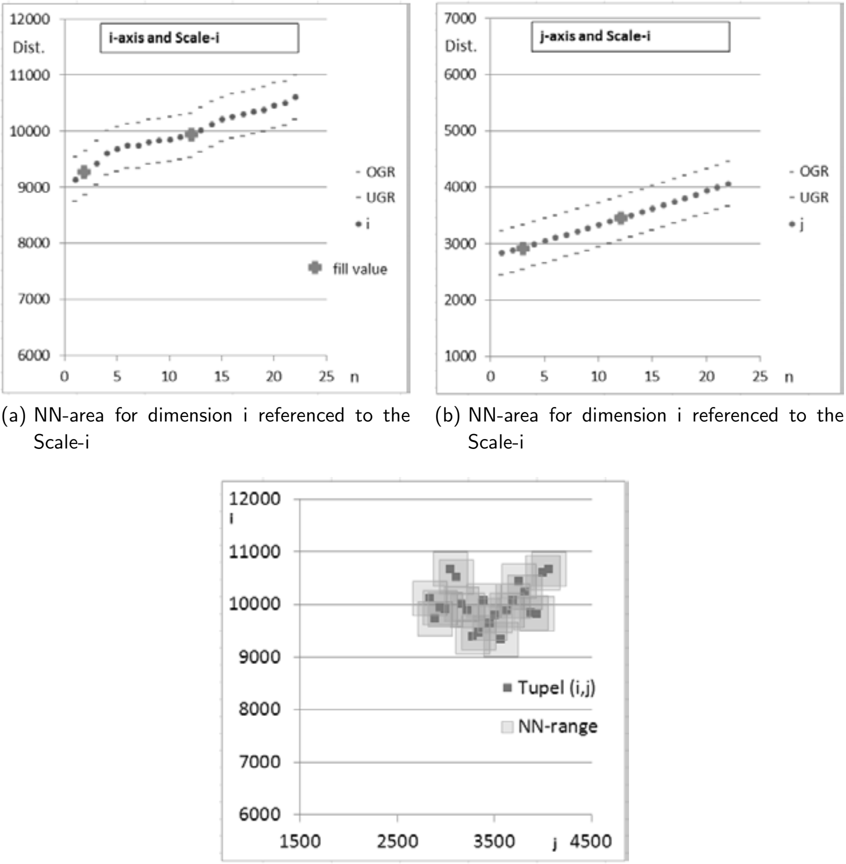


Fig. 15: Two dimensional distribution with mit MD-scaling. The NN-area of the DPs is arranged in ascending order

UGR in column 3 and OGR in column 4 in Figure 13 mark the bounds of the receptive area for the two-dimensional MF. The course of UGR marked (red/—) and OGR marked (blue/—) are for dimension *i* in Figure 16a and for dimension *j* in figure 16b . The size of the receptive area is shown for the two dimensional MF in Figure 16c. Due to the Manhattan-distance (L1-norm), the boundary lines of the receptive area are arranged quadratically around the DPs.

Receptive Area for Multiple Expectated Values in One Channel: For each measuring event, data points are distributed around the expectated value μ. Each distribution will be processed following the instructions with steps 1 to 9 in section 4.1. If several measuring events with the related distributions, e. g. K1.1 ... K1.4 in Figure 17 are present, then there are 4 expectated values in the related center of each distribution. All DPs of these distributions identify the object. These DPs K1.1 ... K1.4 are arranged in relation to the origin using distance to the origin, and the characteristic values were determined following the instructions of step 1 ... 10 in section 4.8. They are stored in model M. Therefore, it will later be recognizable if a NN-point is a fill value (red) or a measured value (blue) and it is recognizable to which measuring event (e. g. K1.1) the DP is related or to



(a) NN-area for dimension i referenced to the Scale-i

(b) NN-area for dimension i referenced to the Scale-i

(c) Two dimensional NN-area with MD-scaling, unsorted

Fig. 16: Resulting receptive area for the example with 8 DPs

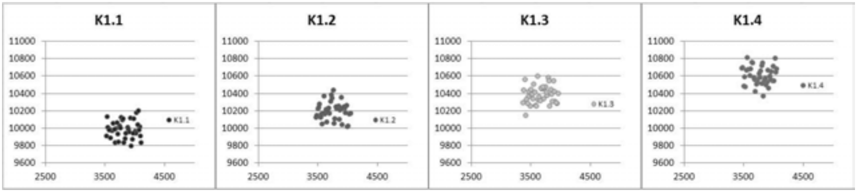
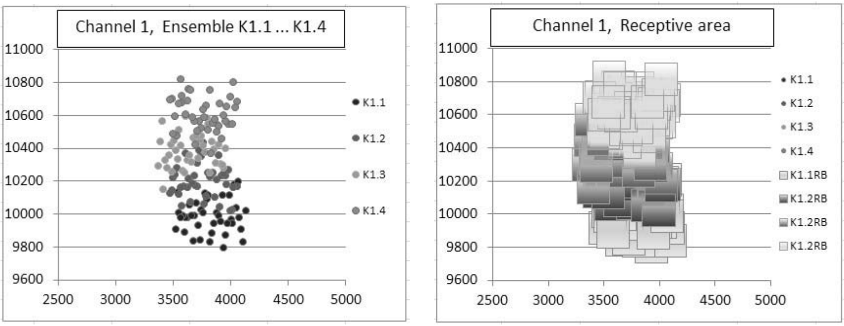


Fig. 17: Measured values of the measuring events K1.1, K1.2, K1.3 and K1.4



(a) Ensemble of measuring events K1.1, K1.2, K1.3 and K1.4

(b) Receptive area in channel 1 for object 1

Fig. 18: Ensemble of expected values, the DPs of its distributions and the related receptive area

which it tends. All points of the channel together, form together the receptive area, which is composed of an ensemble of expected values. In [14], it is explained how similarity is used to determine the membership degree for an object, e. g. a TDS. The receptive area is composed of an ensemble of expected values and the related distributions. Around each expected value there is a region defined by the standard deviation. When 95 % of the signals are to be recognized, so the measure for the receptive area is 2 σ. The characteristic values of this distributions are similar.

Final Note on the Receptive Area: σ was established on the basis of the distances from TP to TP of a data group, e.g. K1.1 . The fill method ensures a smooth MF and fills in missing values. The receptive area around each TP is defined by σ. It consists, for example, of an ensemble of expected values, data sets of the groups K1.1 ... K1.n, which together form the receptive area for the characterized object. With this the hull of the receptive area is the limit within which the new measured values (working phase) are to be expected for a match. They are related to the trained object. In Figures 14c and 16c one data set distributed around the expected value determines the receptive area. In Figure 4c four data sets around an ensemble of expected values determine the receptive area.

6 Skala-i

On the Scale-i, the ensemble of expected values is arranged in sequence.

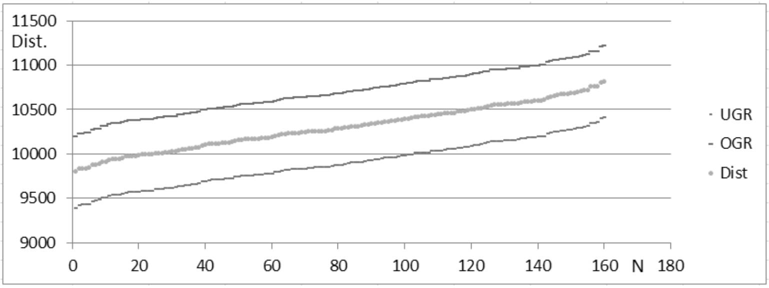


Fig. 19: NN-area and Scale-i

1. The list of distances is now arranged in ascending order to the reference point zero (origin) and includes a smooth MF.
2. The positions of the DPs are entered on the one-dimensional Scale-i. Each position represents a multidimensional DP.
3. The multidimensional structure of the data set is now reduced to the one-dimensional Scale-i, shown in Figure 20.

Summary: The examples used in this paper are intended to serve for explanatory purposes and have been kept simple, respectively to the measured data sets are chosen for practice-relevant conditions. For each measuring situation, the related measured data set will be checked for outliers; these will be eliminated and the set will be prepared for the following process steps. This is essential for real-world applications. For outlier detection a robust estimator (Hodges-Lehmann-estimator) was used to determine the median. The median, which is

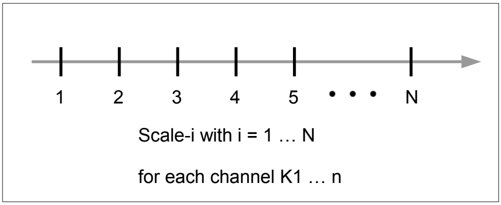


Fig. 20: Representation of the TPs with the one-dimensional Scale-i

independent of measured values and therefore of incorrect or inaccurate values, functions as a multidimensional median and as a reference point for the arrangement of all distance values. The distance values of each DP to each other DP will be calculated and listed in a distance matrix. All values will be arranged in ascending order and referenced to the multidimensional median so that all large distances are located in the upper range of the list. For the determination of outliers only the distance value to the median is from importance. With this method it will be ensure, that in the event of a large number of outliers a outlier-free TDS will be generated. With the choice of the value range below the half-width, conditions will be created, for estimating the standard deviation σ with the relationship $\sigma \approx FWHM/2.3348$ and for determining the threshold for outlier detection and elimination. After elimination, which is described in section 4.4, the TDS is free of outliers and the characteristic values of the TDS can be determined. The determination of the distance to the origin with the method described in section 4 ensures that each DP on the Scale-i can be clearly identified.

Scale-i: For the MAE-method the detection of all DPs and their arrangement in ascending order it is important to make it possible to include all points in the MF, and the position number on the one-dimensional Scale-i can be entered. All points of the TDS in the space will be arranged one beneath the other like on a string of beads in ascending-distance order. Around this series of data, the receptive area of the data model is defined. The prevention of gaps in the manifold (MF) ensures numerical stability and the formation in space and the associated distribution its related and a defined accuracy for the generalization in the depiction space (Z-space). When later in the working phase points appear transversely, like in Figure 5b, in any event the distance to the NN will be used and not the distance to the origin.

Sequence of the DP-numbers on the Scale-i: On the Scale-i all DPs of the TDS are sorted in ascending-distance order with respect to the origin. One can recognize through the sequence of the DP-numbers on the Scale-i how the DPs are arranged in the NN-area during classification of a new enquiry vector (DP X_a) . When, for example by using a dimensionality reduction (DR), the structure of the MF is reconstructed with too few dimensions, cross points will appear in the course of the Scale-i. This is expressed in a discontinuous sequence of the DPs in the NN-area. Instead of four NN with the numbers 91, 92, 93 and 94, there for example the DP-numbers 91, 92, 1934, 95 in the NN-area of X_a will be found. The DP with the DP-number originates out of a region of the MF which crosses the MF at 91 - 95. Before the dimensionality reduction is carried out, it is unclear, how many dimensions are necessary to reconstruct the inner structure of the MF. To avoid crossings in the structure of the reduced MF, the number of dimension must be chosen one or more times higher, and then the calculation of the DR has to be repeated[12].

References

[1] Andrews, D. (1972). *Plots of high-dimensional data*, Biometrics 28: 125-136.

[2] Backhaus, K.; Erichson, B.; Plinke, Wulff; Weiber, R.: *Multivariate Analysemethoden*. 11. Auflage Springer, Berlin, Heidelberg, New York, 2005

[3] Breuer, Dirk: *Abstandsmaße für die multivariate adaptive Einbettung*. Fern-Universität in Hagen, 2014

[4] Bronstein, I.N.; Semendjajew, K.A.; Musiol, G.; Mühlig, H.: *Taschenbuch der Mathematik*. 4.Aufl., Frankfurt am Main; Thun: Verlag Harri Deutsch, 1999

[5] Caspary, Wilhelm: *Der Bruchpunkt von Schätzern – Universität der Bundeswehr*. München. Internet 2017 (`www.unibw.de/IfG/Org/schriftenreihe/pdf-ordner/heft.../87-caspary-39-46.pdf`)

[6] Diehl, Joerg M.; Kohr, Heinz: *Deskriptive Statistik*.13. Auflage, ISBN 3-88074-110-7, Verlag Dietmar Klotz, Eschborn, 2004

[7] Faloutsos, Christos: *Searching Multimedia Databases by Content*, The Kluwer International Series on Advances in Database Systems. Dordrecht: Kluwer Academic Publishers, 1996

[8] Gottwald,S.; Köstner, H.; Hullwich, M.: *Handbuch der Mathematik*. VEB Bibligraphisches Institut Leipzig, 1986

[12] In the literature on this subject, the count for the required dimensionality will be decided by the number of eigenvalues which are greater than a specified value. This subject is discussed in [14].

[9] Kriegel, J.; Gröger, P.; Schubert, E.: *Interpreting and Unifying Outlier Scores*. In: Proc. 11th SIAM International Conference on Data Mining. 2011, `http://epubs.siam.org/doi/pdf/10.1137/1.9781611972818.2`

[10] Mildenberger, Otto: *System- und Signaltheorie*. 3. Aufl., Braunschweig; Wiesbaden: Friedr. Vieweg&Sohn Verlagsgesellschaft mbH, 1995

[11] Patzelt, Rupert: *Elektrische Meßtechnik*. Springer Wien, New York, 1996

[12] Pfeifer, Tilo; Profos, Paul: *Handbuch der industriellen Messtechnik*. Oldenbourg Industrieverlag, 2008

[13] Rosenkranz, Gerd, K.: *A note on the Hodges Lehmann estimator*. In: Pharmaceut. Statist. 9:162-167 (2010) Published online 29 August 2009 in Wiley InterScience; DOT: 10.1002/pst.387

[14] Sartorius, Gerhard: *Multivariate Adaption mit künstlichen neuronalen Netzen*. VDI-Verlag, 2009

[15] Sartorius, Gerhard: Metrology and Machine Learning. In: *Autonomous Systems 2017*, Proceedings of the 10th Conference. VDI-Verlag GmbH, 2017

[16] Weiber, Rolf; Mühlhaus, Daniel: *Strukturgleichungsmodellierung*. 2. Aufl., Berlin, Heidelberg: Springer Verlag, 2014

Graph-based Text Mining and Information Retrieval with Neo4j

Mario Kubek

Chair of Communication Networks
FernUniversität in Hagen, Germany

Abstract:
This talk gives an introduction to the graph database Neo4j and explores its usefulness in searching for text documents while relying on graph-based Text Mining techniques to inherently support this task. It also discusses approaches to programmatically access Neo4j using its application programming interfaces. In doing so, Neo4j's core graph algorithms will be highlighted as they can be applied on large graphs as well.

Prime Clocks

Michael Stephen Fiske

Aemea Institute, San Francisco, California

Abstract: Physical implementations of digital computers began in the latter half of the 1930's and were first constructed from various forms of logic gates. Based on the prime numbers, this paper introduces prime clocks and prime clock sums, where the clocks utilize time and act as computational primitives instead of gates. The prime clocks generate an infinite Abelian group, where for each n, there is a finite subgroup S such that for each Boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$, there exists a finite prime clock sum in S that can represent and compute f. A parallelizable algorithm, implemented with a finite prime clock sum, is provided that computes f. In contrast, the negation $\neg$, conjunction $\wedge$, and disjunction $\vee$ operations generate a Boolean algebra. In terms of computation, Boolean circuits computed with logic gates `NOT`, `AND`, `OR` have a depth. This means that a completely parallel computation of Boolean functions is not possible with these gates. Overall, some new connections between number theory, Boolean functions and computation are established.

1 Introduction

1.1 Notation and Preliminaries

Symbol $\mathbb{Z}$ denotes the integers and $\mathbb{N}$ the non-negative integers. For any $n \in \mathbb{N}$ such that $n \geq 2$ and $a \in \mathbb{N}$ such that $0 \leq a \leq n-1$, consider the equivalence class $[a] = \{a + kn : k \in \mathbb{Z}\}$ that is a subset of $\mathbb{Z}$. Let $\mathbb{Z}_n = \{[0], [1], \ldots, [n-1]\}$. a mod n is the remainder when a is divided by n. In the standard manner, $(\mathbb{Z}_n, +_n)$ is an Abelian group, where binary operator $+_n$ is defined as $[a] +_n [b] = \big[(a+b) \bmod n\big]$. The brackets are sometimes omitted and $[a] \in \mathbb{Z}_n$ is represented with the integer a, satisfying $0 \leq a \leq n-1$. The set of all functions $f : \mathbb{N} \rightarrow \mathbb{Z}_n$ is denoted as ${\mathbb{Z}_n}^{\mathbb{N}}$. Symbol $\overline{c}$ is the constant function $f : \mathbb{N} \rightarrow \mathbb{N}$ where $f(m) = c$ for all $m \in \mathbb{N}$. The set of all n-bit strings is $\{0,1\}^n$. It is convenient to identify the 2 bits in $\{0,1\}$ with the elements $[0]$ and $[1]$ in $\mathbb{Z}_2$.

The least common multiple of positive integers a and b is $\text{lcm}(a, b)$. Let $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7, \ldots$ where the nth prime number is p_n. Let p be an odd prime. p is called a 3 mod 4 prime if $\frac{p-1}{2}$ is odd. p is called a 1 mod 4 prime if $\frac{p-1}{2}$ is even.

1.2 Intuition and Motivation for Prime Clocks

Physical implementations of digital computers began in the latter half of the 1930's and early designs were based on various implementations of logic gates [1, 4, 5, 17–19] (e.g., mechanical switches, electro-mechanical devices, vacuum tubes). The transistor was conceptually invented [9, 10] in the late 1920's, but the first working prototype [2, 13] was not demonstrated until 1947. Transistors act as building blocks for logic gates when they operate above threshold [11]. The transistor enabled the invention of the integrated circuit [8, 12], which is the physical basis for modern digital computers.

As an alternative to gates, prime clocks are based on the prime numbers and the notion of a common clock. Consider the prime number 2 and the clock $[2, 0]$. The 2 means that the clock has two states $\{0, 1\}$ and the 0 means that the clock starts ticking from state 0 at time 0. Shown in column 2 of table 1, the clock $[2, 0]$ ticks $0, 1, 0, 1$, and so on. In column 3 of table 1, the clock $[3, 1]$ has 3 states $\{0, 1, 2\}$ and ticks $1, 2, 0, 1, 2, 0$ and so on.

Table 1: Some Prime Clocks and Sums in $\mathbb{Z}_2{}^{\mathbb{N}}$

Time	$[2,0]$	$[3,1]$	$[2,0] \oplus [3,1]$	$[7,3]$	$[13,6]$	$[7,3] \oplus [13,6]$
0	0	1	**1**	3	6	**1**
1	1	2	**1**	4	7	**1**
2	0	0	**0**	5	8	**1**
3	1	1	**0**	6	9	**1**
4	0	2	**0**	0	10	**0**
5	1	0	**1**	1	11	**0**
...						

Expressed as $\oplus$ in table 1, two or more prime clocks can be added and their sum can be projected into $\mathbb{Z}_2{}^{\mathbb{N}}$. The fourth column of table 1 shows the sum of clocks $[2, 0]$ and $[3, 1]$, projected into $\mathbb{Z}_2{}^{\mathbb{N}}$. This paper primarily focuses on prime clock sums, projected into $\mathbb{Z}_2{}^{\mathbb{N}}$, since they can compute Boolean functions.

Prime clocks sums projected into ${\mathbb{Z}_2}^{\mathbb{N}}$ have a mathematical property that has a practical application. This property is formally stated in theorem 6: for every natural number n, every Boolean function $f : \{0,1\}^n \to \{0,1\}$ can be computed with a finite prime clock sum that lies inside the infinite Abelian group $({\mathbb{Z}_2}^{\mathbb{N}}, \oplus)$. This means prime clocks can act as computational primitives instead of gates [14, 15]. A computer can be built from physical devices that implement prime clocks and their sums. In this regard, theorem 6 establishes a connection between number theory, Boolean functions and computation.

Prime clock addition $\oplus$ is associative and commutative. These two group properties enable prime clocks to compute in parallel, while gates do not have this favorable property. For example, $\neg(x \wedge y) \neq (\neg x) \wedge y$ because $\neg(0 \wedge 0) = 1$ while $(\neg 0) \wedge 0 = 0$. The unary operation $\neg$, conjunction operation $\wedge$, and disjunction operation $\vee$ form a Boolean algebra [7], so circuits built from the `NOT`, `AND`, and `OR` gates must have a depth.

Shown in the last column of table 1, the clock sum $[7,3] \oplus [13,6]$, helps illustrate the disparity between the *parallelization of prime clock sums* versus the *circuit depth of gates*. Figure 1 shows a gate-based circuit with depth 5 that computes $[7,3] \oplus [13,6]$ on $\{0,1\}^4$. This circuit computes Boolean function $h : \{0,1\}^4 \to \{0,1\}$, where $h(x_0\ x_1\ x_2\ x_3) = \big[\big(\neg(x_0 \wedge x_1)\big) \wedge (\neg x_2) \wedge x_3\big] \vee \big[x_0 \wedge x_1 \wedge x_2 \wedge (\neg x_3)\big] \vee \big[(\neg x_2) \wedge (\neg x_3)\big]$. Note $\big([7,3] \oplus [13,6]\big)(m) = h(x_0\ x_1\ x_2\ x_3)$, whenever $m = x_0 + 2x_1 + 4x_2 + 8x_3$.

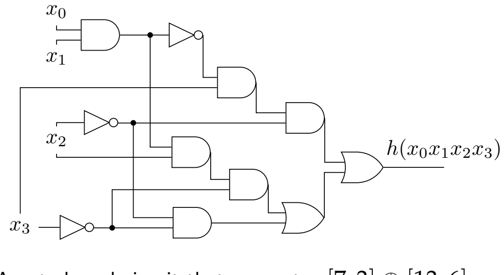


Fig. 1: A gate-based circuit that computes $[7,3] \oplus [13,6]$ on $\{0,1\}^4$.

This disparity enlarges for Boolean functions $f : \{0,1\}^n \to \{0,1\}$ as n increases. Informally, Shannon's theorem [14] implies that most functions $f : \{0,1\}^n \to \{0,1\}$ require on the order of $\frac{2^n}{n}$ gates. More precisely, let $\beta(\epsilon, n)$ be the number of distinct functions $f : \{0,1\}^n \to \{0,1\}$ that can be computed by circuits with at most $(1-\epsilon)\frac{2^n}{n}$ gates built from the `NOT`, `AND`, and `OR` gates. Shannon's theorem states for any $\epsilon > 0$, then $\lim\limits_{n\to\infty} \frac{\beta(\epsilon,n)}{2^{2^n}} = 0$.

Let the gates of a circuit be labeled as $\{g_1, g_2, \ldots, g_m\}$ where m is about $\frac{2^n}{n}$. The graph connectivity of the circuit specifies that the output of gate g_1 connects to the input of gate g_{k_1}, and so on. Shannon's theorem implies that for most of these Boolean functions the graph connectivity requires an exponential (in n) amount of information. This is readily apparent after comparing the number of symbols used in $[7,3] \oplus [13,6]$ versus the symbolic expression $\big[\big(\neg(x_0 \wedge x_1)\big) \wedge (\neg x_2) \wedge x_3\big]$ $\vee \big[x_0 \wedge x_1 \wedge x_2 \wedge (\neg x_3)\big] \vee \big[(\neg x_2) \wedge (\neg x_3)\big]$.

Consider a cryptographic application that uses a function $h : \{0,1\}^{20} \to \{0,1\}^{20}$, where $h = (h_0, \ldots, h_{19})$ and each $h_i : \{0,1\}^{20} \to \{0,1\}$ is highly nonlinear [6]. Then over 1 million gates can be required to compute h, since $\frac{2^{20}}{20} = 52428$ and there are 20 distinct h_i functions. Using the first 559 prime numbers (i.e., all primes ≤ 4051), finite prime clock sums can compute any function $f_{20} : \{0,1\}^{20} \to \{0,1\}$ even though there are $2^{2^{20}} = 2^{1048576}$ distinct functions. This means a physical realization[1] with prime clocks may use the first 599 prime numbers to implement an arbitrary $h : \{0,1\}^{20} \to \{0,1\}^{20}$.

Lastly, the structure of our paper is summarized. Section 2 provides formal definitions of a prime clock, prime clock sums, and some results about the periodicity of finite prime clock sums. Section 3 covers prime clock sums projected into $\mathbb{Z}_2^{\mathbb{N}}$, where the main theorem is that any Boolean function $f : \{0,1\}^n \to \{0,1\}$ can be computed with a finite prime clock sum. Section 4 provides a parallelizable algorithm for computing a Boolean function with prime clock sums.

2 Prime Clocks

Definition 1 *Prime Clocks*

Let p be a prime number. Let $t \in \mathbb{N}$ *such that* $0 \leq t \leq p-1$. *Define* $[p,t] : \mathbb{N} \to \mathbb{N}$ *as* $[p,t](m) = (m+t) \bmod p$. *Function* $[p,t]$ *is called a p-clock that starts ticking with its hand pointing to t.*

Herein the expression *prime clock* $[p,t]$ always assumes that $0 \leq t \leq p-1$. Thus, if $p \neq q$ or $s \neq t$, then prime clock $[p,s]$ is not equal to $[q,t]$; equivalently, if $p = q$ and $s = t$, then $[p,s] = [q,t]$.

[1] Prospective physical realizations of prime clocks are beyond the scope of this paper.

For the nth prime p_n, let $\mathcal{P}_n = \{[p_n, 0], [p_n, 1], \ldots, [p_n, p_n - 1]\}$ be the distinct p_n-clocks. The set of all prime clocks is defined as

$$\mathcal{P} = \bigcup_{n=1}^{\infty} \mathcal{P}_n \qquad (1)$$

For $n \geq 2$, let $\Omega_n = \mathbb{Z}_n{}^{\mathbb{N}}$. Define $\pi_n : \mathcal{P} \to \Omega_n$ as the projection of each p-clock into Ω_n where $\pi_n([p,t](m)) = \big([p,t](m)\big) \bmod n$.

Definition 2

Let $n \in \mathbb{N}$ such that $n \geq 2$. On the set $\mathcal{P}$ of all prime clocks, define the binary operator $\oplus_n$ as $\big([p,s] \oplus_n [q,t]\big)(m) = \big([p,s](m) + [q,t](m)\big) \bmod n$, where $+$ is computed in $\mathbb{Z}$. Observe that $[p,s] \oplus_n [q,t] \in \Omega_n$.

Definition 3 *Finite Prime Clock Sum*

Similarly, with prime clocks $[q_1, t_1]$, $[q_2, t_2] \ldots$ and $[q_l, t_l]$, the function $[q_1, t_1] \oplus_n [q_2, t_2] \cdots \oplus_n [q_l, t_l] : \mathbb{N} \to \mathbb{Z}_n$ can be constructed. For each $m \in \mathbb{N}$, define $([q_1, t_1] \oplus_n [q_2, t_2] \oplus_n \cdots \oplus_n [q_l, t_l])(m) = \big([q_1, t_1](m) + [q_2, t_2](m) + \cdots + [q_l, t_l](m)\big) \bmod n$, where $+$ is computed in $\mathbb{Z}$. $[q_1, t_1] \oplus_n [q_2, t_2] \cdots \oplus_n [q_l, t_l]$ is called a finite prime clock sum in Ω_n.

Table 2 shows a *finite prime clock sum* in Ω_5.

Table 2: Some Prime Clocks and their Sum in Ω_5

Time	$[5,3]$	$[7,6]$	$[11,3]$	$[13,0]$	$[5,3] \oplus_5 [7,6] \oplus_5 [11,3] \oplus_5 [13,0]$
0	3	6	3	0	2
1	4	0	4	1	4
2	0	1	5	2	3
3	1	2	6	3	2
4	2	3	7	4	1
5	3	4	8	5	0
6	4	5	9	6	4
7	0	6	10	7	3
8	1	0	0	8	4
...					

Definition 4 *Let $r_1, \ldots r_k$ be k prime numbers and $q_1, \ldots q_r$ be r prime numbers. Let $f = [r_1, s_1] \oplus_n [r_2, s_2] \cdots \oplus_n [r_k, s_k]$. Let $g = [q_1, t_1] \oplus_n [q_2, t_2] \cdots \oplus_n [q_l, t_l]$. Define $f \oplus_n g$ in Ω_n as $\left(f \oplus_n g\right)(m) = f(m) +_n g(m)$, where $+_n$ is the binary operator in the group $(\mathbb{Z}_n, +_n)$.*

Definition 4 is well-defined with respect to definition 3 (i.e., $f \oplus_n g = [r_1, s_1] \oplus_n [r_2, s_2] \cdots \oplus_n [r_k, s_k] \oplus_n [q_1, t_1] \oplus_n [q_2, t_2] \cdots \oplus_n [q_l, t_l]$) because $(m_1 + m_2) \bmod n = \big((m_1 \bmod n) + (m_2 \bmod n)\big) \bmod n$ for any $m_1, m_2 \in \mathbb{N}$.

Remark 1 $(m_1 + m_2) \bmod n = \big((m_1 \bmod n) + (m_2 \bmod n)\big) \bmod n$ *for any* $m_1, m_2 \in \mathbb{N}$.

PROOF. Euclid's division algorithm implies $m_1 = k_1 n + r_1$ and $m_2 = k_2 n + r_2$, where $0 \leq r_1, r_2 < n$. Thus, $(m_1 + m_2) \bmod n = \big((k_1 + k_2)n + r_1 + r_2\big) \bmod n = (r_1 + r_2) \bmod n = \big((m_1 \bmod n) + (m_2 \bmod n)\big) \bmod n$ □

The binary operator $\oplus_n$ can be extended to all of Ω_n. For any $f, g \in \Omega_n$, define $\left(f \oplus_n g\right)(m) = f(m) +_n g(m)$. The associative property $(f \oplus_n g) \oplus_n h = f \oplus_n (g \oplus_n h)$ follows immediately from the fact that $+_n$ is associative. The zero function $\bar{0}$, where $\bar{0}(m) = 0$ in $\mathbb{Z}_n$, is the identity in Ω_n. For any f in Ω_n, its unique inverse f^{-1} is defined as $f^{-1}(m) = -f(m)$, where $-f(m)$ is the inverse of $f(m)$ in the group $(\mathbb{Z}_n, +_n)$. The commutativity of $\oplus_n$ follows from the commutativity of $+_n$, so $(\Omega_n, \oplus_n)$ is an Abelian group.

Let $\mathcal{Q}$ be a subset of the prime clocks $\mathcal{P}$. Using the projection π_n of $\mathcal{Q}$ into Ω_n, define $S_{\mathcal{Q}} = \{H : H \supseteq \pi_n(\mathcal{Q}) \texttt{ and } H \texttt{ is a subgroup of } \Omega_n\}$. The subset $\mathcal{Q}$ generates a subgroup of $(\Omega_n, \oplus_n)$. Namely,

$$\bigcap_{H \in S_{\mathcal{Q}}} H \tag{2}$$

This paper focuses on subgroups of Ω_2, generated by a finite number of prime clocks; consequently, the more natural symbol $\oplus$ is used instead of $\oplus_2$.

Definition 5 *Periodic Functions*

$f \in \Omega_n$ is a periodic function if there exists a positive integer b such that for every $m \in \mathbb{N}$, then $f(m) = f(m + b)$. Furthermore, if a is the smallest positive integer such that $f(m) = f(m + a)$ for all $m \in \mathbb{N}$, then a is called the period of f. After k substitutions of $m + a$ for m, this implies for any $m \in \mathbb{N}$ that $f(m) = f(m + ka)$ for all positive integers k.

Table 3: Some 2-Clocks, 3-Clocks and Sums in Ω_2

Time	$[2,0]$	$[2,1]$	$[3,0]$	$[3,1]$	$[2,0] \oplus [3,0]$	$[2,1] \oplus [3,0]$	$[2,0] \oplus [3,1]$
0	0	1	0	1	**0**	**1**	**1**
1	1	0	1	2	**0**	**1**	**1**
2	0	1	2	0	**0**	**1**	**0**
3	1	0	0	1	**1**	**0**	**0**
4	0	1	1	2	**1**	**0**	**0**
5	1	0	2	0	**1**	**0**	**1**
6	0	1	0	1	**0**	**1**	**1**
7	1	0	1	2	**0**	**1**	**1**
...							

Table 3 shows that both prime clocks $[2,0]$ and $[2,1]$ projected into Ω_2 have period 2. Both prime clocks $[3,0]$ and $[3,1]$ projected into Ω_2 have period 3. Each prime clock sum $[2,0] \oplus [3,0]$, $[2,1] \oplus [3,0]$ and $[2,0] \oplus [3,1]$ has period 6.

For any $f \in \Omega_n$, define the relation $\underset{f}{\sim}$ on $\mathbb{N}$ such that $x \underset{f}{\sim} y$ if and only if for all $m \in \mathbb{N}$, $f(m) = f(m + |y - x|)$. Trivially, $\underset{f}{\sim}$ is reflexive and symmetric.

To verify transitivity of $\underset{f}{\sim}$, suppose $x \underset{f}{\sim} y$ and $y \underset{f}{\sim} z$. W.L.O.G., suppose $x \leq y \leq z$. (The other orderings of x, y and z can be handled by permuting x, y and z in the following steps.) This means for all $m \in \mathbb{N}$, $f(m + y - x) = f(m)$; and for all $k \in \mathbb{N}$, $f(k) = f(k + z - y)$. This implies that for all $m \in \mathbb{N}$, $f(m + z - x) = f(m + z - y + y - x) = f(m + y - x) = f(m)$. Thus, $\underset{f}{\sim}$ is an equivalence relation.

When f is periodic with period a, each equivalence class is of the form $[k] = \{k + ma : m \in \mathbb{N}\}$, where $0 \leq k < a$. Thus, f has period a implies there are a distinct equivalence classes on $\mathbb{N}$ with respect to $\underset{f}{\sim}$.

Remark 2 *If a is the period of f and b is a positive integer such that $f(m) = f(m + b)$ for all $m \in \mathbb{N}$, then a divides b.*

PROOF. First, verify that $a \underset{f}{\sim} b$. By the definition of period, $a \leq b$ and for all $m \in \mathbb{N}$, then $f(m + b - a) = f(m + a + b - a) = f(m + b) = f(m)$. From the prior observation, a lies in $[0]$ and b also lies in $[0]$. Thus, $b = ma$ for some positive integer m. □

Lemma 1 *If $f, g \in \Omega_n$ are periodic, then $f \oplus_n g$ is periodic. Further, if the period of f is a and the period of g is b, then $f \oplus_n g$ has a period that divides* $\text{lcm}(a, b)$.

PROOF. Let a be the period of f and b the period of g. Let $l_{a,b} = \text{lcm}(a, b)$. $l_{a,b} = ia$ and $l_{a,b} = jb$ for positive integers i, j. For any $m \in \mathbb{N}$, $\big(f \oplus_n g\big)(m)$ $= f(m) +_n g(m) = f(m + ia) +_n g(m + jb) = f(m + l_{a,b}) +_n g(m + l_{a,b}) = \big(f \oplus_n g\big)(m + l_{a,b})$. Thus, $f \oplus_n g$ is periodic and remark 2 implies its period divides $l_{a,b}$. □

In regard to lemma 1, if $g = -f$, then the period of $f \oplus_n g$ is 1.

Remark 3 *There are n^a distinct periodic functions $f \in \Omega_n$ whose period divides a.*

PROOF. Since f is periodic and its period divides a, the values of $f(0), f(1), \ldots,$ $f(a - 1)$ uniquely determine f. There are n choices for $f(0)$. There are n choices for $f(1)$, and so on. □

Remark 4 *Suppose p is prime. There are $n^p - n$ distinct periodic functions $f \in \Omega_n$ with period p.*

PROOF. Consider a finite sequence $c_0, c_1, \ldots, c_{p-1}$ of length p where each $c_i \in \mathbb{Z}_n$ This sequence uniquely determines a periodic f such that $f(m + p) = f(m)$ for all $m \in \mathbb{N}$. In particular, $f(0) = c_0$, $f(1) = c_1, \ldots, f(p - 1) = c_{p-1}$. There are n^p periodic functions with a period that divides p. If the period of f is less than p, then remark 2 implies f has period 1 since p is prime. There are n distinct, constant (period 1) functions in Ω_n Thus, the remaining $n^p - n$ periodic functions have period p. □

Remark 5 *The prime clock $[p, t]$, projected into Ω_n, has period p.*

PROOF. Since p is prime, this follows immediately from remark 2. □

Theorem 1 *Finite Prime Clock Sums are Periodic*

Any finite sum of prime clocks $[q_1, t_1] \oplus_n [q_2, t_2] \oplus_n \cdots \oplus_n [q_l, t_l]$ is periodic.

PROOF. Use induction and apply remark 5 and lemma 1. □

3 Prime Clock Sums in Ω_2

Remark 6 *$[p,t] \oplus [p,t] = \overline{0}$ for any prime clock $[p,t]$.*

Per definition 2, $\big([p,k] \oplus [p,k]\big)(m) = \big([p,k](m) + [p,k](m)\big) \bmod 2 = 0$ in $\mathbb{Z}_2$.

Remark 7 *Let p be an odd prime. If p is a $3 \bmod 4$ prime, then $[p,0] \oplus [p,1] \oplus \cdots \oplus [p,p-1] = \overline{1}$. If p is a $1 \bmod 4$ prime, then $[p,0] \oplus [p,1] \oplus \cdots \oplus [p,p-1] = \overline{0}$.*

PROOF. $\big([p,0] \oplus [p,1] \oplus \cdots \oplus [p,p-1]\big)(0) = (0 + 1 + \cdots + p - 1) \bmod 2 = \frac{1}{2}(p-1)p \bmod 2$. For each $m > 0$, $\big([p,0] \oplus [p,1] \oplus \cdots \oplus [p,p-1]\big)(m)$ is a permutation of the sum inside $(0 + 1 + \cdots + p - 1) \bmod 2$. □

For the special case $p = 2$, observe that $[2,0] \oplus [2,1] = \overline{1}$.

Definition 6 *A finite sum $[q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_l,t_l]$ of prime clocks is non-repeating if $i \neq j$ implies $[q_i,t_i]$ is not equal to $[q_j,t_j]$.*

Remark 8 *Any finite sum $[q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_l,t_l]$ of prime clocks in Ω_2 can be reduced to a non-repeating finite sum $[q_{i_1},t_{i_1}] \oplus [q_{i_2},t_{i_2}] \oplus \cdots \oplus [q_{i_r},t_{i_r}]$, where $r \leq l$ such that for any $m \in \mathbb{N}$, $\big([q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_l,t_l]\big)\ (m) = \big([q_{i_1},t_{i_1}] \oplus [q_{i_2},t_{i_2}] \oplus \cdots \oplus [q_{i_r},t_{i_r}]\big)(m)$.*

PROOF. Since $(\Omega_2, +_2)$ is Abelian, if necessary, rearrange the order of $[q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_l,t_l]$, so that the prime clocks are ordered using the dictionary order. If two or more adjacent prime clocks are equal, then the associative property and remark 6 enables the cancellation of even numbers of equal prime clocks. This reduction can be performed a finite number of times so that the resulting sum is non-repeating. □

Definition 7 *Let p be a prime. A finite sum of prime clocks $[p,t_1] \oplus [p,t_2] \oplus \ldots [p,t_{l-1}] \oplus [p,t_l]$ is called a p-clock sum of length l if for each $1 \leq i \leq l$, the clock $[p,t_i]$ is a p-clock and the sum is non-repeating. The non-repeating condition implies $l \leq p$.*

Lemma 2 *Let p be a prime. A p-clock sum with length p has period 1. A p-clock sum with length l such that $1 \leq l < p$ has period p.*

PROOF. When $p = 2$, the 2-clock sum $[2, 0]$ has period 2 and the 2-clock sum $[2, 1]$ also has period 2. Recall that $[2, 0] \oplus [2, 1] = \overline{1}$. For the remainder of the proof, it is assumed that p is an odd prime.

Let $[p, t_1] \oplus [p, t_2] \oplus \ldots [p, t_{l-1}] \oplus [p, t_l]$ be a p-clock sum. When $l = p$, remark 7 implies that $[p, t_1] \oplus [p, t_2] \oplus \ldots [p, t_{l-1}] \oplus [p, t_l]$ has period 1. Lemma 1 and remark 5 imply that $[p, t_1] \oplus [p, t_2] \oplus \ldots [p, t_{l-1}] \oplus [p, t_l]$ has period p or period 1. The rest of this proof shows that $1 \leq l \leq p - 1$ implies that the p-clock sum cannot have period 1.

Thus, it suffices to show that $1 \leq l < p$ implies that $\big([p, t_1] \oplus [p, t_2] \oplus \cdots \oplus [p, t_l]\big)(m) \neq \big([p, t_1] \oplus [p, t_2] \oplus \cdots \oplus [p, t_l]\big)(m + 1)$ for some $m \in \mathbb{N}$. If needed, the p-clock sum may be permuted so that $[p, s_1] \oplus [p, s_2] \oplus \cdots \oplus [p, s_l] = [p, t_1] \oplus [p, t_2] \oplus \cdots \oplus [p, t_l]$ and the s_i are strictly increasingly. (Strictly increasing means $0 \leq s_1 < s_2 \ldots s_{l-1} < s_l \leq p - 1$.)

Case A. l is odd. If $s_l < p - 1$, then $\big([p, s_1] \oplus [p, s_2] \oplus \cdots \oplus [p, s_l]\big)(0) = \sum\limits_{i=1}^{l} s_i \bmod 2 \neq \sum\limits_{i=1}^{l} (s_i + 1) \bmod 2 = \big([p, s_1] \oplus [p, s_2] \oplus \cdots \oplus [p, s_l]\big)(1)$ because l is odd. Otherwise, $s_l = p - 1$. Set $s_0 = 0$. (The auxiliary index $s_0 = 0$ handles the case $s_{k+1} - s_k$ for all k such that $1 \leq k < l$.) Set $m = \max\big\{k \in \mathbb{N} : (s_{k+1} - s_k) \geq 2 \text{ and } 0 \leq k < l\big\}$. Since $s_0 = 0$ and $1 \leq l < p$, the pigeonhole principle implies m exists. Before the $\bmod 2$ step, the difference between $\sum\limits_{i=1}^{l} \big((s_i + l - m + 1) \bmod p\big)$ and $\sum\limits_{i=1}^{l} \big((s_i + l - m) \bmod p\big)$ equals l. Thus, $\big([p, s_1] \oplus [p, s_2] \oplus \cdots \oplus [p, s_l]\big)(l - m) \neq \big([p, s_1] \oplus [p, s_2] \oplus \cdots \oplus [p, s_l]\big)(l - m + 1)$.

Case B. l is even. Set $j = (p - 1) - s_l$. Before the mod 2 step, the sum $\sum\limits_{i=1}^{l} \big((s_i + j) \bmod p\big)$ differs from the sum $\sum\limits_{i=1}^{l} \big((s_i + j + 1) \bmod p\big)$ by an odd number. Thus, $\big([p, s_1] \oplus \cdots \oplus [p, s_l]\big)(j) \neq \big([p, s_1] \oplus \cdots \oplus [p, s_l]\big)(j + 1)$. $\square$.

Definition 8 *Let p be prime. Suppose the times are strictly increasing: that is, $s_1 < s_2 \cdots < s_l$ and $t_1 < t_2 \cdots < t_m$. Suppose $\max\{l, m\} \leq p$. Then p-clock sum $[p, s_1] \oplus \cdots \oplus [p, s_l]$ is distinct from p-clock sum $[p, t_1] \oplus \cdots \oplus [p, t_m]$ if $l \neq m$ or for some i, $s_i \neq t_i$.*

7-clock sum $[7, 0] \oplus [7, 2] \oplus [7, 3]$ is distinct from $[7, 1] \oplus [7, 2] \oplus [7, 3]$. Table 4 shows that these distinct 7-clock sums are not equal.

Table 4: Two distinct 7-clock sums that are not equal in Ω_2

Time	$[7,0]$	$[7,1]$	$[7,2]$	$[7,3]$	$[7,0] \oplus [7,2] \oplus [7,3]$	$[7,1] \oplus [7,2] \oplus [7,3]$
0	0	1	0	1	1	0
1	1	0	1	0	0	1
2	0	1	0	1	1	0
3	1	0	1	0	0	1
4	0	1	0	0	0	1
5	1	0	0	1	0	1
6	0	0	1	0	1	1
...						

Theorem 2 *For any* 3 mod 4 *prime p, if two p-clock sums are distinct, then they are not equal in Ω_2. The theorem also holds for $p = 2$.*

PROOF. The special case $p = 2$ can be verified by examining the second and third columns of table 3.

Let p be a 3 mod 4 prime. Assume p-clock sum $[p, s_1] \oplus \cdots \oplus [p, s_l]$ is distinct from p-clock sum $[p, t_1] \oplus \cdots \oplus [p, t_m]$. By reductio absurdum, suppose

$$[p, s_1] \oplus \cdots \oplus [p, s_l] = [p, t_1] \oplus \cdots \oplus [p, t_m]. \tag{3}$$

For each $s_i \in \{t_1, \ldots, t_m\}$, the operation $\oplus[p, s_i]$ in Ω_2 can be applied to both sides of equation 3. Similarly, for each $t_j \in \{s_1, \ldots, s_l\}$, the operation $\oplus[p, t_j]$ can be applied to both sides of equation 3. Since $(\Omega_2, \oplus)$ is an Abelian group, equation 3 can be simplified to $[p, s_1] \oplus \cdots \oplus [p, s_L] = [p, t_1] \oplus \cdots \oplus [p, t_M]$ such that $\{s_1, \ldots, s_L\} \cap \{t_1, \ldots, t_M\} = \emptyset$ and $M + L \leq p$.

Set $f = [p, s_1] \oplus \cdots \oplus [p, s_L]$. Apply $f\oplus$ to both sides of $[p, s_1] \oplus \cdots \oplus [p, s_L] = [p, t_1] \oplus \cdots \oplus [p, t_M]$. This simplifies to $f \oplus [p, t_1] \oplus \cdots \oplus [p, t_M] = \overline{0}$. Lemma 2 implies that $L + M = p$. Since $L + M = p$ and $\{s_1, \ldots, s_L\} \cap \{t_1, \ldots, t_M\} = \emptyset$ and p is a 3 mod 4 prime, remark 7 implies that $f \oplus [p, t_1] \oplus \cdots \oplus [p, t_M] = \overline{1}$. This is a contradiction, so $[p, s_1] \oplus \cdots \oplus [p, s_l]$ is not equal to $[p, t_1] \oplus \cdots \oplus [p, t_m]$ in Ω_2. □

Let $\mathcal{S}_l$ be the set of all p-clock sums of length l, where $1 \leq l \leq p$. There are $\binom{p}{l}$ distinct p-clock sums in each set $\mathcal{S}_l$. Set $G_p = \bigcup_{l=1}^{p} \mathcal{S}_l \cup \{\overline{0}\}$. For any $f, g \in G_p$, remark 6 implies $f \oplus g^{-1}$ in G_p. Thus, $(G_p, \oplus)$ is an Abelian subgroup of Ω_2. Set $B_p = \{0,1\}^p$. For any $a_1 \dots a_p \in B_p$ and $b_1 \dots b_p \in B_p$, define $a_1 \dots a_p +_2 b_1 \dots b_p = c_1 \dots c_p$, where $c_i = (a_i + b_i) \bmod 2$. $(B_p, +_2)$ is an Abelian group with 2^p elements. When p is a 3 mod 4 prime, define the function $\phi : G_p \to B_p$ where $\phi(\bar{0}) = 0 \dots 0 \in B_p$ and $\phi([p,t_1] \oplus [p,t_2] \oplus \dots [p,t_l]) = c_1 \dots c_p$ where $c_i = \big([p,t_1] \oplus [p,t_2] \oplus \dots [p,t_l]\big)(i)$. ϕ is a group isomorphism:

Theorem 3 *Let p be a* 3 mod 4 *prime. The subgroup G_p of Ω_2, generated by the p-clocks $[p,0], [p,1], \dots [p,p-1]$ has order 2^p and is isomorphic to $(B_p, +_2)$.*

PROOF. Theorem 2 implies ϕ is a group isomorphism. □

Table 5: The 5-clocks projected into Ω_2

Time	$[5,0]$	$[5,1]$	$[5,2]$	$[5,3]$	$[5,4]$	$[5,0] \oplus [5,1]$	$[5,2] \oplus [5,3] \oplus [5,4]$
0	0	1	0	1	0	1	1
1	1	0	1	0	0	1	1
2	0	1	0	0	1	1	1
3	1	0	0	1	0	1	1
4	0	0	1	0	1	0	0
5	0	1	0	1	0	1	1
...							

Theorem 2 does not hold when p is a 1 mod 4 prime. Table 5 shows $[5,0] \oplus [5,1]$ equals $[5,2] \oplus [5,3] \oplus [5,4]$ in Ω_2.

Theorem 4 *For any* 1 mod 4 *prime p, if two p-clock sums are distinct and their respective lengths L and M are both $\leq \frac{p-1}{2}$, then these two p-clock sums are not equal in Ω_2.*

PROOF. The proof is almost the same as the proof in theorem 2. The conditions $L \leq \frac{p-1}{2}$ and $M \leq \frac{p-1}{2}$ and the reduction $[p,s_1] \oplus \dots \oplus [p,s_L] \oplus [p,t_1] \oplus \dots \oplus [p,t_M] = \bar{0}$ leads to an immediate contradiction: $L + M \leq p - 1$ and $\{s_1, \dots, s_L\} \cap \{t_1, \dots, t_M\} = \emptyset$ means lemma 2 implies $[p,s_1] \oplus \dots \oplus [p,s_L] \oplus [p,t_1] \oplus \dots \oplus [p,t_M]$ has period p. □

Remark 9 *Let p be a* 1 mod 4 *prime. Let $f = [p, s_1] \oplus \cdots \oplus [p, s_l]$ for some $1 \leq l \leq \frac{1}{2}(p-1)$. Set $T = \{0, 1, \ldots, p-1\} - \{s_1, \ldots, s_l\}$. Now $T = \{t_1, \ldots t_m\}$, where $l + m = p$. Set $g = [p, t_1] \oplus \cdots \oplus [p, t_m]$. Then $f = g$ in Ω_2.*

PROOF. Since p is a 1 mod 4 prime, $\big(f \oplus g\big)(0) = \sum\limits_{0}^{p-1} k \bmod 2 = 0$ in $\mathbb{Z}_2$. When $k > 1$, the sum of the elements of $f \oplus g$ before projecting into Ω_2 is a permutation of the elements $\{0, 1, \ldots, p-1\}$. Thus, for all $k > 1$, $\big(f \oplus g\big)(k) = 0$ in $\mathbb{Z}_2$. This means $g = f^{-1}$. Lastly, $f = f^{-1}$ in Ω_2, so $f = g$ in Ω_2. □

Let p be a 1 mod 4 prime. Set $H_{p-1} = \bigcup\limits_{l=1}^{\frac{1}{2}(p-1)} \mathcal{S}_l \cup \{\overline{0}\}$. Observe that $|H_{p-1}| = \sum\limits_{l=1}^{\frac{1}{2}(p-1)} \binom{p}{k} + 1 = 2^{p-1}$. To verify that $(H_{p-1}, \oplus)$ is a subgroup of $(\Omega_2, \oplus)$, let $f, g \in H_{p-1}$. Since $g = g^{-1}$ in $(\Omega_2, \oplus)$, it suffices to show that $f \oplus g$ lies in H_{p-1}. If f or g equals $\overline{0}$, closure in $(H_{p-1}, \oplus)$ holds. Otherwise, $f = [p, s_1] \oplus \cdots \oplus [p, s_l]$ for some $1 \leq l \leq \frac{1}{2}(p-1)$ and $g = [p, t_1] \oplus \cdots \oplus [p, t_m]$ for some $1 \leq m \leq \frac{1}{2}(p-1)$. As mentioned before, the sum $f \oplus g$ may be reduced to $[p, s_1] \oplus \cdots \oplus [p, s_L] \oplus [p, t_1] \oplus \cdots \oplus [p, t_M]$, where $\{s_1, \ldots, s_L\} \cap \{t_1, \ldots, t_M\} = \emptyset$ and $L + M \leq p$. If $L + M \leq \frac{1}{2}(p-1)$, closure in $(H_{p-1}, \oplus)$ holds. Otherwise, if $L + M > \frac{1}{2}(p-1)$, remark 9 implies that there is a p-clock sum $h = f \oplus g$, where h's length is $p - (L + M)$ and $p - (L + M) \leq \frac{1}{2}(p-1)$.

Similar to the group isomorphism ϕ, define $\psi : H_{p-1} \rightarrow B_{p-1}$ such that $\psi(\overline{0}) = 0 \ldots 0 \in B_p$. For each p-clock sum in $\mathcal{S}_l$, where $1 \leq l \leq \frac{1}{2}(p-1)$, define $\psi([p, t_1] \oplus [p, t_2] \oplus \ldots [p, t_l]) = c_1 \ldots c_{p-1}$ where $c_i = \big([p, t_1] \oplus [p, t_2] \oplus \ldots [p, t_l]\big)(i)$. It is straightforward to verify that ψ is a group isomorphism onto B_{p-1}. The group isomorphism $\psi : H_{p-1} \rightarrow B_{p-1}$ leads to the following theorem.

Theorem 5 *Let p be a* 1 mod 4 *prime. The subgroup H_{p-1} of Ω_2, generated by the p-clocks $[p, 0], [p, 1], \ldots [p, p-1]$ has order 2^{p-1} and is isomorphic to $(B_{p-1}, +_2)$.*

Theorem 6 *For positive integer n and any function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, there exists a finite sum of prime clocks in Ω_2 that can compute f.*

PROOF. Euclid's second theorem implies there is a prime $p > 2^n$, where p is a 3 mod 4 or 1 mod 4 prime. Thus, theorem 3 or 5 completes the proof. □

Furthermore, finding a finite prime clock sum that computes f is Turing computable and there are efficient Turing computable algorithms that can decide whether a natural number n is prime [3].

4 Computing Boolean Functions with Prime Clock Sums in Ω_2

Let F_n denote the set of all Boolean functions in n variables. Formally, the set $F_n = \{f \mid f : \{0,1\}^n \to \{0,1\}\ \}$ and F_n contains 2^{2^n} distinct functions. For prime clock sums, it is convenient to think of $f \in F_n$ as a binary string of length 2^n, called the *truth-table* of f. Table 6 shows all 16 Boolean functions in F_2, their truth tables and corresponding prime clock sums that compute each function.

Consider $[p,s] \oplus [q,t]$ in Ω_2. The *first* 2^n *elements* of $[p,s] \oplus [q,t]$ refer to the bit string $\big([p,s] \oplus [q,t]\big)(0),\ \big([p,s] \oplus [q,t]\big)(1),\ \ldots,\ \big([p,s] \oplus [q,t]\big)(2^n - 1)$ of length 2^n. The first 2^n elements of $[p,s] \oplus [q,t]$ represent a Boolean function $f \in F_n$. In the general case, if $q_1, \ldots, q_L$ are primes, the first 2^n elements of $[q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_L,t_L]$ also represent a Boolean function $f_n \in F_n$. Consider the first 2^n elements of prime clock sum $[q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_L,t_L]$. Algorithm 1 computes the ith element of this truth table in F_n.

Algorithm 1: A Prime Clock Sum in Ω_2 Computes a Boolean Function

```
INPUT:   $i$
  set $r_1 = (t_1 + i) \bmod q_1$
  set $r_2 = (t_2 + i) \bmod q_2$
  . . .
  set $r_L = (t_L + i) \bmod q_L$
  set $y = (r_1 + r_2 + \cdots + r_L) \bmod 2$
OUTPUT:  $y$
```

The ith element of $\big([q_1,t_1] \oplus [q_2,t_2] \oplus \cdots \oplus [q_L,t_L]\big)$'s truth table is stored in the variable y when algorithm 1 halts. Algorithm 1 is presented in a serial form. Nevertheless, the computation of the L instructions `set` $r_k = (t_k + i) \bmod q_k$, where $1 \le k \le L$, can be computed in parallel when there is a separate physical device for each of these L prime clocks $[q_1,t_1]$, $[q_2,t_2]$ … $[q_L,t_L]$. Subsequently, the parity of y can be determined in a second computational step that executes a parallel add of $r_1 + r_2 + \cdots + r_L$, followed by setting y to the least significant bit of the sum $r_1 + r_2 + \cdots + r_L$.

As an alternative implementation of algorithm 1, when there is a more suitable physical device for prime clocks, the kth clock can compute the kth bit $b_k = \big((t_k + i) \bmod q_k\big) \bmod 2$ and then a parallel exclusive-or [16] can be applied to the L bits $b_1, b_2, \ldots, b_L$. In contrast, a gate-based Boolean circuit requires at least d computational steps where d is the depth of the circuit.

Table 6: $f_k : \{0,1\}^2 \to \{0,1\}$.

Boolean Function	Truth Table	Prime Clock Sum
$f_1(x,y) = 1$	1111	$[2,0] \oplus [2,1]$
$f_2(x,y) = 0$	0000	$[2,0] \oplus [2,0]$
$f_3(x,y) = x$	0011	$[2,1] \oplus [3,1]$
$f_4(x,y) = y$	0101	$[2,0]$
$f_5(x,y) = \neg x$	1100	$[2,0] \oplus [3,1]$
$f_6(x,y) = \neg y$	1010	$[2,1]$
$f_7(x,y) = x \wedge y$	0001	$[2,0] \oplus [3,0]$
$f_8(x,y) = x \vee y$	0111	$[2,0] \oplus [3,2]$
$f_9(x,y) = \neg x \vee y$	1101	$[3,0] \oplus [3,1]$
$f_{10}(x,y) = x \vee \neg y$	1011	$[3,1] \oplus [3,2]$
$f_{11}(x,y) = (x \wedge y) \vee \neg(x \vee y)$	1001	$[3,1]$
$f_{12}(x,y) = (x \vee y) \wedge \neg(x \wedge y)$	0110	$[3,0] \oplus [3,2]$
$f_{13}(x,y) = \neg(x \vee y)$	1000	$[2,1] \oplus [3,2]$
$f_{14}(x,y) = \neg(x \wedge y)$	1110	$[2,1] \oplus [3,0]$
$f_{15}(x,y) = \neg x \wedge y$	0100	$[3,0]$
$f_{16}(x,y) = x \wedge \neg y$	0010	$[3,2]$

The truth table for $\{0,1\}^2$ is ordered as $\{00, 01, 10, 11\}$.

EXAMPLE This example demonstrates 2-bit multiplication with prime clock sums, computed with algorithm 1. In table 7, for each $u \in \{0,1\}^2$ and each $l \in \{0,1\}^2$, the product $u * l$ is shown in each row, whose 4 columns are labelled by $\mathcal{M}_3$, $\mathcal{M}_2$, $\mathcal{M}_1$ and $\mathcal{M}_0$. With input i of 4 bits (i.e., u concatenated with l), the output of the 2-bit multiplication is a 4-bit string $\mathcal{M}_3(i)\ \mathcal{M}_2(i)\ \mathcal{M}_1(i)\ \mathcal{M}_0(i)$, shown in each row of table 7.

One can verify that, according to algorithm 1, prime clock sum $[2,0] \oplus [7,3] \oplus [7,4] \oplus [7,5] \oplus [11,10]$ computes function $\mathcal{M}_0 : \{0,1\}^2 \times \{0,1\}^2 \to \{0,1\}$. Similarly, $[2,0] \oplus [2,1] \oplus [3,0] \oplus [5,2] \oplus [11,0] \oplus [11,1]$ computes function $\mathcal{M}_1$. Prime clock sum $[5,0] \oplus [7,0] \oplus [7,2] \oplus [11,4]$ computes function $\mathcal{M}_2$. Lastly, $[2,1] \oplus [5,0] \oplus [11,1] \oplus [11,6]$ computes function $\mathcal{M}_3$.

Table 7: 2-Bit Multiplication.

Multiplication functions $\mathcal{M}_i : \{0,1\}^4 \to \{0,1\}$.

u	l	$\mathcal{M}_3$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_0$
00	00	0	0	0	0
00	01	0	0	0	0
00	10	0	0	0	0
00	11	0	0	0	0
01	00	0	0	0	0
01	01	0	0	0	1
01	10	0	0	1	0
01	11	0	0	1	1
10	00	0	0	0	0
10	01	0	0	1	0
10	10	0	1	0	0
10	11	0	1	1	0
11	00	0	0	0	0
11	01	0	0	1	1
11	10	0	1	1	0
11	11	1	0	0	1

Acknowledgement

Words cannot express the depth of my gratitude for Joanne and Haley.

References

[1] H. Aiken and G. Hopper. The Automatic Sequence Controlled Calculator, reprinted in B. Randell, ed., The Origins of Digital Computers. Berlin: Springer Verlag, 203–222, 1982.

[2] J. Bardeen and W. H. Brattain. The Transistor, A Semi-Conductor Triode. Physical Review, **74**, 230, July 15, 1948.

[3] Manindra Agrawal, Neeraj Kayal, and Nitin Saxena. Primes is in P. Annals of Mathematics, **160**, 781–793, 2004.

[4] A.W. Burks and A.R. Burks, The ENIAC: First General Purpose Electronic Computer, Annals of the History of Computing, **3**, 4, 310–399, 1981.

[5] A.W. Burks and A.R. Burks, The First Electronic Computer: The Atanasoff Story. Ann Arbor: Univ. of Michigan Press, 1988.

[6] T.W. Cusick and P. Stanica. Cryptographic Boolean Functions and Applications. Academic Press, 2009.

[7] P. Halmos, S. Givant. Logic as Algebra. MAA, 1998.

[8] Jack Kilby. Miniaturized Electronic Circuits. U.S. Patent 3,138,743. 1959.

[9] J.E. Lilienfeld. Method and apparatus for controlling electric currents. U.S. Patent 1,745,175: January 28, 1930. October 8, 1926.

[10] J.E. Lilienfeld. Device for controlling electric current. U.S. Patent 1,900,018: March 7, 1933. March 28, 1928.

[11] Carver Mead. Analog VLSI and Neural Systems. Addison-Wesley, 1989.

[12] Robert N. Noyce. Semiconductor Device-and-Lead Structure. U.S. Patent 2,981,877. 1959.

[13] M. Riordan, Lillian Hoddeson, and Conyers Herring. The invention of the transistor. Reviews of Modern Physics, vol. 71, no. 2, Centenary 1999. American Physical Society, 1999.

[14] Claude Shannon. The synthesis of two-terminal switching circuits. Bell Systems Technical Journal. **28**, 59–98, 1949.

[15] H. Vollmer. Introduction to Circuit Complexity. Springer, Heidelberg, 1999.

[16] Hao Yan, Liping Feng, Thomas H. LaBean and John H. Reif. Parallel Molecular Computations of Pairwise Exclusive-Or (XOR) Using DNA ”String Tile” Self-Assembly. J. Am. Chem. Soc., **125**, 47, 14246–14247, 2003.

[17] Alan Turing. Proposals for Development in the Mathematics Division of an Automatic Computing Engine (ACE), Report E882. NPL, 1945.

[18] Konrad Zuse. Patentanmeldung Z-2391, German Patent Office, 1941.

[19] Konrad Zuse. Der Computer mein Lebenswerk. Springer-Verlag, 1970.

MAXCUT and Variants of it

Hanno Lefmann

Fakultät für Informatik, TU Chemnitz, Germany

Abstract: Approximation algorithms for the MAXCUT problem and some of its variants are considered and their worst case qualities are analyzed.

1 Introduction

Let $G = (V, E)$ be a graph with vertex set V and edge set $E \subseteq [V]^2$. The MAXCUT problem asks for a given graph $G = (V, E)$ for a partition $V = V_1 \cup V_2$ of the vertex set into two classes V_1 and V_2 such that the number of edges $e \in E$ with one vertex in V_1 and the other in V_2 is maximum.

In general, for $\ell \geq 2$, the ℓ-MAXCUT problem asks for a given graph $G = (V, E)$ for a partition $V = V_1 \cup \cdots \cup V_\ell$ into ℓ classes such that the number of edges $e \in E$ with one vertex in V_i and the other in V_j, $i \neq j$, is maximum.

The MAXCUT problem is known to be $\mathcal{NP}$-hard, so one cannot expect deterministic polynomial time algorithms for solving it. Therefore, one is looking for *approximation algorithms* for solving such $\mathcal{NP}$-hard problems. These are deterministic polynomial time algorithms that produce a feasible approximate solution. The *approximation ratio* $AR_A(I)$ of an approximation algorithm A (for a maximization problem) for an input I is the quotient of the value $OPT(I)$ of an optimal solution and the value $A(I)$ of the solution produced by A

$$AR_A(I) = \frac{OPT(I)}{A(I)}.$$

In Section 2 we discuss approximation algorithms for the MAXCUT problem for arbitrary input graphs, and in Section 3 we restrict to the class of F-free input graphs, where F is a fixed graph. In Section 4 we consider the online situation for this problem, and we finish with some concluding remarks in Section 5.

2 Approximation

The set $N(v)$ of *neighbors* of a vertex $v \in V$ in a graph $G = (V, E)$ is the set of all vertices $w \in V$ with $\{v, w\} \in E$.

An easy approximation algorithm for MAXCUT for a given input graph $G = (V, E)$ works as follows. Take an arbitrary partition $V = V_1 \cup V_2$ into two classes of the vertex set of G. As long as there is a vertex v, say $v \in V_1$, that has more neighbors in its own class V_1 of the current partition than in the other class V_2, move this vertex to the other class V_2, and get the partition $V = (V_1 \setminus \{v_1\}) \cup (V_2 \cup \{v_1\})$. With every move to another class one gains at least one edge between the two classes, thus alltogether the number of moves to another class is at most $|E|$, i.e., this is polynomial. The algorithm stops, when no more moves to another class are possible. Then, at least $|E|/2$ of all edges are between the two classes. The approximation ratio of this algorithm is therefore at most 2.

For the ℓ-MAXCUT problem a similar strategy can be applied. Starting with an arbitrary partition $V = V_1 \cup \cdots \cup V_\ell$ into ℓ classes, move a vertex to another class V_j, in which it has a smaller number of neighbors than in its own class. The approximation ratio of this strategy is at most $\ell/(\ell - 1)$.

A *bipartite graph* $G = (V, E)$ allows a partition $V = V_1 \cup V_2$ with two classes such that all edges $e \in E$ have one vertex in V_1 and the other in V_2. One might suspect that the above mentioned approximation algorithm finds for bipartite graphs an optimal partition for the MAXCUT problem. However, for the cycle on four vertices and with four edges, if the initialization $V = V_1 \cup V_2$ is such that every vertex has one neighbor in its own class and another one in the other class, no move of a vertex to another class will be made, and we get a solution with two edges, where four edges is the optimum solution.

Using semidefinite programming, Goemans and Williamson [3] gave a randomized approximation algorithm for the MAXCUT problem with approximation ratio 1.14, which was later derandomized to a deterministic approximation algorithm with the same quality. It is also known from results by Håstad [4], that under the assumption $\mathcal{P} \neq \mathcal{N}P$ there is no approximation algorithm with approximation ratio at most 1.06.

A graph $G = (V, E)$ on n vertices is called *dense*, if it has at least $\Omega(n^2)$ edges.

For MAXCUT for dense graphs a PTAS is known, but there cannot be a PTAS for MAXCUT for arbitrary graphs unless $\mathcal{P} = \mathcal{N}P$. For dense graphs also strategies using the (weak) regularity lemma are known from the work of Frieze and Kannan [1].

3 Approximation for F-free Graphs

For a fixed graph F we consider F-free graphs, that do not contain F as a subgraph. Bipartite graphs are triangle-free, they do not contain a copy of the complete graph K_3 on three vertices. Among all triangle-free graphs on n vertices that graph with the largest number of edges is the Turán graph (extremal graph), which is the unique balanced, complete, bipartite graph, i.e., the sizes of the two classes V_1 and V_2 are as equal as possible.

More generally, for the complete graph $F = K_{\ell+1}$ on $\ell + 1$ vertices the unique extremal graph for $F = K_{\ell+1}$ on n vertices has ℓ vertex classes with sizes as equal as possible, and all possible edges between distinct classes.

The largest number of edges in an n-vertex F-free graph is denoted by $ex(n, F)$, thus $ex(n, K_3) = \lfloor n^2/4 \rfloor$.

The class of $K_{\ell+1}$-free graphs has the property of *stability*. This means, the more edges a $K_{\ell+1}$-free graph on n vertices has, the more similar is it to the extremal graph for $K_{\ell+1}$ on n vertices. This has been observed first by Simonovits [5] as follows.

Theorem 7 *For all $\delta > 0$ and for all $\ell \geq 2$ there exists $\varepsilon > 0$ such that for all $n \geq n_0$ the following holds. For every $K_{\ell+1}$-free graph $G = (V, E)$ on n vertices and with at least $ex(n, K_{\ell+1}) - \epsilon n^2$ edges there exists a partition $V = V_1 \cup \cdots \cup V_\ell$ such that at most δn^2 of its edges are contained in classes $V_1, \ldots, V_\ell$.*

Recently, Füredi [2] strengthened this as follows.

Theorem 8 *For all $\ell \geq 2$ and for all $n, t \geq 1$ the following holds. For every $K_{\ell+1}$-free graph $G = (V, E)$ on n vertices and with at least $ex(n, K_{\ell+1}) - t$ edges there exists a partition $V = V_1 \cup \cdots \cup V_\ell$ such that at most t of its edges are contained in classes $V_1, \ldots, V_\ell$.*

The *degree* $d(v)$ of a vertex v is the number of neighbors in G, i.e., the number of edges $\{v, w\} \in E$ in G.

The proof of Theorem 8 can be turned into a deterministic polynomial time algorithm. For an input $K_{\ell+1}$-free graph $G = (V, E)$ take a vertex v_1 of maximum degree. Let $N(v_1)$ be its set of neighbors and $NN(v_1)$ be its set of non-neighbors. Set $V_1 := NN(v_1) \cup \{v_1\}$ and $V := N(v_1) = V \setminus V_1$. Then, iterate this on the subgraph of G induced by the set $N(v_1)$ of neighbors of vertex v_1. The algorithm stops in the j'th step, $j \leq \ell$, when $N(v_j) = \emptyset$. This way we get a partition $V = V_1 \cup \cdots \cup V_\ell$ with the desired properties.

For triangle-free graphs this can be seen as follows. Let v be a vertex of maximum degree Δ in a triangle-free graph $G = (V, E)$ on n vertices with $|E| \geq ex(n, K_3) - t$. Let $V_1 := NN(v) \cup \{v\}$ and $V_2 := V \setminus V_1$. As G is triangle-free, there are no edges contained in V_2. Let c_1 be the number of edges $e \in E$ that are contained in V_1. Then we have

$$ex(n, K_3) - t \leq |E| \;=\; \sum_{v \in V_1} d(v) - c_1 \leq (n - \Delta) \cdot \Delta - c_1 \leq ex(n, K_3) - c_1,$$

which implies $c_1 \leq t$ as desired.

Thus, for the MAXCUT problem for triangle-free graphs on n vertices with at least $ex(n, K_3) - t$ edges one can find easily a partition $V = V_1 \cup V_2$ such that at most t of its edges are contained in class V_1 or V_2, hence this approximation algorithm has an approximation ratio of at most

$$\frac{ex(n, K_3) - t}{ex(n, K_3) - 2t},$$

which is $1 + o(1)$ for $t = o(n^2)$.

For $K_{\ell+1}$-free graphs and the ℓ-MAXCUT problem the analysis is similar.

4 Online Situation

So far we considered the *offline situation* for MAXCUT, where the whole input graph is known from the beginning. In the *online situation* an enemy chooses an input graph $G = (V, E)$ and fixes an order for presenting the vertices. Initializing $V_1 = V_2 = \emptyset$, vertices are presented one by one according to the fixed order, and an algorithm A has to decide whether a presented new vertex should be added to V_1 or V_2. This decision is final. Visible to the algorithm A are only those edges contained in the set of all presented vertices. The enemy can stop the process any time.

The *quality* $q(A)$ of an online algorithm A is the quotient of the value of an optimum solution for MAXCUT for the subgraph G' induced by the set of all presented vertices and the value of the by A achieved solution for G'.

If an algorithm puts a presented new vertex v to class V_1, if v has more (visible) neighbors in V_1 than in V_2, else in class V_2, then its worst case quality can be bounded from above by $q(A) \leq 2$.

Now consider a graph $G = (V, E)$ with order of vertices $v_1 < v_2 < \ldots < v_n < \ldots$. There is an edge between the vertices v_1 and v_2. Each vertex v_n, $n \geq 3$, has only the vertices v_1 and v_2 as neighbors. Any algorithm A will put vertices v_1 and v_2 (a) either into a single class (b) or into distinct classes. If A puts both vertices v_1 and v_2 into a single class, the enemy stops, and the quality is $1/0$, so infinite. If vertices v_1, v_2 are put into distinct classes, and the algorithm stops when vertex $v_n, n \geq 3$, has been presented, the achieved value for this MAXCUT problem is $n - 1$, while the optimum is $2n - 4$, hence the quality is

$$\frac{2n-4}{n-1} \longrightarrow 2$$

and can be arbitrary close to 2 with growing values of n. Is there in the online situation a corresponding example for triangle-free graphs?

5 Concluding Remarks

For the class of F-free graphs in the offline situation, where F is a complete graph, by using the concept of stability, one can achieve for MAXCUT problems easily good worst case approximation ratios by using a majority strategy.

In the talk we show further results for classes of F-free graphs as well as for some classes of F-free hypergraphs.

References

[1] Frieze, A., Kannan, R.: The regularity lemma and approximation schemes for dense problems, *37th FOCS*, pp. 12–20, 1996

[2] Füredi, Z.: A proof of the stability of extremal graphs, Simonovits's stability from Szemerédi's regularity, *J. Combinatorial Theory Series B* 115, pp. 66–71, 2015

[3] Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfyability problems using semidefinite programming, *J. of the ACM* 42(6), pp. 1115–1145, 1995

[4] Håstad, J.: Some optimal inapproximability results, *J. of the ACM* 48, pp. 798–869, 2001

[5] Simonovits, M.: A Method for Solving Extremal Problems in Graph Theory, Stability Problems, *Theory of Graphs (Proc. Colloq., Tihany, 1966)*, Academic Press, New York, pp. 279–319, 1968

Resilience and Mobility: Keeping Graphs Connected

Dimitri Samorukov

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract:
Applications running on top of p2p-networks consist often of distributed graph which is spreaded over several peers. In case of churn the graph nodes have to be relocated on other peers, which may lead to losing the connection to the adjacent graph-nodes, located on several other peers. The presented approach helps to keep the graph connected while the graph-nodes are migrating over the p2p-network. This is done by adapting forwarding pointer approach, known from mobile agent research, to the given problem of always connected graph. The presented method keeps the graph connected with probability of 0.8 up to 397500 migrations. It is independent of p2p-topology and provides a lock-freedom, which allows a simultaneous migration of adjacent graph-nodes.

Chaotic Central-frequency Modulation Effective for EMI Suppression in Switching Converters

Yuhong Song[1], Zhong Li[2], Junying Niu[1] and Wolfgang A. Halang[2]

[1]Faculty of Electronic and Information Engineering, Shunde Polytechnic, China

[2]Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: A chaotic signal has characteristics of fluctuant and continuous spectrum due to its pseudo-randomness. The frequency, corresponding to the largest peak on the power spectrum of a chaotic signal, is defined as the central frequency. The central frequency is determined by some interior parameter(s) of a system, which is verified through Chua's circuit. Further, taking the central frequency into account for electromagnetic interference (EMI) suppression in switching converters under chaotic duty modulation will be discussed. Simulations and experimental verification will be conducted on a DC-DC Buck converter.

1 Introduction

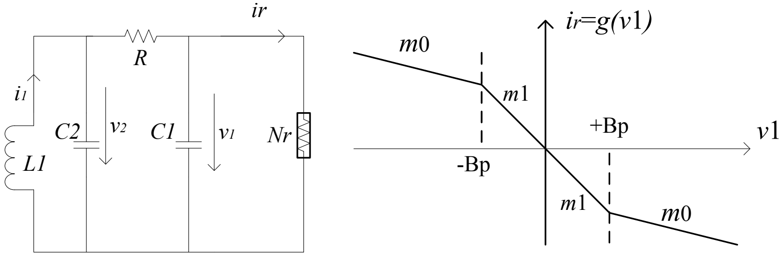
Switching converters are commonly applied in industry and daily life. However, the serious electromagnetic interference (EMI) is produced because of high change rates of current and voltage by applying PWM technology to control switching actions of transistors, which impairs other devices' performance and harms human being's health. Therefore, it is obliged to suppress EMI for application of switching converters.

A Chaotic signal is characterized by the pseudo-randomness and the continuous power spectrum. It is normally applied in chaos modulation to suppress EMI in switching converters by spreading the spectra of the output signals over the whole frequency band [1–3]. However, the chaotic signal is not the ideal wide spectrum signal [4], because its spectrum is not even with some peaks and a limited scope. **In order to conduct a quantitative analysis, define the central frequency of the chaotic signal as the specific frequency with the greatest peak located at the spectrum, and denote the central frequency as f_0.** The central frequency can be observed from the signal spectrum, on which the greatest peak appears. It is hardly concerned by the literatures whether the specific frequency affects the chaos application or not.

For a system, its EMI energy can be estimated according to the spectral amplitudes of the fundamental component and harmonics [1, 3], which is the basic of mathematical analysis and experimental test for EMI. This paper focuses the central frequency and its effect on EMI reduction. It is verified that the central frequency of the chaotic signal is decided by the interior parameter(s) of the system through the Chua's circuit. The further study is to discuss that the central frequency of the chaotic signal influences EMI reduction in the converter with chaos modulation. Simulation and experimental tests are given to complete the verification.

2 Chua's Circuit and the Central Frequency

The Chua's circuit is composed of a sine oscillator and a voltage-controlled nonlinear resistor [5], which includes three energy storing elements, an inductor and two capacitors. Figure 1 (a) shows the Chua's circuit, and i_1 is denoted as the current through the inductor L_1, v_1 and v_2 as the voltages across the capacitors C_1 and C_2. Figure 1 (b) presents the $v - i$ characteristic curve of the nonlinear resistor N_r with the outer line slope m_0, the internal line slope m_1 and the abscissas $+\mathrm{B_P}$, $-\mathrm{B_P}$ of the two breakpoints. The Chua's circuit is expressed by the following equations.



(a) Chua's circuit

(b) $v - i$ characteristic

Fig. 1: Chua's circuit and the characteristic of the nonlinear resistor

$$\begin{cases} \frac{\mathrm{d}v_1}{\mathrm{d}t} = \frac{1}{C_1}[(v_2 - v_1)G - f(v_1)], \\ \frac{\mathrm{d}v_2}{\mathrm{d}t} = \frac{1}{C_2}[(v_1 - v_2)G + i_1], \\ \frac{\mathrm{d}i_1}{\mathrm{d}t} = \frac{1}{L_1}v_2 \end{cases} \tag{1}$$

where $G = 1/R$, $f(v_1)$ is the current i_r through N_r, which is the piecewise linear function of v_1. It is expressed as:

$$f(v_1) = m_0 v_1 + 0.5(m_1 - m_0)[|v_1 + \mathrm{B_P}| - |v_1 - \mathrm{B_P}|]. \tag{2}$$

2.1 Unitary Processing

A unitary processing is made for Eqs. (1) and (2) [6]. Set the parameters as follows:

$$\begin{gathered} x = \frac{v_1}{\mathrm{B_P}}, y = \frac{v_2}{\mathrm{B_P}}, z = \frac{Ri_1}{\mathrm{B_P}}, \tau = \frac{t}{RC_2}, \\ \alpha' = \frac{C_2}{C_1}, \beta' = \frac{C_2 R^2}{L_1}, a = Rm_1, b = Rm_0. \end{gathered} \tag{3}$$

A set of dimensionless equations (4) are obtained for Chua's circuit.

$$\begin{cases} \frac{dx}{d\tau} = \alpha'(y - x - f(x)), \\ \frac{dy}{d\tau} = x - y + z, \\ \frac{dz}{d\tau} = -\beta' y \end{cases} \tag{4}$$

where

$$f(x) = bx + 0.5(a - b)[|x + 1| - |x - 1|]. \tag{5}$$

The expression (5) is a three-segment linear function, which has two slopes of a and b.

For the equations (4) and (5), simulations are conducted under the parameter set: $\alpha' = 9.0$, $\beta' = 14.87$, $a = -1.183$ and $b = -2.3114$. The evolving waveform of v_2 and the trajectory of $v_2 - v_1$ are shown in Fig. 2.

It can be seen that the Chua's circuit holds the chaos as long as the parameter set remains invariant. According to the equation (3), there is no effect on α', β', a and b when increasing or decreasing C_1, C_2 and L_1 simultaneously.

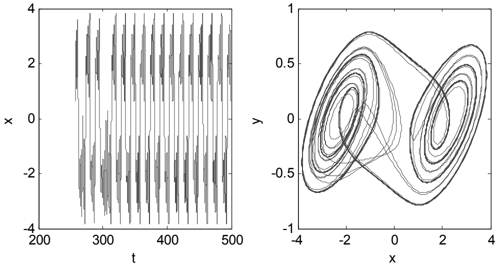


Fig. 2: The chaotic signal of Chua's circuit and the trajectory of $v_2 - v_1$

2.2 Central Frequency

The equation (5) can be described as a third-order linear polynomial function, which is the complex frequency domain as expressed [7],

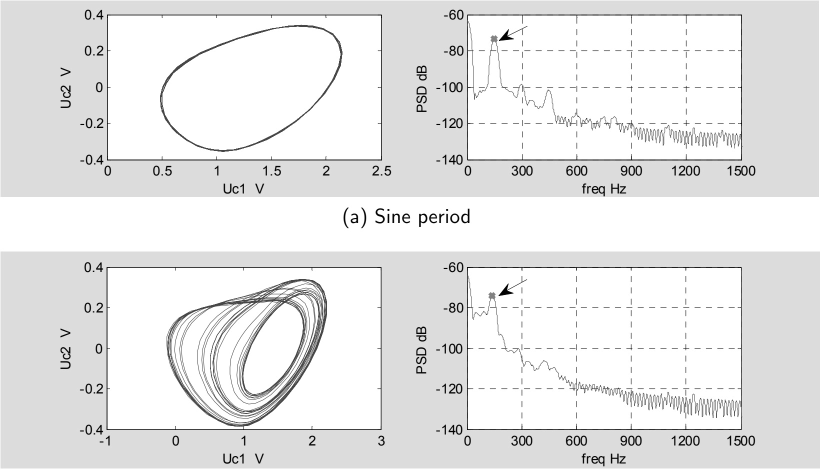
$$s^3 + q_2 s^2 + q_1 s + q_0 = 0 \tag{6}$$

where s is the complex frequency, q_0, q_1 and q_2 are decided by C_1, C_2, G, L_1 and $m_i (i = 0, 1)$. That is, $q_0 = \frac{G+m_i}{L_1 C_1 C_2}$, $q_1 = \frac{1}{L_1 C_2} + \frac{G+m_i}{C_1 C_2}$ and $q_2 = \frac{G}{C_2} + \frac{G+m_i}{C_1}$.

When the oscillator of the Chua's circuit runs in sine periodic status, the oscillating frequency remains constant, on which the energy concentrates. Hence, the spectrum of the oscillating signal is discrete. The component of the oscillating frequency is demonstrated in Fig. 3(a). According to the oscillating condition, the oscillating frequency of the Chua's circuit is estimated as[7],

$$f = \frac{1}{2\pi\sqrt{L_1 C_2}} \sqrt{1 + \frac{G L_1 m_i}{C_1}}, i = 0, 1. \tag{7}$$

Once the Chua's circuit runs into chaos, the energy is scattered at each component. However, the energy is relatively concentrated on the limited scope, which centers on the oscillating frequency, as displayed in Fig. 3 (b). Hence, the oscillating frequency is the central frequency, which is determined by the device parameters.



(a) Sine period

(b) Chaos

Fig. 3: The trajectory of Chua's circuit and the PSD of the signal

3 Effect of the Central Frequency on EMI Reduction

From the above analysis of the Chua's circuit, the central frequency is an attribute of the chaotic signal, which describes the evolving speed of the chaotic signal, and is the greatest spectral component of the chaotic signal.

From Eq. (7), different central frequency can be obtained by adjusting the parameters of C_1,C_2 and L_1 while keeping G and $m_i (i = 0, 1)$ constant [6]. Table 1 lists some parameter combinations and their respective central frequencies.

A DC-DC Buck converter is shown in Fig. 4 and the switching frequency f_s is configured as 48 kHz. The external chaotic signal v_{ch} is drawn from the voltage v_2 of Chua's circuit, which is injected into the pulse-width modulation module of the converter, realizing the chaotic duty modulation. With different central frequencies of chaotic signals, simulations are conducted to compare EMI reduction effects. The parameters of Chua's circuit are set up according to table 1. Taking the amplitude of the fundamental component as the object, tables 2 list the PWM under different central frequencies. Further, it can also be observed in tables 2 that FFT fundamental amplitudes vary along with the central frequency. The valley appears when $f_o \approx 0.5 f_s$, implying the optimal EMI suppression.

Table 1: Combinations of parameters and the relative central frequency

parameters	$f_o(\mathrm{kHz})$
$C_1 = 44\mathrm{nF}, C_2 = 396\mathrm{nF}, L_1 = 88\mathrm{mH}$	0.7
$C_1 = 22\mathrm{nF}, C_2 = 198\mathrm{nF}, L_1 = 44\mathrm{mH}$	1.4
$C_1 = 11\mathrm{nF}, C_2 = 99\mathrm{nF}, L_1 = 22\mathrm{mH}$	2.8
$C_1 = 5.5\mathrm{nF}, C_2 = 49.5\mathrm{nF}, L_1 = 11\mathrm{mH}$	5.6
$C_1 = 2.7\mathrm{nF}, C_2 = 25\mathrm{nF}, L_1 = 5.5\mathrm{mH}$	11.2
$C_1 = 2.45\mathrm{nF}, C_2 = 22.7\mathrm{nF}, L_1 = 5\mathrm{mH}$	12.5
$C_1 = 1.70\mathrm{nF}, C_2 = 15.8\mathrm{nF}, L_1 = 3.5\mathrm{mH}$	17
$C_1 = 1.45\mathrm{nF}, C_2 = 13.5\mathrm{nF}, L_1 = 3\mathrm{mH}$	20.5
$C_1 = 1.20\mathrm{nF}, C_2 = 11.2\mathrm{nF}, L_1 = 2.5\mathrm{mH}$	24
$C_1 = 0.95\mathrm{nF}, C_2 = 9.0\mathrm{nF}, L_1 = 2\mathrm{mH}$	29
$C_1 = 0.80\mathrm{nF}, C_2 = 7.65\mathrm{nF}, L_1 = 1.7\mathrm{mH}$	34.57
$C_1 = 0.76\mathrm{nF}, C_2 = 7.20\mathrm{nF}, L_1 = 1.65\mathrm{mH}$	36

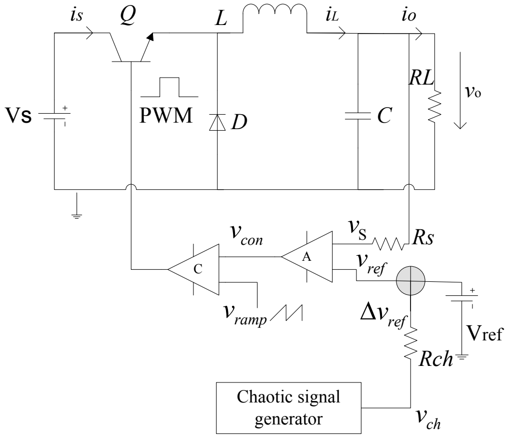
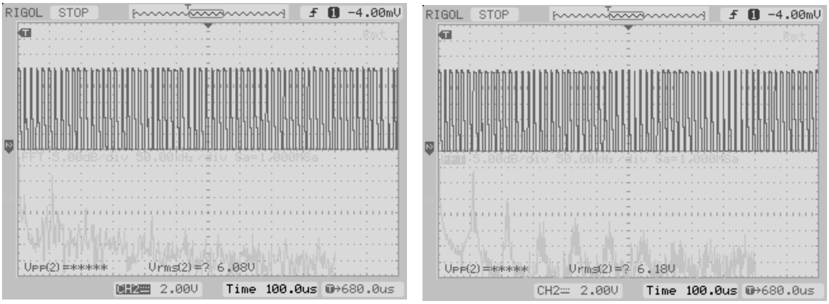


Fig. 4: The schematic diagram of chaotic duty modulation

Table 2: The fundamental component of PWM FFT (unit:dB)

f_0 (kHz)	0.7	1.4	2.8	5.6	11.2	12.5	13.5	17	20.5
PWM FFT	18.10	18.17	18.38	18.04	17.93	17.66	17.77	17.52	17.36
decrement	1.12	1.05	0.84	1.18	1.29	1.56	1.45	1.70	1.86
f_0 (kHz)	24	29	34.5	36	40	48.5	54	63	no chaos
PWM FFT	17.34	15.78	17.11	17.57	17.67	18.55	19.32	19.41	19.22
decrement	1.88	3.44	2.11	1.65	1.55	0.67	-0.1	-0.19	

Assuming $RL = 1$ Ohm, $L = 1$ mH, $C = 470$ uF, $\mathrm{V_L} = 0$ V, $\mathrm{V_U} = 3$ V, $\mathrm{V_s} = 25$ V, $\mathrm{V_{ref}} = 2.4$ V, $R_{ch} = 10$ k Ohm, and the switching frequency $f_s = 25$ kHz, the scheme in Fig. 4 is realized with hardware. Chaotic signals with $f_o = 24$ kHz and $f_o = 12.5$ kHz are respectively produced by adjusting the parameters of C_1, C_2 and L_1 in Chua's circuit according to Table 1. PWM waveforms and their spectra are shown in Fig. 5. It can be observed that the amplitudes of the fundamental and harmonics are reduced more when $f_o = 12.5$ kHz than when $f_o = 24$ kHz.



(a) $f_o = 12.5\,\mathrm{kHz}$ (b) $f_o = 24\,\mathrm{kHz}$

Fig. 5: The experimental waveforms for switching pulse

4 Conclusions

A chaotic signal characterizes the central frequency over its spectrum, which is testified through the Chua's circuit in this study. The further research proves that the central frequency makes an effect on the EMI suppression in switching converters with chaotic duty modulation. Simulations and experiments verifies that it is optimal when the central frequency of chaotic signal is close to one half of the switching frequency, which provides the basis or reference to apply chaotic signal in chaos modulation.

References

[1] Balestra, Michele and Lazzarini, Marco and Setti, Gianluca and Rovatti, Riccardo: Experimental performance evaluation of a low-EMI chaos-based current-programmed DC/DC boost converter, *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, 1489–1492, 2005

[2] Amini, AA and Nazarzadeh, J: Improvement behavior and chaos control of buck converter in current mode controlled, *IEEE Int. Conference on Industrial Technology*, 1–6, 2008

[3] Li, Hong and Li, Zhong and Zhang, Bo and Tang, Wallace KS and Halang, Wolfgang: Suppressing electromagnetic interference in direct current converters, *Circuits and Systems Magazine, IEEE*, 9, 4, 10–28, 2009

[4] Ruan, Lixin and Chen, Kangsheng: Chua's circuit chaotic signal spectrum distribution characteristics and its application in the circuit design, *Journal of circuits and systems*, 3, 1, 8–13, 1998

[5] Jackson, LB and Lindgren, AG and Kim, Y and others: A chaotic attractor from Chua¡s circuit, *IEEE Transactions on Circuits and Systems*, 31, 12, 1984

[6] Kennedy, Michael Peter: Three steps to chaos. II. a Chua's circuit primer, *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, 40, 10, 657–674, 1993

[7] Elwakil, Ahmed S and Kennedy, Michael Peter: Improved implementation of Chua's chaotic oscillator using current feedback op amp, *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, 47,1, 76–79, 2000

Secure Filtering Techniques for Instant Messaging

Günter Fahrnberger

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: Many scientific disquisitions have proposed ways to sustain data security in outsourced unconfident or semiconfident domains. These for secure numeric calculations try to explore feasible homomorphic functions. Most of those for secure string computations tackle issues in enciphered databases and keyword search in encrypted documents, mainly by dint of trapdoor functions. Both function types – homomorphic and trapdoor ones – form the heart algorithms of blind computing. Blind computing indicates that a cloud application computes ciphertext data without becoming aware of the meaning of input, output, and intermediate results.

Despite the bulk of scholarly publications about blind computing, the problem has not been comprehensively solved so far where an untrusted IM (Instant Messaging) platform must securely sift passing messages, especially those that are addressed to children. A solution does not only require an amalgamation of a searchable encryption scheme and a topic detection technique. Moreover, only offensive parts shall be blocked rather than the entire encircling message.

Currently, there exist some concepts about chat and IM filters, a whole slew of topic detection methods, and a plurality of papers about keyword detection in searchable encryption schemes. A secure IM filter as a combination of these research threads has not been engineered until now, much less a needful cryptosystem to safely find and delete bad words in instant messages. Therefore, this contribution recaps an accomplished academic project for the development and implementation of such a sifter system, which regards the security targets authenticity, integrity, privacy, and resilience.

1 Blind Computing

The outsourcing of own computing power to external places shifts (the responsibility of) the maintenance efforts of the outsourced facilities to a service provider and has been a valid strategy since the commercial use of computers [5]. It began with central computations in mainframes and was supplemented with the hosting of foreign IT-hard- and software. The will to commercialize less utilized or even idle computing resources and the possibility of virtualized hardware have driven the emergence of the topical cloud computing business with its three major service models *Infrastructure as a Service (IaaS)*, *Platform as a Service (PaaS)*, and *Software as a Service (SaaS)*.

These days, private individuals and, particularly, enterprises have to administer and process huge masses of data [7]. Either they house and compute their binary goods in their private computer centers respectively in their home workstations, or they resort to private, public, or hybrid cloud computing. For private users, on the one hand, it is common practice to have their files and electronic mail accounts outsourced in public clouds due to comfortable accessibility from nearly all over the world [8]. On the other hand, organizations mainly benefit from low or even no initial costs and standing expenses beside the ability to save maintenance efforts, because almost all cloud providers offer their resources by means of the pay-as-you-use model [5]. Especially, many start-up companies cannot predict their commercial growth and size in future despite of well-prepared business plans [4]. Therefore, it is hard for them to scale their necessary IT-infrastructure optimally. In case of under-sizing, the risk of crucial outages due to less redundancy and shortage of resources is high. On the contrary, an over-sized infrastructure leads to unused resources. Crucial outages may harm expected earnings, and fallowed resources cause unnecessary capital and operational expenditures. Clouds have remedied such optimization problems, because they offer the required scalability mostly coupled with the pay-as-you-use model.

In-house solutions let their owners know where their data reside, but redundancy through geographical in-house replicas and scalability lack or demand expensive expansions [7]. Cloud computing offers the opposite with great flexibility and resilience, but this mitigation comes along with a typical security problem that is evident for all external operating solutions [4]. Compared to an in-house solution, an ordinary cloud user does not have a lock on own applications and confidential data anymore, because they are physically dislodged at a more or less unknown data center where they are only protected by a generic non-disclosure agreement at best. Thence, detention of sensitive data as well as reading or writing activities on them become nontransparent and, thus, probably dangerous [7].

Daily news about exploitations of sensitive data in public clouds let the private users' security awareness upswing awhile before it reverts to its original level, or even worse, people do not fear for the security of their data because they suppose that their data do not arouse anyone's interest [8]. Even the latter would immediately change their opinion if somebody successfully looted their bank account due to their stored plaintext credentials in an online folder. Companies usually act much more carefully, because lost or exploited business data could entail the complete ruin of reputation or even existence of a firm. An organization will not make use of a public cloud as long as its personnel assess uncertainty for its valuable data there. Unfortunately, wrongdoing corporations take advantage of their prudent background about IT security and the users' ignorance of equivalent ken to exploit user data for entrepreneurial purposes. Ironically, these firms characterize such data abuse as marketing action because the illegally observed user habits permit them to proffer their products more purposively. Albeit such shady actors do not behave viciously, they can easily cause crucial security breaches if they ignore or negate any prophylactic measures or responsive countermeasures against genuine villains.

A decision maker has two options available. Either they entirely confide in a public cloud, or they completely suspect it. On account of this, basically, trust represents a dichotomous property between two objects. Some authors allude to semi-honest [13] or semi-honest-but-curious [2] parties which are only partially trusted. The airy way is to eschew cloud resources at all, but in spite of security risks, they often promise luring commercial advantages [8].

Altogether, the resulting challenge consists in the maximum yield of the advantages of the in-house and of the cloud paradigm accompanied by the eradication or mitigation of the disadvantages [7]. An imaginable approach could focus on costly transformations of the own physical IT-infrastructure into a private cloud to achieve all benefits. The cheaper way is taken by bringing (convenient) security to a public cloud. Security poses as a coarse term and needs further specification. Privacy, integrity, authenticity, and resilience represent intersubjective desirable IT security goals that shall be maintained at all.

Obviously, a cloud user and owner of sensible data needs to engage technical measures to firmly attain these IT security objectives [4].

For the safe bidirectional transport of sensible data between clouds and their users, many cryptosystems have been proposed so far. As mentioned above, clouds cannot be regarded as trustworthy environments, which makes transport ciphering insufficient and requires further efforts. For this reason, a policy maker can consider to involve a public cloud in such a way that an interloper cannot figure out what the cloud stores or computes [8].

Numeric or textual data intended just to be stored in the cloud without any reading or writing computations on them can be easily protected by employing one of the many available sorely approved-as-secure end-to-end cryptosystems [5].

Furthermore, there have been constant inquiries to access and modify encrypted data directly with the so-called blind computing paradigm [3]. Blind computing indicates that a cloud application computes ciphertext data without becoming aware of the meaning of input, output, and intermediate results [7].

For (keyword) searches on encrypted texts, there already exist useful SESs (Searchable Encryption Schemes) in masses [5]. Most of them tackle issues in enciphered databases and in encrypted documents, mainly by converting (key)words into characterizing hash values with the aid of parameterizable (trapdoor) hash functions (see definition 9) [7, 10].

Definition 9 *Let Σ denote an alphabet, let $K \subseteq \Sigma^*$ denote a key space, let $m \in \mathbb{N}$ denote a block size, let $H : \Sigma^* \times K \rightarrow \Sigma^m$ denote a parameterizable (trapdoor) hash function, let $k \in K$ denote a secret key, let $v \in \Sigma^*$ denote a plaintext document, let $E(v,k) : \Sigma^* \times K \rightarrow \Sigma^*$ denote an encryption function that does not only encipher v but additionally attaches the hash values of contained character strings by dint of H to its output, let $u \in \Sigma^*$ denote a sought plaintext character string, let $F(v,u) : \Sigma^* \times \Sigma^* \rightarrow \{false, true\}$ denote a search function that decides if v contains u, and let $G(E(v), H(u)) : \Sigma^* \times \Sigma^m \rightarrow \{false, true\}$ denote a homomorphic search function that decides if v contains u.*
Then the homomorphic function $G(E(v), H(u))$ forms a searchable encryption scheme if $(\forall v)(\forall u)\ G(E(v), H(u)) = F(v,u)$.

Blind operations on numerical values require the employment of disguising techniques or of homomorphic cryptosystems [8, 10].

While existent homomorphic encryption schemes will play an important role in future, they do not support calculations on encrypted character strings [3, 4].

Character strings are sequences of symbols from an alphabet Σ (also known as a character set). Although character strings can be represented by positive binary, octal, decimal, or hexadecimal numbers, a string function $H() : \Sigma^* \rightarrow \Sigma^*$ can neither be substituted by an arithmetic function $C() : \mathbb{N} \rightarrow \mathbb{N}$ nor by a homomorphic arithmetic function $G(E()) : \mathbb{N} \rightarrow \mathbb{N}$, so that $D(G(E()) = C()$ if $E() : \mathbb{N} \rightarrow \mathbb{N}$ denotes an encryption function, and $D() : \mathbb{N} \rightarrow \mathbb{N}$ denotes a decryption function [8]. From this it follows that representative codings of strings can be only used for identification. Their numerical values do not designate any rank or quantity [3, 4].

The use of trusted third parties, multiple parties, or cryptographic hardware has led to appropriate solutions, but obviously all of them depend on substantial hardware efforts [5]. Reasons, like the reduced or lost flexibility of clouds, higher costs, or the unwillingness of integration by the cloud providers, make the use of additional hardware unattractive, in particular for operators of secure IM services [11].

On this account, the proximate section 2 delivers insight into the history and hazards of IM.

2 Instant Messaging

Communication has always accompanied group-living beings as their tool to consign warnings of threats, signs of water or nutrition (yet edible or intended for hunting), and pairing advances [6]. Beyond that, higher developed creatures (like humans) also communicate to pass down knowledge and to entertain themselves mutually. Beside impersonal, undirected, and non-interactive entertainment (like television or theater), folks gratify their sensation mongering by exchanging past or imminent blatant happenings in their environment by a variety of messages. Phonemes started out as the first message bearer, complemented through pictographs. The evolution and bidirectional translation abilities of literary languages steadily endorsed the communication of scattered tribes more and more. Letters set out as the precursor of written messages. The emergence of electrical signal transmission amplified letters to telegrams and personal conversations to telephone and video calls. The digital age let dedicated protocols arise for the transfer of instant messages between chatters. Nowadays, we experience the convergence of acoustic and visual contents through all-round multimedia-facilitating Internet-protocols that convey them within common transmission channels and paths between two or more parties in subsecond timescale. Originally, users joined virtual chat rooms with adjustable anonymous nicknames instead of concrete authentication credentials.

Until today, instant messengers have ousted classic chats, fostered dialogs and multiple-user conversations, and respected present-day safety requirements by extorting fitting login data from the registered users in order to provide them communication with their (confirmed) virtual contacts. As a consequence, instant messaging now represents the most prominent way to transact communication between two or more people [9]. Luckily, on the one hand, chatters in democracies can freely express their opinions in instant messages without needing to fear political persecution and punishment. Unfortunately, on the other hand, all IM mediums suffer of two chief classes of vulnerabilities if evildoers can exploit this freedom because appropriate countermeasures have been neglected or even desisted [6].

The first kind of vulnerabilities concerns the wrapped contents of messages [6]. Boorish or even vicious talkers or chatters fabricate undesirable dubious utterances. Manifold factors cause the generation of annoying expressions. Not only but especially children are swamped with harassing sentences due to their little experience of life. Tokunaga corroborated this anxiety with the fact that more than 97% of youths in the United States are associated with Internet for some reason and, on account of that, represent potential victims [14]. Undesirable consequences from such associations are at an alarming rate, and major parts of such consequences are risky offenses against children and teenagers [11]. Least severe but already remarkable may be imprudent oral or written online dialogs between a child and, for instance, a neighbor kid that drops some unseemly cusses [6]. The underage listener or reader adopts swearwords either subliminally or consciously. The problem aggravates and drifts to bullying if conversations entail insult at the recipient's end rather than changing their habitual language use. Cyberbullying as the online version of bullying recently has been receiving a fair amount of attention [11]. It has evolved as a buzzword for deliberate repeated harm or harassment in online communication media and occurs in variously severe sorts [9]. Already undesired online contacting (via instant messengers) can be considered as cyberstalking. Flaming, the use of assaultive or rude language in heated or intense instant messages, implies hostile and insulting interaction among humans. Cyberharassment means offensive instant messages targeted at one or several individuals. Cyberstalking, flaming, and cyberharassment exemplify just three kinds of cyberbullying. In case of children and teens, cyberbullying effects are of serious concern and could paralyze future generations [11]. An investigative study on cyberbullying uncovered a fact that approximately 20–40% of youths encountered cyberbullying at least once in their life as of 2010 [11, 14]. The rate of cyberbullying is on its nurture and needs immediate attention [11]. Fatal cases of cyberbullying (particularly against teenagers) with suicide of the aggrieved party emphasize not to underestimate this type of oppression [6]. Further exacerbation incurs if ostensible smalltalk just aims at criminal objectives. Felons intend to win the victim's confidence in order to exploit them sexually or monetarily. Usually, culprits start with the broadsword and create huge masses of contact requests, mainly in written form via junk mails or friendship requests in social media. Responders are treated with the scalpel, which means that they become entangled in intensive seesaws until the villain succeeds or (unfortunately rarer) gets seized.

Home office task forces for child protection over the Internet were designed to train parents and children on sharing out no personal information or passwords, arranging no face to face meetings, and ignoring messages that make them feel threatened or uncomfortable [11, 15, 17]. Blocking the causing IM accounts seems to be the most obvious technical approach to inhibit further affronts, but nothing can prevent the miscreants to spawn new accounts, (re)inspire the mobbed sufferers with trust, and continue to afflict them [9]. A more promising resort would be a text filter that inspects all instant messages addressed to cyberbullying-endangered IM receivers. The sifter either cleans objectionable messages or drops them before their consignees get hold of them. Of course, adult addressees respectively legal guardians of minor recipients must consciously agree to such a screening rather than being forced to it. If they have decided to employ a text sieve, then it shall always be on duty irrespective of their location and used IM client application. Aside from location-independence, they anticipate a sieve that also includes colloquialisms, vogue terms, and cutting-edge swearwords in its decisions. These two requirements motivate a centralized, updatable screening solution (best directly coupled with the IM core platform) with proper resilience rather than a trivial client-side approach.

At this point, the second ilk of vulnerabilities arises. Public clouds usually shelter IM core platforms due to their demand of scalability for the incalculable user growth. Thus, affiliated screeners would also be located in public clouds. Cloud providers, intelligence agencies, and even the IM operators themselves have a field day retrieving private data out of instant messages for individual and big data analyses, to modify, or to stash them away. If IM users knew the risk for their confidential data in cloud services, they would never opt for (online filtering of their) instant messages. Hence, a prudent sifting design (that also abets the IT security goals authenticity, integrity, privacy, and resilience) must do the trick.

How do examples for violations of these targets for voice and text broadcasts look like [11]? A covertly impersonating speaker or writer harms the authenticity. A dork or killjoy who adds, modifies, or suppresses information during Chinese whispers offends the integrity, just as deliberately carving out voice sequences of a phone call in a switchgear or manipulating instant messages in a proxy. Wiretaps intentionally destroy the privacy. Eventually, overwhelming a communication service with a DoS (Denial of Service) attack through a single entity or even a DDoS (Distributed Denial of Service) attack through a herded bot-net firmly defeats its resilience. Historically, about 90% of computer security research delved into privacy and integrity problems, about 9% dealt with authenticity issues, and approximately 1% only addressed shortcomings in resilience [1]. Actual strikes and companies' defense expenses tend to be the other way round: more funds flow into the retention of resilience than into the other three security aims together [1].

The remainder of this document recapitulates methods that merge the insinuated challenge of blind computing on character strings in section 1 and an approach against both described ilks of vulnerabilities to instant messaging in section 2.

Section 3 treats SecureString 1.0 as the first approach for blind computing on encrypted character strings that perpetuates their secrecy even during modifications on them [3].

Section 4 dedicates itself to the advanced successor SecureString 2.0 that resolves emerged functional and security-relevant defects of SecureString 1.0 [4, 5, 7].

Section 5 embraces SecureString 3.0 that avoids any ciphertext repetitions and, in addition, scores by bearing up all IT security demands (authenticity, integrity, privacy, and resilience) [8, 10].

Section 6 exhibits a novel conception of a secure IM sieve based on whitelisting [9].

Section 7 concludes this scholarly piece by analyzing and comparing the IT security of all introduced methods.

3 SecureString 1.0 – The First Approach for Blacklisting

SecureString 1.0 was the first blacklisting approach to overcome extra hardware resources with a pure software solution [3]. It belongs to the kind of cryptosystems for blind computing on nonnumerical data. It bases upon a topical underlying symmetric cryptosystem and polyalphabetical encryption. The ciphering scheme encrypts each n-gram (character string of length n) of a word together with the beginning position of the n-gram within the word. Thereby, the start index of each n-gram stipulates the applied alphabet (alphabet = encryption transformation) on it. SecureString 1.0 brings the ciphertext n-grams out of sequence after their encryption without losing the possibility to operate on them because their order can be restored after their decryption through their enveloped position information.

Due to the finiteness of character string lengths that can be fully supported through querying and replacing operations, every SecureString 1.0-object intendedly contains exactly one word. Disadvantageously, this discloses the string boundaries and abets repetition pattern attacks on them.

4 SecureString 2.0 – The More Flexible Successor

The detection of these considerable limitations in SecureString 1.0 ended up in the improved version 2.0 of SecureString [4, 5, 7]. Among other improvements, the succeeding SecureString 2.0 overcame this limitation to permit an arbitrary number of cohered words per SecureString 2.0-object. In detail, SecureString 2.0 heads for monoalphabetical (substitutional) encryption within each character string, but every utilized alphabet must not become effective for more than one word. Normally, each encryption transformation depends on a dedicated key, but SecureString 2.0 enciphers each character of the same plaintext character string together with an identical salt (salt = arbitrary nonce). If the salt always transmutes from word to word, the encryption transformation also changes from word to word, even in case the same key would be employed for all character strings. Automatic salt updating [1, 6] is applicable, which means that each salt serves as input for a hash function that outputs the salt for the encryption of the successive plaintext word. This non-size-preserving behavior of SecureString 2.0 entails the advantage that the decryption scheme can simply ignore salts rather than must care about them. Optionally, for flow control, the salts can be shortened to make room for an appended sequential number that becomes verified during the decryption scheme.

The most recent treatise about SecureString 2.0 suggested to develop an advancement of the cryptosystem that converts every arbitrary plaintext into pangram ciphertext and, thereby, impedes any recurrences of ciphergrams [7].

5 SecureString 3.0 – The Abolishment of Ciphertext Repetitions

On these grounds, SecureString 3.0 as the next generation of the cryptosystem hardened SecureString 2.0 with better authenticity, integrity, and privacy [8, 10]. As its predecessors SecureString 1.0 and SecureString 2.0, SecureString 3.0 poses a non-size-preserving cryptosystem. The encryption scheme does nothing else but to salt each individual plaintext character of a string, i.e. to append an arbitrary character sequence, and to encipher the amalgamation of plaintext character and salt with a contemporary high-performing cryptosystem in ECB (Electronic CodeBook) mode [12]. The encryption algorithm may prolong as many characters with the same salt as no ciphertext repetition appears. Therefore, before it scrambles a dedicated character with an equal salt twice, it must switch to another salt and, thereby, defuses the hazardousness of repetitions despite the usage of ECB mode. Automatic salt updating by inputting the recent salt in a (trapdoor) hash function denotes an easy and secure alternative to obtain a fresh salt from an RNG (Random Number Generator) [1]. This kind of salt production does not only perpetuate backward security, i.e. no salt can be concluded from its successors, but also lets a TTPG (Trusted Third Party Generator) prepare homomorphic computations for an untrustworthy CSP (Communication Service Provider) just with a starting salt per cohesive text body.

To make a long story short, SecureString 3.0 achieves non-repetitive ciphertexts with controlled randomness attained through hashed salts in lieu of haphazardly compassed ciphertexts through the contents of foregoing blocks. SecureString 3.0 allows feasible secure blacklisting of instant messages.

6 Secure Whitelisting of Instant Messages

Blacklisting connotes the search and eradication of unsolicited terms. Unfortunately, already explicit phrases in spaced form can circumvent any blacklist filters, not to mention resourceful fake character insertions and uppercase-lowercase-combinations [16]. For this purpose, a secure whitelisting technique emerged that scans instant messages securely and only forwards those with solely harmless words to their destinations. The novel creation amalgamates the following six principles.

White- instead of Blacklisting to impede circumvention: Only instant messages with merely approved words inclusive of their declined or conjugated variants become delivered to their intended targets.

Secret sharing between IM relay and IM filter to maintain privacy: An IM relay only becomes aware of the source and the destination accounts of instant messages, but not of their plaintext content. An IM filter just processes a huddle of addressless character strings symmetrically encrypted with the secret key of the IM relay.

PKI (Public Key Infrastructure) to sustain anonymity and authenticity: Alice separately encrypts an accrued whole plaintext instant message and its bulk of individual words in a hybrid ilk, i. e. she symmetrically enciphers the instant message with a random secret key and the heap of its individual words with another secret key before she veils the secret keys with the public key of Bob. Hybrid cryptography joins the velocity and unlimited input length of symmetric ciphering and the benefit of unequal keys for encryption and decryption through asymmetric cryptographic functions. Moreover, Alice convinces Bob of her authenticity by signing a hash value of the ciphertext instant message with her private key. Altogether, Alice remains anonymous for the sieve and can be sure at the same time that merely the possessors of the appropriate private keys can confidently decrypt her ciphertexts.

Training mode to enforce resilience: Bob can covertly recommend vocables to the IM filter which should be included in the whitelist in future or excluded from there.

One-time secret keys to assure privacy: Alice and Bob draw on a unique arbitrary secret key for the symmetric encipherment of each instant message respectively on another one for the symmetric scrambling of a batch of individual words to reach nondeterministic encryption and to avert collision attacks which would be a consequence of ciphertext repetitions.

Onionskin paradigm to warrant authenticity, integrity, and privacy: A topical hybrid cryptosystem between all involved components ensures an extra umbrella against menaces during data transport.

Table 1: Security Analysis

Technique	**Authenticity**	**Integrity**	**Privacy**	**Resilience**	**Payload Conformity**
No Encryption	✗	✗	✗	✗	Blacklisting Whitelisting ✓
Transport Encryption (AES-128)	✗	✗	✗	✗	Blacklisting Whitelisting ✓
SecureString 1.0 (Section 3)	✗	✗	✗	✗	Blacklisting ✗
SecureString 2.0 (Section 4)	✗	✗	✓	✗	Blacklisting ✗
SecureString 3.0 (Section 5)	✓	✓	✓	✓	Blacklisting ✗
Secure Whitelisting of Instant Messages (Section 6)	✓	✓	✓	✓	Whitelisting ✓

7 Conclusion

Table 1 demonstrates that each exhibited technique in this paper superiorly accomplishes the contrived requirements than its antecessor.

SecureString 1.0 flops across the board by contributing too less to privacy and nothing to the other IT security requirements (authenticity, integrity, and resilience). Above all, it does not enclose whitelisting.

Although SecureString 2.0 is also restricted to blacklisting, it chalks up with strikingly topped privacy.

SecureString 3.0 satisfies all IT security objectives. Its sole shortfall consists in the absence of whitelisting functionality.

The secure whitelisting approach outplays all antecedent techniques. It tackles all threats and, due to that, bears up all IT security goals as well as payload conformity.

For future work, the improvement of this secure whitelisting technique ought to be the best advice.

Acknowledgments

Many thanks to Bettina Baumgartner from the University of Vienna for proofreading this paper!

References

[1] R. J. Anderson. *Security engineering - a guide to building dependable distributed systems (2. ed.)*. Wiley, jan 2008.

[2] Q. Chai and G. Gong. Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. In *Communications (ICC), 2012 IEEE International Conference on*, pages 917–922, jun 2012.

[3] G. Fahrnberger. Computing on encrypted character strings in clouds. In C. Hota and P. K. Srimani, editors, *Distributed Computing and Internet Technology*, volume 7753 of *Lecture Notes in Computer Science*, pages 244–254. Springer Berlin Heidelberg, feb 2013.

[4] G. Fahrnberger. Securestring 2.0 - a cryptosystem for computing on encrypted character strings in clouds. In G. Eichler and R. Gumzej, editors, *Networked Information Systems*, volume 826 of *Fortschritt-Berichte Reihe 10*, pages 226–240. VDI Düsseldorf, jun 2013.

[5] G. Fahrnberger. A second view on securestring 2.0. In R. Natarajan, editor, *Distributed Computing and Internet Technology*, volume 8337 of *Lecture Notes in Computer Science*, pages 239–250. Springer International Publishing, feb 2014.

[6] G. Fahrnberger. SIMS: A comprehensive approach for a secure instant messaging sifter. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2014 IEEE 13th International Conference on*, pages 164–173, sep 2014.

[7] G. Fahrnberger. Repetition pattern attack on multi-word-containing securestring 2.0 objects. In R. Natarajan, G. Barua, and M. R. Patra, editors, *Distributed Computing and Internet Technology*, volume 8956 of *Lecture Notes in Computer Science*, pages 265–277. Springer International Publishing, feb 2015.

[8] G. Fahrnberger. *A Detailed View on SecureString 3.0*, pages 97–121. Springer Singapore, Singapore, jan 2016.

[9] G. Fahrnberger. *Secure Whitelisting of Instant Messages*, volume 648, pages 90–111. Springer International Publishing, jun 2016.

[10] G. Fahrnberger and K. Heneis. Securestring 3.0 - a cryptosystem for blind computing on encrypted character strings. In R. Natarajan, G. Barua, and

M. R. Patra, editors, *Distributed Computing and Internet Technology*, volume 8956 of *Lecture Notes in Computer Science*, pages 331–334. Springer International Publishing, feb 2015.

[11] G. Fahrnberger, D. Nayak, V. S. Martha, and S. Ramaswamy. Safechat: A tool to shield children's communication from explicit messages. In *Innovations for Community Services (I4CS), 2014 14th International Conference on*, pages 80–86, jun 2014.

[12] ISO/IEC 10116:2006. Information technology – security techniques – modes of operation for an n-bit block cipher. International Organization for Standardization, feb 2006.

[13] C. Örencik and E. Savas. An efficient privacy-preserving multi-keyword search over encrypted cloud data with ranking. *Distributed and Parallel Databases*, 32(1):119–160, mar 2014.

[14] R. S. Tokunaga. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26(3):277–287, may 2010.

[15] U.S. Department of Education. Parents guide to the internet, nov 1997.

[16] J. M. van der Zwaan, V. Dignum, C. M. Jonker, and S. van der Hof. On technology against cyberbullying. In S. van der Hof, B. van den Berg, and B. Schermer, editors, *Minding Minors Wandering the Web: Regulating Online Child Safety*, volume 24 of *Information Technology and Law Series*, pages 211–228. T.M.C. Asser Press, mar 2014.

[17] J. Williams. Promoting internet safety through public awareness campaigns - guidance for using real life examples involving children or young people, nov 2005.

A Crowd-verifiable Microprocessor

Marcel Schaible

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: Off-the-shelf microprocessors with their steadily increasing complexity prevent the construction of verifiable systems. The deployment of these processors in safety critical environments, where potentially faulty systems can harm major infrastructures and in worst case scenarios result in the loss of human lifes, must be considered negligent. Formal, mechanised theorem proving software cannot solve this problem, because verification is, just like mathematical proofs, a social process which relies on the consensus of their community members. Hope and pray is not an option. This paper advocates abandoning superfluous complexity and suggest a microprocessor architecture with the major design goal: simplicity. The whole design will be accessible online for crowd-based evaluation and verification.

1 Introduction

Contemporary microprocessor architectures are optimized for instruction throughput and with the advent of the "mobile revolution", the minimisation of the overall power consumption started to get more attention. These endeavours cumber the verification of processor designs, because of the inherent increasing complexity. Theorem proving software is applied to master this complexity. This approach has some serious problems which are discussed in chapter 3.

Safety-critical systems are the main target of the proposed architecture. Especially these systems are demanding for correctness, reliability, availability and deterministic time behaviour [9, 20, 21].

The VLIW (Very Long Instruction Word) architecture is a variant of a instruction set architecture (ISA) with the goal to increase the execution speed by issuing multiple operations in parallel at the instruction level. A compiler analyses sequential programs and determines operations, which can be performed in parallel. These operations are grouped together and encoded in the instruction word. The size of the various groups is determined by the available functional units.

The proposed architecture adopts this idea by composing an instruction word with several operations not to increase the execution performance but rather to simplify the instruction decoder.

The control unit (CU) is the central part of a microprocessor. It initiates sequences of controlling signals in a predefined order to coordinate the signal flow of the various parts of the processor. Control units are designed with hardwired or microprogrammed control logic [28]. Hardwired control units, especially with a significant amount of instructions, are tedious to verify, because of their hardly comprehensible wired connections between different parts. On the other hand microprogrammed control units model most of their logic in a read-only control store (CS). The sequence of signal transitions is stored in a control store as a two dimensional table and each operation is represented by a set of control signals.

The major requirements of the control unit presented in chapter 2.1 are *consensus-oriented* and *crowd-verifiable* [5, 12]. Both terms are defined in chapter 3.

2 Architecture

The proposed microprocessor architecture (see fig. 1) consists (like the Harvard approach) of strictly separated program (ROM: read only memory) and transient memories (RAM: random access memory), the control unit (CU) and several basic functional units (FU). In contrast to the Harvard architecture the execution of all instructions is strictly sequential to simplify the understanding of the internal state changes of the processor. Because ROM and RAM are housed completely on-chip the memory access is relieved from complicated external memory control logic. The functional units are implementing the basic operations e. g. adding or comparing two data values.

The processor can address 2^m memory cells. The program memory is n-bit wide where n depends on the number of operations (see fig. 4). The transient memory consists of m-bit wide cells and reserves a part of the memory address space to support memory mapped input/output.

The presented microprocessor architecture includes the following properties:

- Instructions are executed strictly sequentially.

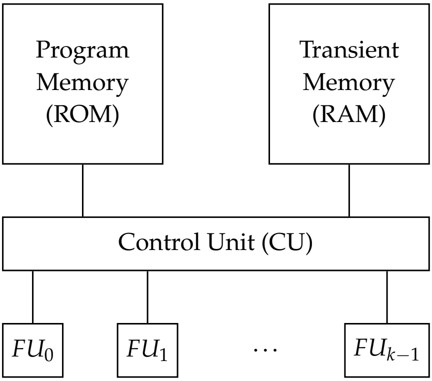


Fig. 1: Architecture Overview

- Code and data memory are housed on-chip.
- Code and data memory are completely separated for safety reasons (Harvard Architecture [11]) .
- Code memory is read-only during execution time.
- The table-driven control unit is designed to guarantee deterministic execution time per instruction.
- The instruction decoding is simplified by defining all instructions equally long.
- Each instruction points directly to a distinct row of the control store.

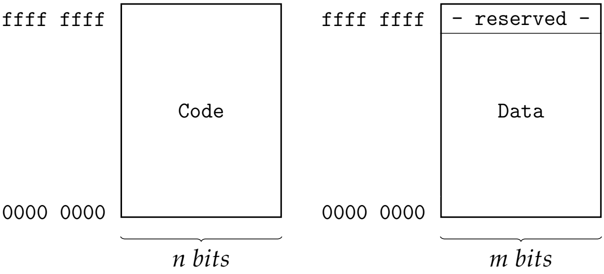


Fig. 2: Memory Layout

2.1 Datapaths and Control Units

The data path (DP) in fig. 3 consists of a set of functional units (FU).

The transient memory is connected through the memory address register (MAR) and the memory content register (MDR) to the data bus. The registers A,B,C, program counter (PC) and the program status word (PSW) are also connected to the data bus. The code memory interfaces via the code address register (CAR) and the instruction register (IR) to the various functional units (FU).

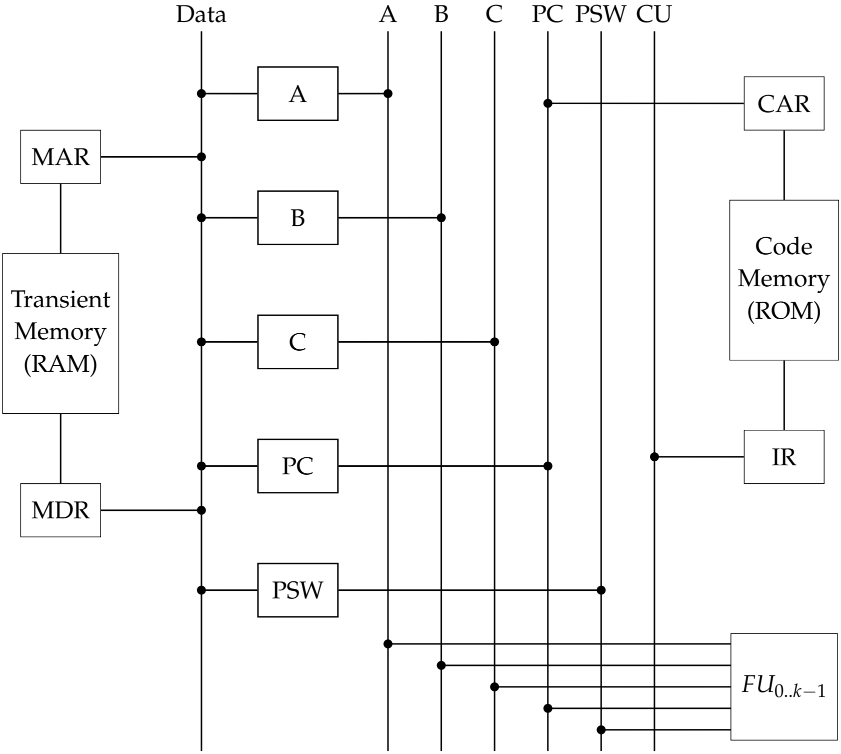


Fig. 3: Datapaths

The major part of the control unit (CU) consists of the control store (CS). CS is a read-only table of control signals where the columns define a set of control signals for enabling certain registers and setting read- or write-flags.

The general execution of the CU is described by algorithm 2.

The control unit is designed with the following properties:

1. The first three operations are stored in ROM cells and each instruction is implicitly prefixed with them.
2. Fetch cycle: The program counter (PC) points to the next instruction and is transferred into the program address register (CAR). This initiates the read of the memory content addressed by CAR. After completion the memory content is available in the instruction register (IR). After loading the IR register the PC is incremented.
3. Execute cycle: Loop over the operations and enable the various functional units (see fig. 5).

A decode phase is unnecessary, because each operation points directly at exactly one row of the control store. All unused operation codes will immediately lead to a processor halt when loaded into the IR register of the control unit.

Algorithm 2: Execution Model

```
Reset;
while forever do
    foreach $i \in \{-3, -2, -1, 0, ..., l - 1\}$ do
        $CU_{addr} \leftarrow \text{Operation}_i$
        Apply selected control signals which e.g. enable $FU$
    end
end
```

2.2 Instructions and Operands

All instructions share the uniform layout in fig. 4 and consists of $0 \ldots l - 1$ operations followed by three operands.

The m-bit wide operands are embedded in the instruction word and can contain constants, direct or indirect memory addresses.

Table 1 list all supported addressing modes.

$$\underbrace{\boxed{opcode_0}\ \cdots\ \boxed{opcode_{l-1}}}_{Operations}\underbrace{\boxed{Operand\ A}\boxed{Operand\ B}\boxed{Operand\ C}}_{Operands}$$

Fig. 4: Instruction Format

Table 1: Addressing Modes

Addressing mode	Description
constant	$register \leftarrow operand$
direct	$register \leftarrow M[operand]$
indirect	$register \leftarrow M[M[operand]]$

Table 2: *Opcodes*

Opcode		Mnemonic	Description
binary	sedecimal		
0000 0000	00	`SKIP`	Skip instruction
0000 0001	01	`LDPC`	$CAR \leftarrow PC$
0000 0010	02	`LDIR`	$IR \leftarrow CDR$
0000 0011	03	`INCPC`	$PC \leftarrow PC + 1$
0000 0100	04	`MARD`	$MAR \leftarrow MDR$
0000 0101	05	`LDAC`	$A \leftarrow constant$
0000 0110	06	`MARA`	$MAR \leftarrow A$
0000 0111	07	`AMDR`	$A \leftarrow MDR$
0000 1000	08	`LDBC`	$B \leftarrow constant$
0000 1001	09	`MARB`	$MAR \leftarrow B$
0000 1010	$0a$	`BMDR`	$B \leftarrow MDR$
0000 1011	$0b$	`LDCD`	$C \leftarrow addr$
0000 1100	$0c$	`MARC`	$MAR \leftarrow C$
0000 1101	$0d$	`CMDR`	$C \leftarrow MDR$
⋮	⋮	⋮	⋮
0010 0000	20	`ADD`	$C \leftarrow A + B$
⋮	⋮	⋮	⋮
0100 0000	40	`COPR`	$M[C] \leftarrow R$
⋮	⋮	⋮	⋮
1111 1111	ff	`Halt`	Halts the execution

The encoding of the various addressing modes is done by issuing the appropriate operation like $LDAC$ for loading the register A with a constant.

2.3 Instruction Register

The instruction register (IR) stores the active set of operations and operands. Because one requirement of this architecture is to relieve the CU from complex logic, the fetch phase is integrated into the IR. The opcodes with indices -3 through -1 implement the fetch of the next instruction and are hardcoded. With this approach the CU does not implement a distinct fetch phase.

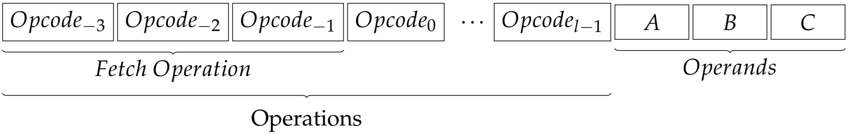


Fig. 5: Instruction Register

2.4 Functional Units

All functional units (see fig. 6) share the following structure:

- Inputs: Datapath to the operands A and B, program counter (PC) and program status word (PSW)
- Processor clock
- Enable signal activates the processing of the FU.
- Outputs: Datapath to the result register, PC and PSW

Because some functional units do not use all of the connections (e.g. the adder does not need to access the PC) it is feasible not to implement these paths.

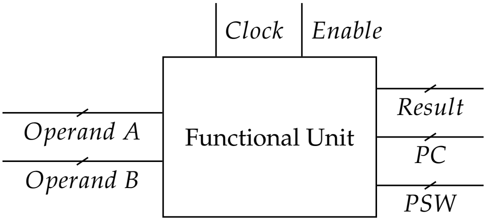


Fig. 6: Functional Unit

2.5 Example

Consider the following instruction for adding the constant `#42` to the value at memory address `$10` and storing the result pointed to by the address `$20`.

```
ADD #42, $10, [$20]
```

The instruction encoding is described in fig. 7.

LDPC	LDIR	INCPC	LDAC	MARB	BMDR	MARC	MARD	CMDR	⋯	42	10_{16}	20_{16}

Fig. 7: ADD Instruction Format

The execution of this instruction is divided into the following steps:

ADD	#42, $10, [$20]	
LDPC		Load `PC` into `CAR` register
LDIR		Load `CDR` into `IR` register
INCPC		Increment the `PC`
LDAC	42_{10}	Load the `A` register with the constant `42`
MARB	10_{16}	Load 10_{16} into the `MAR` register
BMDR		Store `M[`10_{16}`]` into the `B` register
MARC	20_{16}	Load 20_{16} into the `MAR` register
MARD		Load `M[`20_{16}`]` into `MAR`
CMDR		Store `M[M[`20_{16}`]]` into the `C` register
ADD		Enable the adder
STR		Store the value in `R` into `M[C]`

3 Verification

The standard model of chip verification is described in fig. 8. Starting with the designer's intention a functional specification is produced, which leads to a microarchitecture, a Register-Transfer-Level (RTL) description, a netlist and finally from a physical layout to a silicon dye. In each step the correctness of the transformations must be ensured.

Chip manufactures utilise formal verification methodologies in some areas like floating-point unit for risk minimisation [14]. Because of the complex designs of modern processors, automated theorem proving systems (ATPS) [2, 4, 13, 29] tend to generate large proofs with thousands of deductions which cannot be checked even by human experts. It is the state of the art that sufficiently complex software programs like ATPS are unlikely to be error-free and therefore the confidence into the correctness of the generated proofs is questionable. This implies that errors or misconceptions regarding the specification will remain undetected.

The main drawbacks of automated and mechanised theorem prover are:

- Sophisticated designs produce long and hard-to-follow reasoning.
- Theorem provers like all other software programs are not error free. Therefore the generated proofs can contain errors which are hard to reveal.
- Intentions of designers cannot be formalised.

Understanding the nature of a proof cannot be achieved by reading it line-by-line and checking each line for syntactically correct transformations. A proof is more a guide for reasoning [22, p. 317].

Confidence and trust in technical artefacts can only be gained by transparent designs and hence are the result of a social process. The key is *intelligibility* and as a consequence *simplicity* of the architecture. Comprehensible and verifiable designs by humans are inevitable.

3.1 Consensus-oriented Verification

Because of the clean and straightforward design of the above described architecture the verification will be performed consensus-oriented and crowd-based [5, 12, 24, 25]. Consensus-oriented verification is the process of examination of a community (crowd-based), which comes to a common understanding and

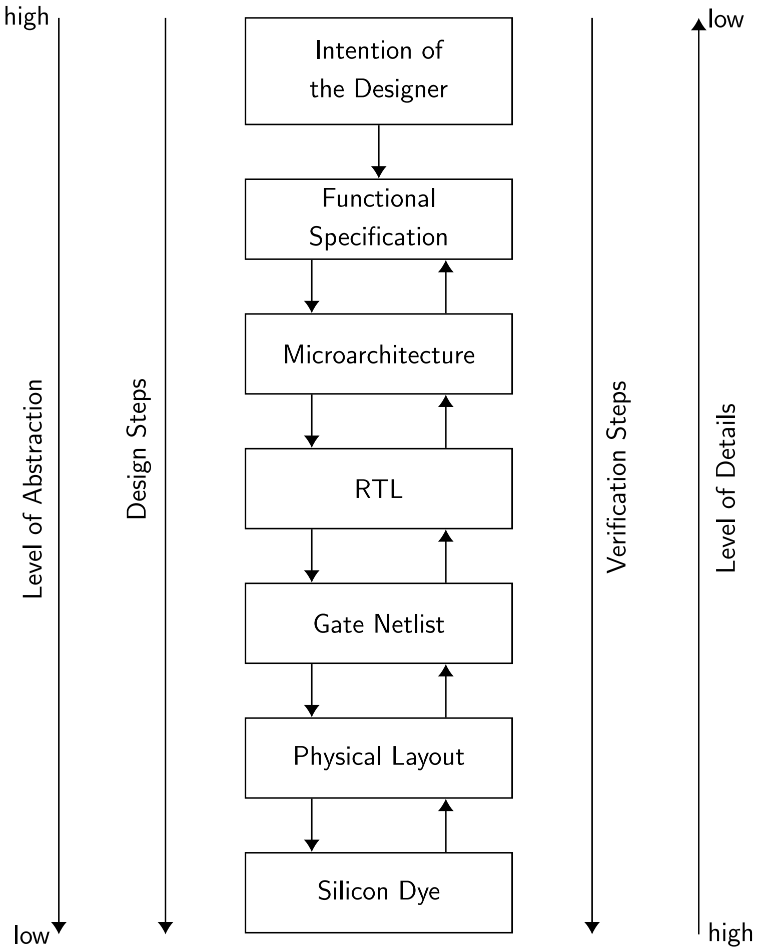


Fig. 8: Verification Model

agreement that a design can be considered correct. This approach is well established in science. Mathematicians have been working consensus-oriented and provided their work to the mathematic community for examination with great success for thousands of years.

The proposed architecture will be published online for crowd-based evaluation and verification.

But because of the inherent design properties of the presented architecture it is feasible to apply the consensus-oriented and crowd-based methodology to a reversed verification process: Starting from the physical layout the verification can be done backwards to the netlist all the way up to the specification. This approach was first applied as a software verification methodology by TÜV Rheinland [19] to acquire permission for commissioning of the nuclear power plant in Halden (Norway). Although this method was used with software it can be adapted to hardware verification. An important ancillary effect of the diverse backward analysis methodology [9] is the intrinsic verification of the correctness of the applied transformation tools.

The core characteristics (see [9, pp. 152-153]) of the diverse backward analysis methodology are

1. The process of verification is diverse in relation to the implementation.
2. The process of verification has the character of a proof.
3. Each step of the verification is strictly documented and checkable.
4. The verifier is not defencelessly delivered to potential systematic errors of the verification process.

4 Summary

The objective of this work is to develop a verifiable microprocessor architecture for usage in safety-critical systems. The architecture has consequently cut out all *artificial complicatedness* [10].

The described control unit architecture with its table-driven composition and strictly sequential execution logic provides a coherent architecture which can be verified by field and non-field experts in a consensus-oriented and crowd-based approach. Due to the characteristics of the table-driven design the verification can be literally performed by checking each set of generated control signals.

The crowd-based verification can be performed both forward and backward in a consensus-oriented discourse with the objective to gain confidence und trust in the correctness of the examined design. Diverse backward analysis as a powerful verification methodology, which establishes trust not only in the correctness of the design itself but also in the tools used to generate it, can be applied.

References

[1] Andersen, B. Scott and Romanski, George: Verification of Safety-critical Software, In: *Queue, ACM*, 9, 8, pp. 50–59, 2011

[2] Berg, C. and Beyer, S. and Jacobi, C. and Kröning, D. and Leinenbach, D.: Formal Verification of the VAMP Microprocessor (Project Status), Symposium on the Effectiveness of Logic in Computer Science (ELICS02), pp. 31–36, 2002

[3] Berg, C. and Jacobi, C. and Kröning, D.: Formal Verification of a Basic Circuits Library, In: *Proc. of the IASTED International Conference on Applied Informatics, Innsbruck (AI 2001)*, pp. 31–36, 2001

[4] Beyer, S. and Jacobi, C. and Kroening, D. and Leinenbach, D. and Paul, W. J. : Putting it all together - Formal Verification of the VAMP, In: *Software Tools for Technology Transfer (STTT), Special Section on Recent Advances in Hardware Verification* , 8, pp. 411–430, 2006

[5] Brabham, Daren C.: Crowdsourcing, *The MIT Press*, 2013

[6] Cohn, A: A proof of correctness of the viper microprocessor: the first level, In: *University of Cambridge, Computer Laboratory*, 1987

[7] Dijkstra, E. W.: The next fifty years, 1996

[8] Gilreath, W. F. and Laplante, P. A.:Computer Architecture: A Minimalist Perspective, *Kluwer Academic Publishers*, 2003

[9] Halang, W. A. and Konakovsky, R.: Sicherheitsgerichtete Echtzeitsysteme, *Oldenbourg*, pp. 150–152, 1999

[10] Halang, W. A.: Simplicity Considered Fundamental to Design for Predictability, In: *Perspectives Workshop: Design of Systems with Predictable Behaviour, 16.-19. November 2003*, 2004

[11] Hennessy, J. and Patterson, D.: Computer Architecture - A Quantitative Approach, Morgan Kaufmann, 2011

[12] Howe, Jeff: Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business, *Crown Business*, 2008

[13] Hunt, W. A. J.:FM8501: A Verified Microprocessor, In: *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence*, Springer, 1994

[14] Intel Corp.: *Intel Xeon Processor E3-1200, v3 Product Family, Specification Update*, 2015

[15] Joyce, J.J.: Formal specification and verification of microprocessor systems, In: *Technical report (University of Cambridge. Computer Laboratory)*, 1998

[16] von Kaenel, P. A.: Designing and testing a control unit, In: *J. Comput. Sci. Coll.*, 19, 5, pp. 228–237, 2004

[17] Kaivola, Roope et al.: Replacing Testing with Formal Verification in Intel® *Core*^{TM} i7 Processor Execution Engine Validation, In: *Proceedings of the 21st International Conference on Computer Aided Verification*, Springer Verlag, pp. 414–429, 2009

[18] Knuth, D. E.: An Empirical Study of FORTRAN Programs, In: *Softw., Pract. Exper.*, 1, 2, pp. 105–133, 1971

[19] Krebs, H. and Haspel, U.: Ein Verfahren zur Software-Verifikation. In: *Regelungstechnische Praxis*, rtp 26, pp. 73â78, 1984

[20] Laprie, J. C., Avizienis, A. and Kopetz, H.: Dependability: Basic Concepts and Terminology, Springer Verlag, 1992

[21] Laprie, J.-C. and Béounes, C. and Kanoun, K.: Definition and Analysis of Hardware- and Software-Fault-Tolerant Architectures, In: *Computer, IEEE Computer Society Press*, 23, 7, pp. 39–51, 1990

[22] MacKenzie, D.: Mechanizing Proof: Computing, Risk, and Trust, MIT Press, 2001

[23] Schaible M.: Towards the Verification of a Table-driven Microprocessor Architecture for Safety-critical Systems, In: *Fortschritts-Berichte VDI, Proceedings of the 6th GI Workshop Autonomous Systems 2013*, Vol. 835, pp. 24–30, 2013

[24] Schaible M.: A Consensus-oriented Crowd-verifiable Microprocessor Architecture, In: *Fortschritts-Berichte VDI, Proceedings of the 7th GI Workshop Autonomous Systems 2014*, Vol. 835, pp. 209–220, 2014

[25] Schaible M.: Design of a Consensus-oriented and Crowd-verifiable Control Unit, In: *Fortschritts-Berichte VDI, Proceedings of the 8th GI Workshop Autonomous Systems 2015*, Vol. 842, pp. 52–60, 2015

[26] Schaible M.: On the construction of a crowd-verifiable Microprocessor, In: *Fortschritts-Berichte VDI, Proceedings of the 9th GI Workshop Autonomous Systems 2016*, Vol. 848, pp. 46–54, 2016

[27] Stieger, H. and Halang W. A.: Eine hochsprachenorientierte Rechnerarchitektur ohne arithmetische Register, *IFB Verlag*, 2003

[28] Wilkes, M. V. and Stringer, J. B.: Micro-programming and the design of the control circuits in an electronic digital computer, In: *Mathematical Proceedings of the Cambridge Philosophical Society*, 2, pp. 230–238, 1953

[29] Phillip J. Windley: Formal Modeling and Verification of Microprocessors, In: *IEEE Transactions on Computers*, 44, 1, pp. 54–72, 1995

Development of Embedded Systems in the Domain of Measurement Techniques

Irina Kaiser[1], Wolfgang Fengler[1] and Maxi Weichenhain[1,2]

[1]Computer Architecture and Embedded Systems Group
Technische Universität Ilmenau, Germany

[2]BearingPoint GmbH, Germany

Abstract: The measurement technology represents a typical domain for embedded computer systems. A development process intended for this purpose must meet the necessary general and special requirements. A complex and highly efficient information processing is needed for more powerful measuring systems and their functions. It guaranties as well their feasibility.

The developed and described certifiable development process (ZEfIRA), based at a W-model, provides a procedural method for the development of complex embedded systems according to a top-down principle. This principle covers functional design, structural design, partitioning into embedded systems, hardware modules, reconfigurable logic (FPGA) and software modules. Furthermore it supports simulation based testing in an early state and a joint development of measurement and information technology, while constantly taking into account the interactions occurring. A highly flexible and performance FPGA prototype is of particular importance in the development process. In particular, a calibrateable measuring system demands compliance with corresponding regulations, such as the weights and measures act and other guidelines.

Within the framework of the development of measuring instruments, a verification ("certification") is required to ensure the requested uncertainty of measurement under all specified measurement conditions of the respective device (metrological reliability). This applies to the entire measurement chain and the involved domain of functional determining information processing in hardware and software. In this article the common development approach will be expanded with additional requirements for metrological reliability and a solution will be presented to enable the calibration capability and ensure compliance with the above-mentioned guidelines in the early stages of the development process.

When analyzing the requirements, the measuring instrument must be taken into account. This involves the consideration of the functional (predominant in the field of measurement management) and non-functional specifications for information processing. An effective process management must go along with the development process.

1 Introduction and Motivation

The measurement technology represents an important domain for the use of more or less complex embedded computer systems regarding required information processing behavior. Very complex (such as machines for nanopositioning and nanometer measurement [1]) or medium complex systems (such as dynamic weighing technology [2]) have development requirements result from embedded systems and specific measurement technique (such as metrological reliability, calibration capability and so on). Therefore, this domain can very well be used to design a process that corresponds to both types. A general character can be achieved with this process. The following representations summarize a series of works that have been carried out and described as partial results (e.g. [1–3, 5, 9–11]).

Modern measuring technology with global quality standards requires not only a safe system but also a system, which is adapted to measurement uncertainty and time behavior. Continuously growing requirements with regard to precision and dynamics in the metrological domain lead to the need for new approaches to increase performance. The particularly powerful measuring systems in the field of length measurement technique with resolution and measurement uncertainty in the nm boundary region (e.g. nanopositioning and measuring machine (NPMM) [1]) as well as the measuring instruments in dynamic weighing technology [2] require complex and high-performance information processing functions. The complexity of the algorithmic part results from alternative filter, control and signal processing functions with many computational operations, which have to be realized in adapted higher accuracies [3]. The most important objectives of all measures in information processing, including the measuring technology, are to shorten the measuring time, reduce the uncertainty of measurement and increase the resolution. These requirements must be fulfilled in a specific cost frame of the device.

The development costs have to circle round specific range under the requisite short-term development (”time to market”). As state of the art, such complex solutions of information processing are feasible to realize by machine-oriented embedded systems [2, 4]. Many measuring systems are based on the character of complex mechatronic systems. The design of such systems can only be implemented effectively using a model-based design procedure.

In the model-based design, the functional requirements are transferred into a model. After that the model will be transferred into the implementation structure. The validation of the implementation structure against the requirements is complemented by the validation of the model against the requirements and the implementation structure compared to the model. Furthermore the implementation structure can be formally verified against the model if the techniques have been developed or are available for this purpose. Usually, several models are used with different degrees of abstraction and models of different character are also be used.

The development, in this case used in the metrological domain with possibly extended requirements of the calibration capability, uses a goal-driven deployment and systematic use of principles, methods, concepts and tools. The assurance of the propriety functions of the information processing, measuring technical and electrical components plays an important role. For instance, the weighing technology requires proof of the product complies regarding the corresponding calibration laws, standards, directives and standards (e.g. Directive 2004/22 / EC of the European Parliament on measuring instruments and WELMEC) [9]. These requirements must be met by the measuring instruments, which must be documented and adhered during development and operation. As a result this implies that the development methods and tools used must be accepted in a certification process, even if there are no normative rules. Another organization, which requires certified processes besides the national calibration authority, is the Food and Drug Administration (FDA). In particular, special requirements on specific weighing classes are imposed (for example related to the pharmaceutical and food industry usage).

2 Systematic Development Process

A largely formalized process contributes to the fact that the later product has significantly fewer development defects. Therefore from an economic point of view, the application of a systematic development process is required. However, the pure formalization is not enough. For the considered application domain, particular methods and design methods needs to be supported:

- The overall system design is performed according to the top-down principle, beginning with the highest abstraction level (requirement specification) to the lowest (realization / implementation).
- In the reuse of previous solutions, the meet in the middle approach (combination of top-down and bottom-up principle) can be integrated as well. Both implemented functions and model parts can be used.
- Embedded system (instrumental part) and embedded system (information processing) are developed in parallel and close interaction.
- There is a separation in functional design (algorithms) and structural design (realization and implementation).
- Stages build one after another are carried out model-based. This means that both models are used for the measurement as well as for information processing part.
- The models should be executable (simulatable). This allows an early validation of the partial and total system at different abstraction levels. Test scenarios of the models later provide the basis for implementation testing.
- Possible parts for realization platforms of information processing are the embedded software, hardware and reconfigurable hardware (FPGA, Field Programmable Gate Array).
- In the transition from the functional to the structural models (platform-dependent), partitioning and possibly later repartitioning into the measuring technical and the three information technological variants part has to be carried out. Whereby for alternative assignable model parts the technically and economically most sensible variant is to be chosen.
- In the abstraction level, in contrast to the design, various integrations of the partial solutions, as well as testing itself, integration and acceptance tests are carried out after realization/implementation. This happens with the use of the above-mentioned model-based test scenarios.

The ZEfiRA process, which was developed on the basis of these requirements and is presented in [6] and [11], enables the effective design and testing of complex information processing devices in hardware and software, described using the example of weighing technology. Important aspects are the model-based design, including the integration of verification and validation activities in each design step, which are associated with appropriate approaches. In the approach described, new concepts, ideas and target platforms are first tested on a prototype before a producible solution for serial production is developed. The prototype (the information processing) is defined the following: it enables the investigation, functional and in principle the structural design of the planned product (here the measuring device), especially in the critical features for different variants, such as for the investigation of the measuring technology part. The development of prototypes therefore makes it possible to find a suitable design solution (for example observing the latencies resulting from the frequency spectrum of the measuring technology part, while at the same time minimum resource consumption is fulfilled) from a plurality of implementation and implementation variants [11].

Figure 1 shows the W-model in the development of a prototype. The name given to the model is based on the presentation structure of the procedure [11], the letter W. For the producible solution, a similar procedure can be specified, whereby the prototype with regard to this generally has a greater complexity with fewer claims to the most economical realization. Models, partitioning and fundamental realizations/implementations of building the prototype are the starting point for the producible solution.

The base of this model is a model-based top-down principle with previous test activities already mentioned above as a requirement at the time during the design.

The functional requirements for a system to be designed are transferred into a model (possibly into an executable specification model or a virtual prototype in the technical system design). A tested functional model will be reproduced and improved during the technical design on a realizable model. This is separated into the embedding system (the measuring part) and the embedded system (information processing). All decisions to be made from the technical possibilities (what can be realized in the partition) and the necessities (what has to be realized in the partition) are dependent. If there are several possible variants to choose from, additional criterias such as cost effectiveness of the solution will be taken into consideration. The stages up to the module specification (after the partitioning of the embedding and embedded systems) will be modeled, investigated and tested independent from the target system. The ″testing″ is to be understood as an extension by additional validation and verification activities.

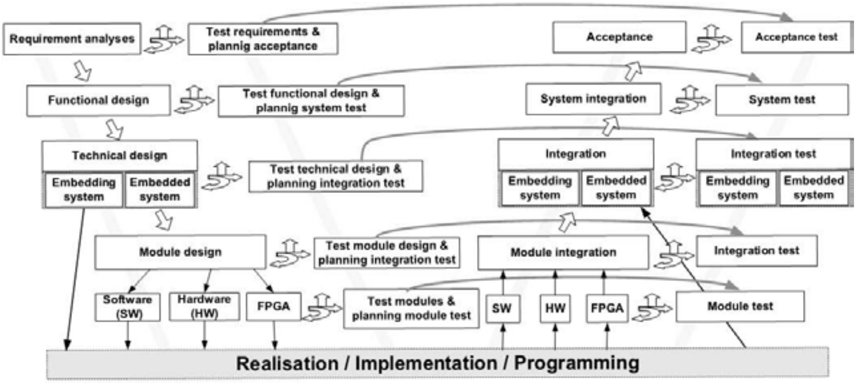


Fig. 1: W-model of the prototypical development

The models to be used must be executable in a suitable tool (testable to the model level). For verification, it is useful to generate models from the simulatable descriptions (e.g. petri nets to extract control flows from data dependencies in data flows).

The right-hand side of the W-model contains the opposite direction with respect to the degree of abstraction during integration and testing of the implementation parts up to the system. The ideal behavior shown, in which all parts are at the same level of abstraction, are usually achieved in several stages.

The analysis of the interaction of the measuring- technical part and the information processing (both parts based on models or in mixed form, "model in the loop", "software in the loop", "hardware in the loop") is possible even before the complete realization is finished. Furthermore it is feasible to achieve this in different variations. By using an information system with performance parameters and functions which are higher than those of a producible solution (here referred to as a prototype), the behavioral and parameter analysis as well as the investigation of different algorithms are supported in the lower abstraction levels. The latter is then already done with its concrete realization. All these mentioned points within the W-model process ensure an effective validation of the correct function and an appropriate documentation [9].

According to [11], existing sub models and implementations from earlier developments in terms of reuse can be used as the starting point of the current project. Depending on the functional proximity of both, they can be the starting, adapted or reused partial solution. In the case of identical or similar implementation platforms, implementation must also be considered. In all the cases the models provide effective access.

On account of the calibration capability, additional measures must be integrated in the solution for information processing in order, for example to prevent manipulation. This has to be modeled and realized on different levels of abstraction (beginning with the early stages).

3 Metrological Reliability and Criteria

The specialized requirements at the development process for metrological reliability will serve as an example.

3.1 Definition

The measurement uncertainty needs to be constantly demonstrated during and afterwards of the development of devices. There is a calibration requirement for a large number of measuring instrument classes, which set the requirements for regularly measurement uncertainty checks during operation. By the approval of the measuring instruments or of the measuring systems a proper operation can be assured. Such complex information processing in these systems requires, in the course of development, an efficiency test of the measurement uncertainty for the applied method. Compliance with qualitative and quantitative criteria is intended to guarantee metrological reliability. In the metrological literature [7, 8] this is diffusely described, which is why it will be defined as following:

> **Definition** *The metrological reliability as part of an information processing system of a measuring device ensures that the correct function taking into account the required measurement uncertainty over the entire measuring range and under all permitted measuring conditions of the measuring device for the duration of the configured or legally permissible operating time. The latter applies to measuring instruments for which verification is required.*

Metrological reliability must always be considered and taken into account in the context of technical and economic requirements. The following list represents a summary of the criteria that must be met:

- Consideration of calibration requirements to the measuring instrument, which consists the measuring system and the information processing.

- Consideration of requirements beyond the calibration capability (standards, regulations, directives and similar),
- Use of a systematic development process, whereby the error probability decreases,
- Model-based development and model-based testing,
- Validation and verification during all development steps and sub processes using appropriate methods,
- Consideration of systematic and random errors for the proof of metrological reliability (error analysis, mathematical modeling, calculation the contribution of each measuring member).

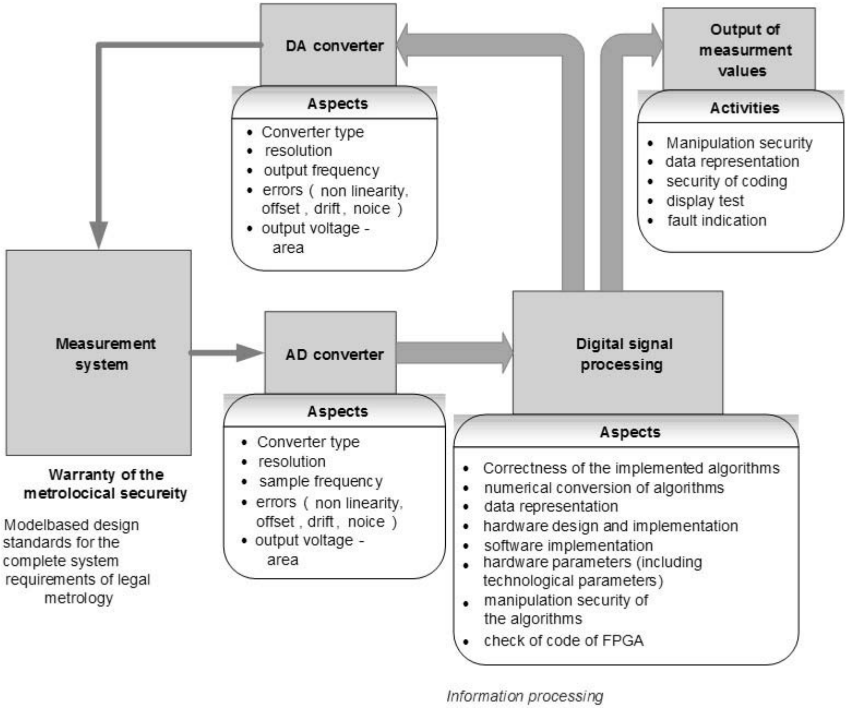

DA converter
Aspects
• Converter type
• resolution
• output frequency
• errors (non linearity, offset , drift, noice)
• output voltage - area
Output of measurment values
Activities
• Manipulation security
• data representation
• security of coding
• display test
• fault indication
Measurement system
AD converter
Aspects
• Converter type
• resolution
• sample frequency
• errors (non linearity, offset , drift, noice)
• output voltage - area
Digital signal processing
Aspects
• Correctness of the implemented algorithms
• numerical conversion of algorithms
• data representation
• hardware design and implementation
• software implementation
• hardware parameters (including technological parameters)
• manipulation security of the algorithms
• check of code of FPGA
Warranty of the metrolocical secureity
Modelbased design standards for the complete system requirements of legal metrology
Information processing

Fig. 2: Aspects of the metrological reliability in the measurement chain

3.2 Consideration of Metrological Reliability

Figure 2 shows the measuring chain with the function blocks of the information processing, which is typically realized by an embedded system. The contributions to metrological reliability are assigned at the appropriate place. The aspects and actions to be considered for each subsystem must be integrated into the development process. The metrological reliability of an information processing system refers to the totality of the individual aspects. In addition, the general requirements for the measuring system ensure as well the metrology reliability.

3.3 Realization of Additional Functions

In order to ensure metrological reliability and fulfill the legal and specific requirements, various additional functions must be implemented in the information processing system and, as a rule, protection of the information processing against external attacks needs to be implemented. These can be developed as re-usable blocks at the level of the model and adapted as a partial realization/implementation, as well as adjusted and integrated during the system design. Furthermore an overview of the points in the measuring chain, which require a special consideration of these aspects in the realization, is given.

- **Protection against manipulation from outside** is provided to prevent the manipulation of transmitted parameters (in signal processing) and/or measurement results. For the realization checksum, digital signature and encryption (using a cryptographic method) can be used.

- **Protection against internal attacks** by identifying malfunctions which can be activated externally or arise during the development and as an outcome would subsequently modify the behavior of the information processing. For the realization formal verification, memory protection against manipulation, checksum or cryptographic hash function can be used to protect sensitive data.

- **Proof of correctness and plausibility** includes detection, display and suppression of the measured value output when a function has been manipulated or does not provide correct results (e.g. due to failure of a hardware block). An online test data generator developed for this purpose can be used to implement the safety and test component. This checks the results with the associated, previously calculated expected values and runs parallel to the actual program sequence.

- **Internal protective measures** must be used to monitor execution times. For this purpose, the realization of time calculation and evaluation components is realized, which compares the time of the calculation of a certain operation or function during the execution with an expected value.

4 Requirements and Project Management

The procedure of the project management in the considered target domain can be roughly divided into five sub steps. In Figure 3 the procedure is presented shortened and the main tasks of the respective project phase are labeled. Application domain specific and general aspects are included.

- Top of the list you can find the requirement collection forms as the basis of project management. This inquires not only the identification or the collection of requirements, but also the derivation of the objectives. The requirements for information processing are the result of the requirements analysis of the entire measuring instrument. An important role in the introduction of a scale is the analysis of influence and disturbance variables, which can play a decisive role in project introduction and selection of the required balancing properties.

  This will be explained briefly on the basis of the functional requirements ″accuracy″ and ″measurement time″. From the requirement ″accuracy″ (in measurement technology: measurement uncertainty) results the demand on the error behavior and the error propagation based on calculations from the ″measurement time″ based on the guaranteed sampling times for the digital signal processing.

  Further requirements arise mainly from process conditions, conditions at the installation site, influence and disturbance variables, errors (systematic or accidental), hardware and software aspects, laws to be complied with, auxiliary equipment, tests, customer specific requirements, economic criteria (decretory) and general considerations.

- The project definition includes the first major planning of the project, which includes the setting of milestones. The domain specific project objectives must be clearly defined and set up.

- In the preplanning of the project, a project effort estimate needs to be carried out for the calculation of the required costs. In addition time management and scheduling is an integral part.

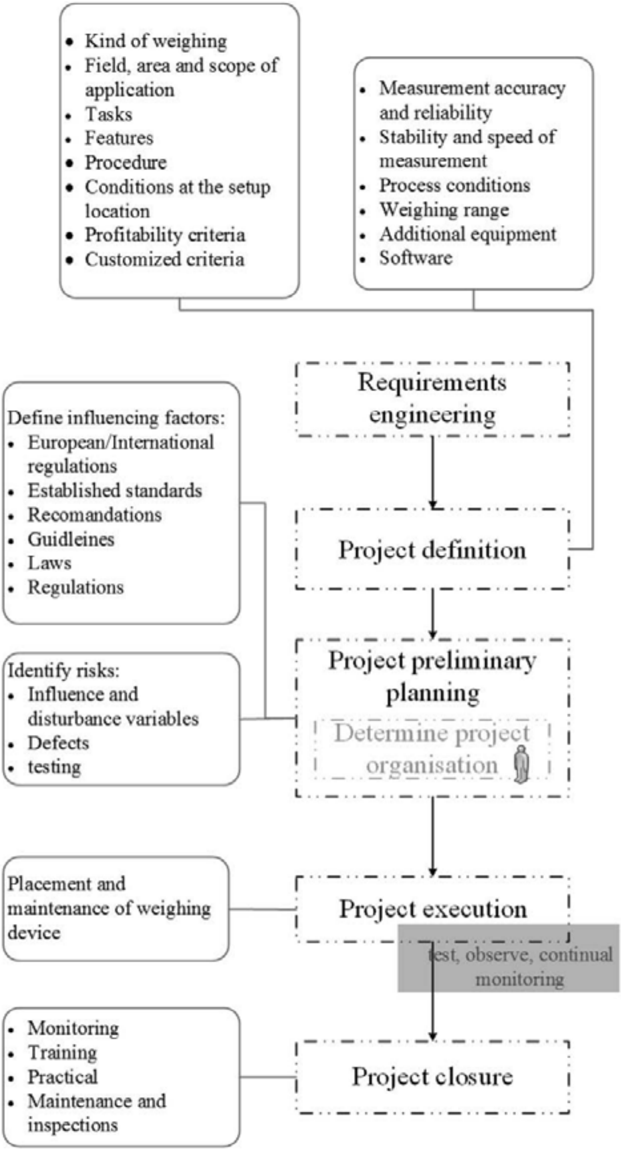

Kind of weighing
Field, area and scope of application
Tasks
Features
Procedure
Conditions at the setup location
Profitability criteria
Customized criteria
Measurement accuracy and reliability
Stability and speed of measurement
Process conditions
Weighing range
Additional equipment
Software
Requirements engineering
Define influencing factors:
European/International regulations
Established standards
Recomandations
Guidleines
Laws
Regulations
Project definition
Identify risks:
Influence and disturbance variables
Defects
testing
Project preliminary planning
Determine project organisation
Placement and maintenance of weighing device
Project execution
test, observe, continual monitoring
Monitoring
Training
Practical
Maintenance and inspections
Project closure

Fig. 3: The procedure of project management in dynamic weighing technology – shortened

- The project organization is rather domain independent and includes change management, configuration management, all administrative work, risk management, group and team formation and the definition of the individual team leader. It represents an independent, comprehensive process with a wide range of tasks, including the assignment of the necessary employees and their responsibilities. For this purpose the capacity testing, the establishment of uniform terminology standards (here also a common conceptual world between the developers of the embedding system, the embedded system and the subsequent user of the device or the customer), the documentation of the project, the determination of the exact flow chart (W-model), the concretization of the milestones, the determination of the to be observed influencing factors and the risk assessment.

- The project implementation is followed by project implementation, which is determined by the configuration, change, deadline, risk, quality, team and resource management. Throughout the project implementation, an iterative testing of requirements, monitoring and constant control of the project framework, budgets, quality and delivery is necessary. The project implementation follows the W development process described in section 2, based on the scheduling, milestones, resources required and special qualification of the employees.

- The project closure (of the prototype) is primarily concerned with the documentation of the project, with the main aim to record the experiences for later projects.

- In this case, the further use in particular of models and developed partial solutions for the product produced and a successor system/device must be prepared from the documentation viewpoint. Provision must also be made for such follow up developments.

5 Summary

The article describes the development of complex information processing solutions in measurement technology. In addition to a systematic approach, a specific requirement for metrological reliability has been presented, which must be taken into account in the context of technical and economic requirements and is one of the essential objectives in the development of a measurement instruments. The underlying development process is evolutionary and includes the steps of model-based, platform-independent design as well as the stage specific design. Metrological reliability and the additional appropriate measures for security against manipulation are intended to ensure the certainty and compliance with standards, regulations, directives and to be supported directly in the design process. In this context, the realization of different security and protection functions in the early stages of a development process plays an important role. Depending on the specific application, the functions described can be combined, supplemented or modified.

References

[1] S. Zschäck, J. Klöckner, I. Gushchina, A. Amthor, C. Ament: ″Control of nanopositioning and nanomeasuring machines with a modular FPGA based data processing system", in: *IFAC Mechatronics*, 23(3): pp. 257–263, 2013

[2] H. Weis, I. Gushchina, A. Amthor; F. Hilbrunner, T. Fröhlich: ″Investigation of digital control concepts for dynamic applications of electromagnetic force compensated balances", in: *2013 NCSLI International Workshop and Symposium*, 2013

[3] M. Müller, J. Klöckner, I. Gushchina, A. Pacholik, W. Fengler, A. Amthor: ″Performance evaluation of platform-specific implementations of numerically complex control designs for nano-positioning applications", in: *IJES*. Vol. 5.2013, 1/2, pp. 95–105

[4] I. Grout: ″Digital Systems Design with FPGAs and CPLDs", Elsevier Ltd, 2008, ISBN-13: 978-0-7506-8397-5

[5] I. Gushchina, W. Fengler, J. Ciemala, D. Streitferdt: ″Certifiable development process for information processing in the measurement domain on the base of 3W-models", in: *Embedded Systems Design at the 37th IEEE Annual Computer Software and Applications Conference*, pp. 685–690, 2013

[6] European cooperation in legal metrology WELMEC: ″Software Guide (Measuring Instruments Directive 2004/22/EC)", WELMEC 7.2, `http://www.welmec.org/fileadmin/user_files/publications/WELMEC_07.02_Issue5_SW_2012-03-19.pdf` (Stand Oktober 2015)

[7] P. Sarhadi, S. Yousefpour: ″State of the art: hardware in the loop modeling and simulation with its applications in design, development and implementation of system and control software", *International Journal of Dynamics and Control*, December 2015, Volume 3, Issue 4, pp. 470–479

[8] n.n.: ″Systemdesignsoftware LabVIEW", National Instruments 2016, `http://www.ni.com/labview/d/` (as October 2016)

[9] I. Kaiser: ″Certifiable Development Process for Information Processing in the Technology of Balances" (in German: ″Zertifizierbarer Entwicklungs-

prozess für komplexe Informationsverarbeitungssysteme in der Wägetechnik"), Dissertation TU Ilmenau, 2015, `https://www.db-thueringen.de/receive/dbt_mods_00027143`

[10] I. Kaiser, W. Fengler, T. Fröhlich: "Metrologic Reliability of an Information Processing System in the Measurement Technology" (in German: "Metrologische Sicherheit eines Informationsverarbeitungssystems in der Messtechnik"), Messunsicherheit praxisgerecht bestimmen und Prüfprozesse in der industriellen Praxis: Braunschweig, 19. und 20. November 2015. - Düsseldorf: VDI-Verl. 2015, pp. 233–240, 2015

[11] D. Grüner, T. Baumann, B. Däne, W. Fengler: "Simulation driven development of complex systems: foundations for simulation based design workflows", in: *Modelling and Simulation 2016* (ESM '2016), pp. 413–419, Las Palmas: October 26-28, 2016

A Data Specification Architecture

Stefan Widmann

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: Embedded systems are used in rising numbers of safety-related applications, e.g. steer-by-wire and brake-by-wire in automotive applications. At the same time, complexity of hard- and software of such systems is increasing and the reduction of the minimum feature sizes of used integrated circuits makes them more sensitive to environmental influences, resulting in a rising number of data flow errors. Instead of trying to detect these errors by even more complex software, a Data Specification Architecture is presented, which adds a comprehensive set of data properties to data values in form of tags. It provides a high level of error prevention and a simple hardware-based detection of 20 identified data-flow-related errors and attacks.

1 Introduction

Safety-related applications that have been realized using mechanical or electromechanical systems before are being more and more realized by programmable systems today. Good examples are automotive and avionic applications like braking, steering and flying: brake-by-wire, steer-by-wire and fly-by-wire replace the proven and reliable mechanical systems by electrical and electronic sensors and actors, the sensors' signals being processed by microprocessors to calculate control signals for the actors [2, 19, 25].

At the same time the complexity of the software used in embedded systems is rising. In 2006, Broy states in [6], that a car contains about 10 million lines of code and that he expects a rise by the factor ten with each car generation. Three years later, in 2009, Broy is cited in [7], stating that the number of lines of code has risen to 100 million. In 2008, a future rise to 200 to 300 millions was estimated in [11].

Complexity of the hardware is rising, too. Over 70 embedded systems were used in a car in 2006 [6] and up to 100 in 2009 [7]. On chip level, the number transistors per die has reached 1.3 billion in 2015 [24]. The integration of growing numbers of components in integrated circuits is only possible due to a continuous reduction of the minimum feature size, making them more sensitive to environmental influences like radiation, especially neutrons, even on ground level [4, 20].

All of this results in rising numbers of software and hardware errors. Since commercial-off-the-shelf (COTS) microprocessors commonly used in such systems are not designed to provide effective means for error prevention and detection on hardware level, more and more complex software is used to detect errors during runtime, e. g. by applying arithmetic coding in the form of Software Encoded Processing (SEP) and Compiler Encoded Processing (CEP) [23]. This additional software complexity results in a higher probability of software errors.

Instead of continuing to focus on error detection on a software level, a Data Specification Architecture [28] is presented, which adds a comprehensive set of data value properties to the data value itself in form of tags. By forcing the system designers to focus on data and its properties in a very early phase of a project, the new architecture provides a high level of error prevention, and the Data Specification Architecture hardware is able to detect 20 data-flow-related error and attack types during runtime in a very simple and reliable way, superior to the state of the art.

2 Identified Data-flow-related Error and Attack Types

To be able to evaluate the state of the art and the effectiveness of the presented Data Specification Architecture, 18 data-flow-related error types have been collected from various sources [8–10, 13–15, 17, 22, 26]. Since embedded systems are facing a rising number of attacks – impressively shown by the Stuxnet worm [12] in 2010 – 2 attack types are added for evaluation, too. The resulting list of the 20 identified data-flow-related error and attack types is shown in table 1, classified in 8 different categories.

3 State of the Art

The state of the art shall be illustrated by three typical examples: conventional architectures like x86 and ARM, ANBD coding and Tagged Memory Architectures. The error detection means provided by each example, the types of errors

Table 1: Collection of data-flow-related error and attack types

Error category	**Error type**
Operand incompatibility	Incompatible data types
	Incompatible units
Violation of range of values and accuracy requirements	Violation of allowed range of values
	Accuracy outside of allowed range
Erroneous operations	Wrong operand selection
	Wrong operation
	Erroneous operation result
Violation of real-time requirements	Violation of deadline
	Premature data update
	Late data update
General data flow errors	Lost update
	Synchronization errors and incomplete data transfers
	Buffer over- or underflows
	Wrong data flow (e. g. wrong receptor)
	Duplicated data
Data corruption by errors or interferences	Data corruption caused by errors or interferences
Erroneous data access	Erroneous data access (missing access rights)
	Usage of non-initialized data
Attack category	**Attack type**
Attacks	Manipulated / crafted data
	Replay attack

they can detect and their limitations are described in the following subsections.

3.1 Conventional Architectures

The conventional architectures x86 and ARM are typical examples for off-the-shelf hardware which is used in millions of applications worldwide. They are designed for maximized data throughput, not for error prevention and detection.

A relatively new protection feature called Memory Protection Extension MPX can be used on x86 systems to detect buffer under- and overflows, but its usage requires a large code overhead. Further, the x86 provides the special instruction `bound` in order to verify a data value to be in a specific range of values. However, this instruction has been omitted in the 64 bit long mode of the processor.

In some systems – especially servers – ECC memory is used to verify the integrity of the stored memory words, realized by using an Extended-(72,64)-Hamming-Code. But only the path between memory, data bus and memory controller is covered by the integrity checks. Processor internal buses and registers remain unsupervised.

3.2 ANBD Coding

ANBD coding is an arithmetic coding, first introduced as AN coding in [5] and extended to ANBD coding in [10]. It can be used to improve error detection on conventional architectures like x86 and ARM. Data values N are coded by adding an integrity check A, an identifier B and a timestamp D resulting in the coded value x_c:

$$x_c = A \cdot x + B + D$$

The integrity of x_c can be checked using a given B and D by verifying, that the equation

$$x_c = A \cdot x + B + D \equiv B + D \mod A$$

is satisfied.

Using ANBD coding, wrong operands, wrong operations and erroneous operations can be detected, but the coding scheme is limited to integer data types. Some coded operations require complex corrections of the results and some operations like divisions are problematic.

3.3 Tagged Memory Architectures

In Tagged Memory Architectures, data values are equipped with additional information in the form of tags, containing additional information about the data value in a hardware-readable format. The basic principle is to add data type information to a memory word, allowing the hardware to verify the compatibility of an operation's operands parallel to the execution of the operation. Additionally, the instruction set can be simplified, since it is no longer necessary to provide dedicated instructions specialized on specific data types.

Descriptor Architectures provide special data types and dedicated instructions to define and safely access arrays in memory, giving the hardware the possibility to check array boundaries, thus buffer under- and overflows can be easily detected by such a hardware.

Capability Architectures add data types to create, identify and protect so called capabilities, which grant subjects – e. g. programs – access to specific objects in memory. This gives the hardware the ability to manage and verify access rights when accessing memory contents.

Known examples of Tagged Memory Architectures are the Telefunken TR 4 [1], the Burroughs B5000 [16] and the Intel iAPX 432 [18].

While tagged memory provides very simple yet powerful means for error detection, most of the features those architectures provided have been forgotten. Recently some projects have been presented using tagged memory to improve data safety. Examples of such projects are Loki [29], CHERI [27], lowRISC [3] and SAFE [21].

4 Data Specification Architecture

Based on the identified error and attack types and the basic principles seen in Tagged Memory Architectures, the following design paradigm has been developed for the Data Specification Architecture – short DSA – to allow a comprehensive and simple hardware based error detection:

All properties describing a data value shall be indivisibly linked to the data value and be transmitted, stored and processed with it; all these properties shall be represented in a hardware readable and verifiable form.

Data was identified to be specified by the following properties:

- the data value,
- the data values accuracy,
- the allowed range of values the data value shall be part of,
- the data type,
- the data value's unit, based on the seven SI base units,
- the access rights regarding readability, writeability and owner of these rights,
- the initialization state, identifying whether or not the data value is valid and can be read,
- the underlying source of the data value or the sources the value is based on, e. g. a sensor,
- the processing path, specifying the permitted path of the data value throughout the system,
- the destination or destinations that are allowed to consume the data value or its processed results, e. g. actors,
- the discrete time step related to the data value,
- the deadline, after which the data value is no longer valid and shall not be processed anymore,
- the cycle time, specifying the earliest and latest moment in time to receive an update of the data value,
- the address or identifier of the data value to be included in integrity checks,
- an integrity check mechanism to verify the data's integrity and
- a cryptographic signature of the data to be able to detect manipulated or crafted data as such.

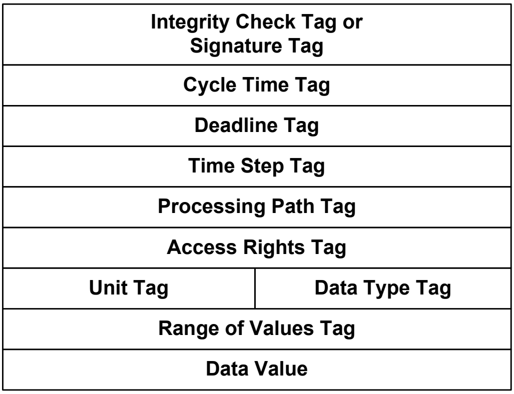


Fig. 1: Data element layout of the DSA

These properties have been added to the data value, resulting in the data element layout shown in figure 1. The name "data element" has been chosen, because a data element is the smallest unit of information and all tags are linked indivisibly to the data value. Instructions are encapsulated in instruction elements, analogue to the data values within the data elements.

A detailed description of all tags in data and instructions elements, their layout and how the DSA hardware evaluates, processes and sets their contents can be found in [28]. In this paper, only the most important novel tags will be presented in detail.

4.1 Range of Values Tag

The Range of Values Tag, shown in figure 2, consists of the two subtags Upper Range Tag URT and Lower Range Tag LRT and specifies the permitted range of values the data value in the data element shall be part of.

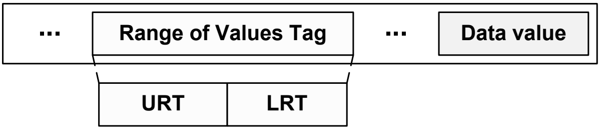


Fig. 2: Data element with Range of Values Tag (with detailed layout) and data value

The tag allows the DSA hardware to do plausibility checks of the data value when reading it from a data element. When writing a new data value to a data element, the DSA hardware verifies that the data value which is to be written is part of the specified range of values. If one of the both checks described fails, the DSA hardware will raise an exception and appropriate error handling can be applied.

4.2 Unit Tag

The Unit Tag, shown in figure 3, specifies the signed exponents of the seven SI base units m, kg, s, A, K, mol and cd, that form the unit of the data value, in seven subtags.

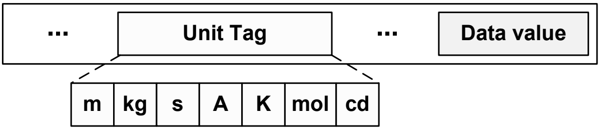


Fig. 3: Data element with Unit Tag (with detailed layout) and data value

While processing data, the DSA hardware will – depending on the type of operation to be applied to the operands – either verify the equality of the units of all source operands, e. g. when executing additions, subtractions and comparisons, or calculate the unit of the operation's result by applying the power laws to the seven subtags of the source operands, e. g. when executing multiplications and divisions.

By calculating and verifying the contents of the Unit Tag, the DSA can reveal the proverbial comparison of "apples to oranges" as an error. If the DSA hardware detects different operand units, it will raise an exception and appropriate error handling can be applied.

4.3 Processing Path Tag

In automation systems, data shall usually follow a pre-defined path from the sensors over processing units to the actors. In a DSA, the data sources specify the intended processing path in the Processing Path Tag. The Processing Path Tag, shown in figure 4, consists of the four subtags Source Subtag S, System

Level Processing Path Subtag PP_{sys}, Local Level Processing Path Subtag PP_{loc} and Destination Subtag D.

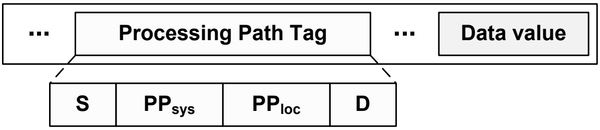


Fig. 4: Data element with Processing Path Tag (with detailed layout) and data value

Every data generating source is assigned a bit position within the Source Subtag, as well as every data sink is assigned a bit position in the Destination Subtag. Every device processing data is assigned a bit position in the System Level Processing Path Subtag. The data processing program instances or hardware blocks within the data processing devices are assigned bit positions within the Local Level Processing Path Subtag. In simple systems, where only one data processing device is present and its internal layout is well known, sources can set the contents of the Local Level Processing Path Subtag, too. In most cases, this isn't possible. Instead the sources do specify any details in the Local Level Processing Path Subtag, leaving it to the individual data processing device to specify the intended local processing path by filling in that information in the Local Level Processing Path Subtag on reception of a data element.

The DSA hardware of each data processing unit will verify that the bit at its bit position within the Processing Path Subtag of each data element being accessed is set. Furthermore, the DSA hardware will verify that the bit at the specific bit position of every data processing instance being executed within such a data processing unit is set, too. Additional checks can be applied by the DSA hardware to verify, that the data values being processed originate from the expected data sources by evaluating the contents of the Source Subtag.

In data sinks, e. g. actors, the DSA hardware will verify that the bit at the sink's bit position is set, indicating that the sink is allowed to consume the data value inside of the data element. Additionally, the DSA hardware of the sinks can compare the data's path throughout the system to the expected path by checking the Source Subtag Processing Path Subtag contents.

If any of the described checks fail, the DSA hardware raises an exception and appropriate error handling can be applied.

4.4 Time Step Tag

In digital systems, data values are generated in discrete time steps. The number of the time step related to the specific data value is stored in the Time Step Tag of the data element in a DSA, as shown in figure 5. It consists of two subtags, the presence bit P and the Time Step Subtag. The P subtag indicates, whether or not the Time Step Subtag contains an evaluable time step.

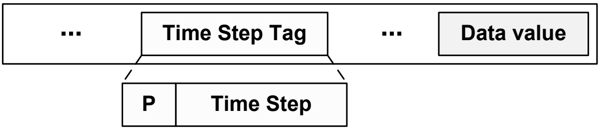


Fig. 5: Data element with Time Step Tag (with detailed layout) and data value

The speciality of the DSA is, that every instruction element has a Time Step Tag, too, as shown in figure 6. It consists of several subtags, two of which will be explained here: the presence bits P and the Δt subtag specifications of temporal relations of the instruction's operands and the Increment Subtag +1.

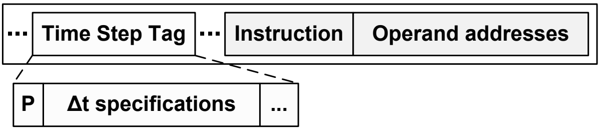


Fig. 6: Instruction element with Time Step Tag (with detailed layout) and instruction

The presence bits P specify, how many temporal relations of operands are to be verified. The Δt subtag specifies the expected time step differences between the contents of the Time Step Tags of the operands of the instruction, which means the temporal relations of the operands. This makes it unnecessary to compare the Time Step Tag contents of the operands to absolute values.

By verifying the temporal relations of the operands, the DSA hardware can detect lost variable updates in loops, filters, sliding averages and a lot of other algorithms. If the expected time step difference is specified as zero, the DSA hardware can reveal synchronization errors as well. If the calculated time step difference doesn't match the difference specified in the Δt Subtag of the Time Step Tag of the instruction, the DSA hardware raises an exception and appropriate error handling can be applied.

4.5 Deadline Tag

Systems with hard real-time constraints are characterized by the need of processing data within specified time frames. Results being available after such a deadline are useless and can lead to dangerous outputs.

In a DSA, the deadline specifying the latest moment in time a data value shall be read is placed in the Deadline Tag, shown in figure 7. On each reading access to a data element, the DSA hardware compares the deadline specified in the Deadline Tag to the current time. This way, a deadline violation will be detected as early as possible and the DSA's hardware will raise an exception, allowing appropriate error handling to be applied.

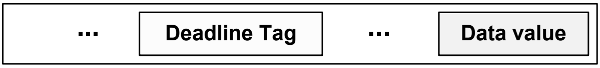


Fig. 7: Data element with Deadline Tag and data value

4.6 Cycle Time Tag

The Cycle Time Tag, shown in figure 8, specifies the earliest – CT_{min} – and latest – CT_{max} – allowed moment in time an update of the specific data value is allowed to be received together with an identifier used to identify the data value.

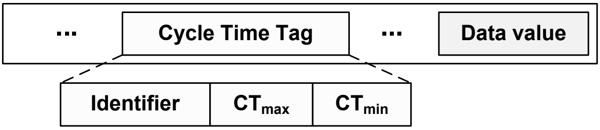


Fig. 8: Data element with Cycle Time Tag (with detailed layout) and data value

The DSA hardware provides a Cycle Supervision Unit which tracks the information provided in the Cycle Time Tag. On each reception of an updated data value the Cycle Supervision Unit verifies that the previously defined earliest allowed moment for an update of that data value – identified using its identifier – has not been violated. Additionally, the Cycle Time Supervision Unit checks whether any of the latest allowed moments in time for a data value's update have passed without receiving a value update. In case of a premature, late or even missing data value update, the Cycle Supervision Unit raises an exception and appropriate error handling can be applied.

4.7 Signature Tag

While some of the Tagged Memory Architectures provided means for checking the integrity of data elements, this doesn't prevent an attacker from crafting or manipulating data elements. To enable the DSA hardware to detect such attempts to influence a system, each device – sensors, data processing units and actors – is assigned a pair of cryptographic keys. Using these keys, each data element is signed and the digital signature is stored within the Signature Tag inside of the data element, as shown in figure 9. This allows the DSA hardware to verify the integrity of the contents of a data element in parallel to using its contents. If signature verification fails, the DSA hardware raises an exception and appropriate error handling can be applied.

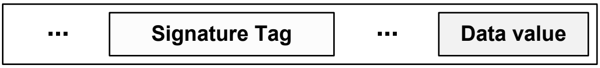


Fig. 9: Data element with Signature Tag and data value

5 Evaluation

The capability of detecting the error and attack types listed in chapter 2 for the state of the art and the new Data Specification Architecture DSA is shown in table 2. The superiority of the DSA is clearly visible, since it is able to detect all of the 20 identified data-flow-related error and attack types.

Table 2: Evaluation of the state of the art and the Data Specification Architecture DSA

Error type	**x86, ARM**	**ANBD**	**TMA**	**DSA**
Incompatible data types	-	-	+	+
Incompatible units	-	-	-	+
Violation of allowed range of values	○ x86	-	-	+
Accuracy outside of allowed range	-	-	-	+
Wrong operand selection	-	(+) BD	-	+
Wrong operation	-	(+) BD	-	+
Erroneous operation result	-	(+) BD	-	+
Violation of deadline	-	-	-	+
Premature data update	-	-	-	+
Late data update	-	-	-	+
Lost update	-	(+) D	-	+
Synchronization errors and incomplete data transfers	-	(+) D	-	+
Buffer over- or underflows	(+) x86	(+) D	+ DS	+
Wrong data flow (e. g. wrong receptor)	-	-	○ BA	+
Duplicated data	-	-	-	+
Data corruption caused by errors or interferences	○	(+) AN	+	+
Erroneous data access (missing access rights)	○	-	+ BA	+
Usage of non-initialized data	-	(+) BD	○	+
Attack type				
Manipulated / crafted data	-	-	-	+
Replay attack	-	-	-	+

Detection: - not possible, ○ partially possible, (+) possible with limitations, + possible
TMA: Tagged Memory Architectures; DSA: Data Specification Architecture

References

[1] AEG Datenverarbeitung: TR 4 Bedienungshandbuch

[2] AIRBUS: Fly-by-wire; http://www.aircraft.airbus.com/innovation/proven-concepts/in-design/fly-by-wire/

[3] A. Bradbury, G. Ferris, R. Mullins: Tagged memory and minion cores in the lowRISC SoC; 2014; http://www.lowrisc.org/docs/

[4] R. C. Baumann, E. B. Smith: Neutron-Induced Boron Fission as a Major Source of Soft Errors in Deep Submicron SRAM Devices; Reliability Physics Symposium, 2000.

[5] D. T. Brown: Error Detecting and Correcting Binary Codes for Arithmetic Operations; IRE Transactions on Electronic Computers; Vol. EC-9, Issue 3; 1960

[6] M. Broy: Challenges in Automotive Software Engineering; ICSE ’06 Proceedings of the 28th international conference on Software engineering, pp. 33–42; 2006

[7] R. N. Charette: This Car Runs on Code; http://spectrum.ieee.org/transportation/systems/this-car-runs-on-code; 2009

[8] CODENOMICON: The Heartbleed Bug; http://heartbleed.com/; 2014

[9] B. Dasarathy: Timing Constraints of Real-Time Systems: Constructs for Expressing Them, Methods of Validating Them; IEEE Transactions on Software Engineering, Vol.11, Issue 1; 80–86; 1985

[10] P. Forin: Vital Coded Microprocessor Principles and Application for Various Transit Systems; 1989; IFAC Control, Computers, Communications; S. 79–84; Paris

[11] S. Ramesh: Software’s Significant Impact on the Automotive Industry; Frost & Sullivan Market Insight; 2008 Proceedings, 38th Annual 2000 IEEE International; pp. 152–157; 2000

[12] D. Kushner: The Real Story of Stuxnet; http://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet

[13] N. G. Leveson, C. S. Turner: An Investigation of the Therac-25 Accidents; Computer, Vol. 26, Issue 7; S. 18–41; 1993

[14] J.-L. Lions et al.: Ariane 501 Inquiry Board report; 1996; http://esamultimedia.esa.int/docs/esa-x-1819eng.pdf

[15] Mars Climate Orbiter Mishap Investigation Board Phase I Report; November 10, 1999; ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf

[16] A. Mayr: The Architecture of the Burroughs B5000 - 20 Years Later and Still Ahead of the Times?; 1982; http://www.smecc.org/The%20Architecture%20%20of%20the%20Burroughs%20B-5000.htm

[17] MITRE Corporation: 2011 CWE/SANS Top 25 Most Dangerous Software Errors; https://cwe.mitre.org/top25/archive/2011/2011_cwe_sans_top25.pdf

[18] G. Myers: Advances in Computer Architecture; 2. Auflage, 1978; John Wiley & Sons; ISBN 0-471-07878-6

[19] NISSAN: Nissan Pivo Concept Press Kit: Overview; http://nissannews.com/en-US/nissan/usa/releases/435dd488-658e-433a-a57a-cd0184e4b51c

[20] E. Normand: Single Event Upset at Ground Level; IEEE Transactions on Nuclear Science, Vol. 43, Issue 6; pp. 2742–2750; 1996

[21] SAFE: A secure computing platform; http://www.crash-safe.org/

[22] SAFECode, S. Simpson et al.: Fundamental Practices for Secure Software Development; 2. Auflage, 2011; http://www.safecode.org/publications/SAFECode_Dev_Practices0211.pdf

[23] U. Schiffel: Hardware Error Detection Using AN-Codes; 2011; PhD thesis; Technische Universität Dresden

[24] SEMI: Why Moore Matters; http://semi.org/en/node/55026; 2015

[25] N. Shimizu: Nissan Puts Steer-by-Wire on the Road: An In-Depth Look at the Technology; Nikkei BP Japan Technology Report / A1403-058-005

[26] J. Teller: Problematik der Datenflussfehler; Angewandte Informatik, Ausgabe 29, Nr. 6; S. 240–247; 1987

[27] University of Cambridge: Capability Hardware Enhanced RISC Instructions (CHERI); https://www.cl.cam.ac.uk/research/security/ctsrd/cheri.html

[28] S. Widmann: Eine Datenspezifikationsarchitektur – Methoden zur Datenflussüberwachung in sicherheitsgerichteten Echtzeitsystemen; Dissertation; FernUniversität in Hagen; 2017

[29] N. Zeldovich, H. Kannan, M. Dalton, C. Kozyrakis: Hardware Enforcement of Application Security Policies Using Tagged Memory; Stanford University; 2008; http://www.scs.stanford.edu/~nickolai/papers/zeldovich-loki.pdf

A Survey of Traffic Flow Modelling Concepts: Pros and Cons of Traditional Models

Nkiedel Alain Akwir[1], Muhindo Kule Mutengi[2], Antoine Kayisu Kazadi[3,4], Meera K. Joseph[4], Jean Chamberlain Chedjou[1], and Kyandoghere Kyamakya[1]

[1]Institute of Smart Systems Technologies, Transportation Informatics Group (TIG), Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

[2]Institut Supérieur de Techniques Appliquées (ISTA), Goma, DR Congo

[3]Université de Kinshasa, Faculté Polytechnique, Kinshasa, DR Congo

[4]University of Johannesburg, Faculty of Engineering and the Built Environment, Johannesburg, South Africa

Abstract: This paper presents a comprehensive survey of traffic flow modelling concepts according to the latest state-of-the-art. Several models have been proposed by the relevant literature depending upon different scenarios, states of the traffic, phenomena and some specific requirements of the modelling process, which have been clearly presented. We mostly consider macroscopic traffic flow models, which present the traffic as a continuum and express the variations of the flow, the density and the speed (the fundamental traffic flow parameters) over space and time through partial differential equations. Different models are presented, whereby both respective strong and weak points are discussed.

1 Introduction

Traffic flow has a very complex and nonlinear behavior since it involves a lot of stakeholders such as: *the interacting different types of vehicles, the unpredictable human behavior, the number of lanes on the road, the conflicting areas like junctions and ramps, the road design, the pedestrians, just to name a few.* These stakeholders lead to different kinds of traffic phenomena like *shockwaves, rarefaction waves, stop-and-go waves, capacity drop, hysteresis, … etc.* Those phenomena lead to different states of the traffic including the under saturation, saturation and over saturation state, the traffic jam, the platoons, the bottleneck, different kind of incidents, … etc. All these phenomena and states make the traffic flow very difficult to model and one can only rely on restrictions depending upon the level of details, the scale of the application, the modelling process (scale of the independent variables or the representation of the process, ... etc.).

The basic/seminal model for traffic flow was proposed by Lighthill-Whitham (1955) [1] and Richards (1956) [2], the so-called LWR's model. This model, also known as first order model, is based on the continuity equation from the compressible dynamics theory, which expresses the conservation of a flowing quantity from one point to another. Despite the fact that the LWR model can reproduce the formation, propagation and evolution of a shockwave, it cannot reproduce the non-equilibrium traffic flow as the speed in the model cannot deviate from the equilibrium speed (variation of the speed over space and time equals zero [3]). We will notice that, nevertheless, this model is still useful when it comes to model complex configurations such as multiclass [4, 5] and/or multilane traffic flow models [6].

To face this model, Payne [7] proposed the first high-order model which in addition to the LWR models a second equation expresses the dynamics of the speed. He has introduced two terms that are responsible of the speed dynamics of the traffic: the relaxation term, which shows how the traffic dynamics always tends to the equilibrium state, and the anticipation term expresses the interaction between cars. The drawback of this model is that it completely follows the behavior of the fluid flow, while the traffic has a different behavior. The so-called anisotropy principle [8] is not respected. To solve these issues, researchers such as Ross (1988) and Del Castillo (1993) have proposed a lot of alternatives focusing on the anticipation term but fail to solve the anisotropy principle. Later, new models that can solve this issue started to be constructed; the first model has been proposed by Zhang [9]. Jiang et al. [10] developed a speed-gradient (SG) model from the full velocity difference model FDV [11]. Gupta and Katiyar [12] developed a modified anisotropy model and some second order models for multiclass respectively multilane configuration have been proposed by Gupta and Katiyar [13] respectively by Tang et al. [14].

In this current work/paper we first propose a classification of traffic flow models, which helps to categorize a model by targeting a specific state, behavior or phenomena. Then, we propose, chronologically, different models focusing on the macroscopic level of details which are represented in form of partial differential equations that can be transformed into a system of conservation laws.

2 A Classification of Traffic Flow Models

Traffic flows models can be classified in different manners. The following are some of the most considered classes:

- **Scale/nature of the independent variables**
  - *Continuous:* The traffic parameters (density, flow and speed) are continuous and the models are generally represented in form of differentials equations (ODEs and/or PDEs).
  - *Discrete:* The space is discretized in cells (cell transmission model [15]).
  - *Semi-discrete:* the space is discretized into cells in which traffic parameters are considered continuous. (e.g. METANET model [16]).
- **Level of details**
  - *Microscopic:* All vehicles are described individually. The models define the behavior of single vehicles such as car-following or lane change models.
  - *Macroscopic:* These models consider the traffic flow as a continuum and use aggregate variables to describe the traffic. They are based on the hydrodynamic theory.
  - *Mesoscopic:* Combination of macroscopic and microscopic models. They are based on the gas-kinetic theory, which allows following one or a group of vehicles in the continuum.
- **Representation of the processes**
  - *Deterministic:* The traffic state or the output of the model remains the same after several simulations while/when using same initial conditions, same inputs and same boundary conditions.
  - *Stochastic:* The traffic state or output of the model gives different results for several simulations under the same inputs conditions. The model contains at least one stochastic parameter.

- **Physical interpretation**
  - *White box:* Deductive approach. Derivation of physical equations that describe the relationships between different states of the traffic. We have the knowledge of all the parameters of the model. They are described in form of PDEs and ODEs.
  - *Black box:* Inductive approach. Input and output data of a system are adjusted in order to fit with the expected model data. We don't have the knowledge of the parameters of the model. The model may take the form of a neural network, a toolbox, and/or a simulation tool.
  - *Grey box:* Combination of Black-Box and White-Box modelling. In this case we exploit mathematical models in which we used inductive approach to derive parameters of the model exploiting measurement data for instance.
- **Scale of application**
  - *Stretches:* A traffic model which considers a portion of road.
  - *Links:* A traffic model which considers a road segment, which connect junctions or nodes.
  - *Intersections:* Models which consider an area which experiences different and conflicting directions of the traffic (junctions or nodes).
  - *Networks:* A traffic model which considers links and intersections following a specific topology.
- **Operationalization**
  - *Analytical:* A traffic flow model expressed in form of a closed mathematical form which is mostly represented in form of a function of a dependent variable like density, flow or speed with respect to independent variable(s) like space and time.
  - *Simulation:* Also known as numerical model, it is expressed in form of a series of numbers that are generally computerized.

3 First-order Models

If one looks into the traffic flow from a very long distance, the flow of fairly heavy traffic appears like a stream of a compressible fluid. This means that the number of vehicles is large enough to consider the system as a continuum, in contrast with microscopic models where each vehicle is followed in its motion. Therefore, a macroscopic theory of traffic can be developed with the help of the hydrodynamic theory of fluids by considering the traffic as an effectively one-dimensional compressible fluid. The traffic state is described based on the density (the number of vehicles per unit of length), the speed (the rate of variation of the position per unit of time) and the flow (the number of vehicles per unit of time).

More than five decades ago, the first traffic flow model was released by Lighthill Whitham and Richard (1956) and named LWR's model. This model is first constituted by the continuity equation as the one which describes the motion of a conservative quantity as follows:

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{1}$$

where k respectively q denote the density respectively the flow. These quantities are involved, along with the speed u, in the fundamental relation of traffic flow which is expressed as follows:

$$q = k\,u \tag{2}$$

Different empirical studies have been conducted during years to derive a model (function) that expresses the relationship between the speed and the density $u = u(k)$ (speed versus density). It has been noticed that generally the function is monotonously decreasing, which expresses the fact that when the density is low the interaction between vehicles is weak and they move freely in the road at a speed u_f called **free flow speed** and when the density is high, the interaction between vehicles is strong, the speed decreases until the vehicles stop when the **jam density** k_j is reached. From the function $u = u(k)$ and Eq. (2); it is easy to derive the functions $q = q(k)$ (flow versus density) *and* $q = q(u)$ (flow versus speed). These last mentioned three functions are called the fundamental diagrams of traffic flow. The seminal models that have been proposed by literature form the Greenshields's model [17] expressed as follows:

$$u = u(k) = u_f - \frac{u_f}{k_j}k \tag{3}$$

$$q = q(k) = k\,u_f - \frac{u_f}{k_j}\,k^2 \tag{4}$$

$$q = q(u), \quad u^2 = uu_f - \frac{u_f}{k_j}\,q \tag{5}$$

Several other models have been proposed such as the Greenberg's model [18], the Underwood's model [19], the energy conservation model [20] etc., just to name a few.

Finally, the LWR's model is expressed as a combination of the continuous equation and the fundamental diagram as follows:

$$\frac{\partial k}{\partial t} + \frac{\partial q(k)}{\partial x} = 0 \tag{6}$$

As mentioned before, this equation meets the hyperbolic systems of conservation laws which are time-dependent systems of partial differential equations [21]. We will see further that most of (all) the macroscopic models can be transformed into hyperbolic systems of conservation of laws and, thus, we can provide the primitive and the conservative form.

Inconvenience of this model (i. e. the LWR's model):

Although this model is very informative, theoretically clear and can reproduce some phenomena such as the formation, propagation and evolution of shockwave, it does have some weaknesses. Indeed, it suffers from a lot of drawbacks. The main drawbacks are the following:

- This model is adapted instantly to the fundamental diagram (the flow-density relationship expresses the equilibrium traffic state), in other words: "the traffic is always at the equilibrium state where the speed is constant". Let us mention that the traffic is at equilibrium state when the following conditions [22] are fulfilled:

  a) Stationarity

$$\begin{cases} \frac{\partial k}{\partial t} & = & 0 \\ \frac{\partial u}{\partial t} & = & 0 \end{cases} \tag{7}$$

b) In addition to (7) (strong equilibrium)

$$\begin{cases} \frac{\partial k}{\partial x} &= 0 \\ \frac{\partial u}{\partial x} &= 0 \end{cases} \tag{8}$$

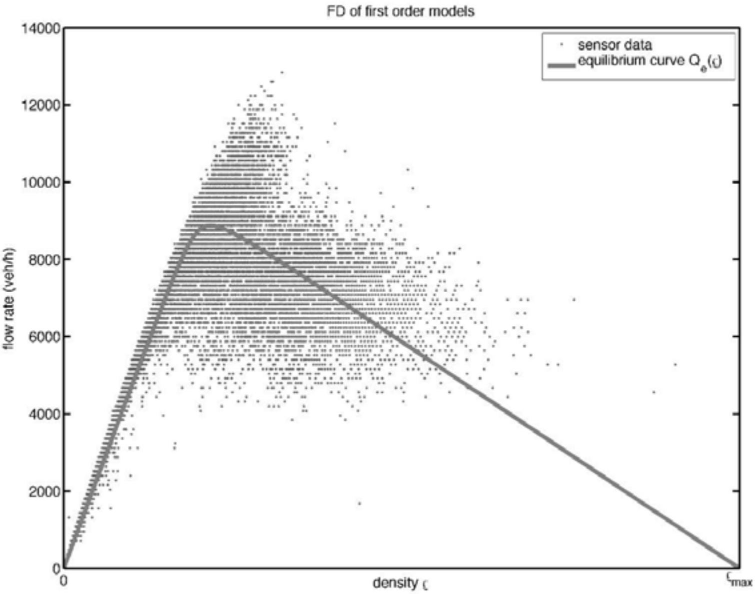


Fig. 1: Fundamental diagram "flow versus density": The dots show the sensor data and the curve shows the equilibrium traffic [23]

Non-equilibrium (realistic traffic flow) traffic conditions show wide scatter in the fundamental diagram as shown in the figure.

- In this model, the change of the density leads instantly to the change of the speed which implies an infinite acceleration. It is, however, more realistic that the traffic flow is adapted after a certain time delay (reaction time).
- This model cannot highlight a lot of traffic phenomena such as stop-and-go-wave, breakdown, bottleneck, traffic hysteresis, just to name a few.

4 Second-order Models

To face the constant speed and the infinite acceleration (due to the change of density) issues, the second-order models have been proposed to express the variation of the traffic speed based on the momentum approach which expresses the acceleration as follows:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = a \tag{9}$$

This expression (Eq. 9), in addition to equation (1) and (2), gives the second order of the traffic flow which expresses the variation of the three fundamentals parameters of the traffic. The question to be asked is to know what influences the variation of the speed in the traffic flow. The two following terms do generally influence the speed of the traffic:

1. The first one is the **relaxation term a_r** which expresses the fact that the traffic always tends towards the equilibrium speed. This term obviously depends on the speed and the density. Another term is the relaxation time τ which is the time it takes from the actual/current speed to the desired one. The relaxation term is expressed as follows:

   $$a_r(k,u) = \frac{V_e(k) - u}{\tau} \tag{10}$$

   where $V_e(k)$ is the desired speed (equilibrium speed from fundamentals diagram like Greenshields).

   Some models which consider only the relaxation term are the following:

   - Phillips's model (1979): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau}$
   - Ross's model (1988): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{u_f - u}{\tau}$
   - Liu et al. (1998): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau(k)}$

2. The second one is the **anticipation term a_a** which expresses the interaction of a vehicle with its surrounding vehicles. Apart from the density k and its partial derivative over space, this term may depend on several parameters according to one or another modelling approach.

Some of those approaches are expressed as follows:

- Payne (1971): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau} + \frac{\boldsymbol{V_e'}}{\boldsymbol{2\tau k}}\,\frac{\boldsymbol{\partial k}}{\boldsymbol{\partial x}}$
- Whitham (1974): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau} - \frac{\boldsymbol{c_0^2}}{\boldsymbol{k}}\,\frac{\boldsymbol{\partial k}}{\boldsymbol{\partial x}}\,,$

  where c_0 is the propagation velocity.
- Phillips (1979): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau} - \frac{\boldsymbol{p'}}{\boldsymbol{k}}\,\frac{\boldsymbol{\partial k}}{\boldsymbol{\partial x}}\,,$

  where p is the traffic pressure (analogue with fluid dynamics) $p = c_0^2 k$.
- Del Castillo (1993):

  $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau} \boldsymbol{-k(V_e'(k))^2 \exp\left(\frac{1}{b}(V_e(x) - u)\right) \frac{\partial k}{\partial x}}$
- Zhang (1998): $\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau} \boldsymbol{-k(V_e'(k))^2\,\frac{\partial k}{\partial x}}$

3. The third is the **vicious term** where there is the viscosity coefficient μ_0 which reflects some observations of vicious traffic flow in reality. The model proposed by Khune et al. (1987) is given as follows:

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = \frac{V_e(k) - u}{\tau} - \frac{c_o^2}{k}\,\frac{\partial k}{\partial x} + \frac{\boldsymbol{\mu_0}}{\boldsymbol{k}}\,\frac{\boldsymbol{\partial^2 u}}{\boldsymbol{\partial x^2}}$$

The main drawback of this model is that it completely follows the behavior of the fluid flow while the traffic has a different behavior. The so-called anisotropy principle as mentioned before [24] is not respected. This principle means that a vehicle cannot interact with all its surrounding vehicles as it is the case for fluid particles but only with the vehicle in front and the consequence is that one shockwave speed may be greater than the vehicles's speed, which is not realistic. This is shown by the characteristic speeds which are greater than the actual speed.

The Payne's Whitham model is used in conservative form to illustrate the anisotropy principle as follows:

The conservative form is given by:

$$\frac{\partial Q}{\partial t} + \frac{\partial F(Q)}{\partial x} = 0\,, \tag{11}$$

where $Q = \begin{pmatrix} k \\ q \end{pmatrix}, \quad F(Q) = \begin{pmatrix} q \\ \frac{q^2}{k} + c_0^2\,k \end{pmatrix}.$

Let us rewrite the conservative form as follows:

$$\frac{\partial Q}{\partial t} + A(Q)\frac{\partial Q}{\partial x} = 0\,, \tag{12}$$

where $A(Q)$ is the jacobian matrix of $F(Q)$ from the primitive form and expressed as follows:

$$A(Q) = \frac{\partial F}{\partial Q} = \begin{pmatrix} 0 & 1 \\ c_0^2 - \frac{q^2}{k^2} & 2\,\frac{q}{k} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ c_0^2 - u^2 & 2\,u \end{pmatrix} \tag{13}$$

Finally, by solving for the eigenvalues from $|A(Q) - \lambda I| = 0$, we obtain the following two distinct and real eigenvalues:

$$\lambda_{1,2} = u \pm c_0 \tag{14}$$

Since c_0 is always positive, one of the eigenvalues $(\lambda_1 = u + c_0)$ is greater than the traffic speed. The anisotropy principle is not respected. The model shows the following behavior:

- Negative speeds
- Perturbations travel faster than the speed of cars

5 Alternatives Taking into Account the Anisotropy Principle

New models that can solve this issue (i. e. respect the anisotropy principle) have been proposed by some researchers such Aw and Rascle [25], by Zhang [26], and Jiang et al. [10], just to name a few. For instance, in the Aw and Rascle's model we exploit the lagrangian point of view (i. e: an internal observer sees a decreasing density in front of him if, for example, cars ahead move faster) instead of the Eulerian point of view (i. e.: an external observer sees an increasing vehicle density in a certain fixed spatial position). By applying the material derivative (convective derivative) to the anticipation term of the Payne-Whitham's model, after mathematical transformation, we obtain the Aw-Rascle model as follows:

$$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial (k\,u)}{\partial x} = 0 \\ \frac{\partial}{\partial t}(u + P(k)) + u\frac{\partial u}{\partial x}(u + P(k)) = 0 \end{cases} \tag{15}$$

where $p(k) = k^{\gamma}$ is a smooth and increasing function of density with $\gamma > 0$.

Its characteristic speeds (eigenvalues of the jacobian of the flux function) are given as follows:

$$\lambda_1 = u - \gamma\, p(k) \quad \text{and} \quad \lambda_2 = u \tag{16}$$

The eigenvalues are less than the traffic speed ($\lambda_1 = u - \gamma\, p(k) < u$ and $\lambda_2 = u$). The anisotropy principle is respected.

In Table 1, some further models which respect the anisotropy principle are presented.

Table 1: Second order traffic flow models with their respective eigenvalues (characteristic speeds)

Models	**Equations**	**Eigenvalues (characteristic speeds)**
Greenberg, 2001	$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial (k\,u)}{\partial x} = 0 \\ \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{V_e - u}{\tau} - k\,V_e' \frac{\partial u}{\partial x} \end{cases}$	$\lambda_1 = u + k\,V_e' \,<\, u, \lambda_2 = u$ anisotropy principle respected
Jiang et al., 2002	$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial (k\,u)}{\partial x} = 0 \\ \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{V_e - u}{\tau} + c_0 \frac{\partial u}{\partial x} \end{cases}$	$\lambda_1 = u - c_0 \,<\, u, \lambda_2 = u$ anisotropy principle respected
Zhang, 2002	$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial (k\,u)}{\partial x} = 0 \\ \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -k\,V_e' \frac{\partial u}{\partial x} \end{cases}$	$\lambda_1 = u - k\,p'(k) \,<\, u,$ $\lambda_2 = u$ anisotropy principle respected
Xue and Dai, 2003	$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial (k\,u)}{\partial x} = 0 \\ \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{V_e - u}{\tau(k)} + \frac{k\,t_r}{\tau(k)}\, V_e' \frac{\partial u}{\partial x} \end{cases}$	$\lambda_1 = u + k\,\frac{t_r}{\tau(k)}\, V_e' \,<\, u,$ $\lambda_2 = u$ anisotropy principle respected
Gupta and Katiyar, 2005 [27]	$\begin{cases} \frac{\partial k}{\partial t} + \frac{\partial (k\,u)}{\partial x} = 0 \\ \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{V_e - u}{\tau} - 2\,\beta\, c(k) \frac{\partial u}{\partial x} + \\ \frac{V_e'}{\tau}\left[\frac{1}{2\,k}\frac{\partial k}{\partial t} + \frac{1}{6\,k^2}\frac{\partial^2 k}{\partial x^2} - \frac{1}{2\,k^3}\left(\frac{\partial k}{\partial t}\right)^2\right] \end{cases}$ where $c^2(k) = -\frac{V_e'}{2\,\tau}$	$\lambda_1 = u + \left(\beta + \sqrt{1+\beta^2}\right)$ $c(k) \,<\, u,$ $\lambda_2 = u + \left(\beta - \sqrt{1+\beta^2}\right)$ $c(k) \,<\, u$ anisotropy principle respected for some values of β which is called the anisotropic term

6 Conclusions

Traffic flow modelling is not an easy task since it involves a lot of stakeholders as mentioned before, amongst which the human behavior which is a very complex system. We only rely on restrictions (from classification (see. Section 2) which can allow us to model specific scenarios according to given requirements. Although the first order model (LWR's model) is unrealistic and made by a lot of assumptions (constant speed, infinite acceleration, . . . , etc), it is still useful when it is about modelling very complex scenarios such as multiple lanes traffic [6], multiclass of vehicles traffic [4] or very complex structures/networks of roads [28]. It is also exploited in grey-box modelling (inverse problems) to determine the traffic parameters which fit with a given scenario for a given model.

The second-order model is basically the most used one since it is much more realistic than the first-order model. A lot of research is still being conducted to improve the second-order model by targeting specific scenarios/states of the traffic. The anisotropy principle is always verified to avoid the "negative speeds" issues which are unrealistic. The continuity equation of the second-order model remains the same except for models with ramps metering [29] where we have terms expressing vehicles entering or living the highway through ramps in the continuity equation. In the second equation of the second-order models (exploiting the momentum approach), which expresses the variation of the speed changes, more and more parameters are added according to a targeted behavior such as: the varying conditions of the traffic [30, 31], models with adaptive cruise control (ACC) and cooperative cruise control [29], or a traffic model with a consideration of driver's reaction time and distance [32].

References

[1] Lighthill MJ, Whitham GB, "On kinematic waves: II. A theory od traffic flow on long crowed roads.", in: *Proceeding Royal Society*, London, 1955.

[2] Richards PI, "Shock waves on the highway", *Operational Research*, vol. 4, pp. 42-51, 1956.

[3] Kessels F W; Lint H; Vuik K and Hoogendoorn S., "Genealogy of traffic flow models", *Euro Journal on Transportation and Logistics*, vol. 4, no. 4, pp. 445–473, 2015.

[4] G.C.K Wong, S.C Wong, "A multi-class traffic flow model – an extension of LWR model with heterogeneous drivers", *Transportation research Part A: policy and Practice*, vol. 36, no. 9, pp. 827–841, 2002.

[5] D. Ngoduy , R. Liu, "Multiclass first-order simulation model to explain non-linear traffic phenomena", *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 667–682, 2007.

[6] M H Kabir and L S Andallah , "NUMERICAL SOLUTION OF A MULTILATRAFFIC FLOW MODEL", *GANIT J. Bangladesh Math. Soc*, vol. 33, pp. 25–32, 2013.

[7] P. HJ., "Models of freeway traffic control", In: *Mathematical models of Public System*, Bekey GA (ed), Simulation Councils, inc.: La Jolla, California, pp. 51–61, 1971.

[8] Beneditto Piccoli and Andrea Tosin, "A review of continuum mathematical models of vehicular traffic", *Mathematics of complexity and Dynamical Systems*, pp. 1748–1770, 2012.

[9] Z. HM, "A theory of nonequilibrium traffic flow", *Transportation Research Part B*, vol. 32, no. 7, pp. 485–498, 1998.

[10] Rui Jiang, Qing-Song Wu , Zuo-Jin Zhu, "A new continuum model for traffic flow and numerical tests", *Transportation Research Part B: Methodological*, vol. 36, no. 5, pp. 405–419, 2002.

[11] Jiang R, Wu QS, Zhu ZJ, "Full velocity difference model for car-following theory", *Physical Reviewn E*, vol. 64, 2001.

[12] Gupta A. K. and Katiyar V.K , "Phase transition of traffic staes with on-ramp", *Physica A: Statistical Mechanics and Its Applications*, vol. 371, no. 2, pp. 674–682, 2006.

[13] Gupta A.K. and Katiyar VK, "A new multi-class continuum model for traffic flow.", *Transportmetrica*, vol. 3, pp. 73–85, 2007.

[14] Tang T.Q Wong S.C., Huang H.J and Zhang P, "Macroscopic modeling of lane-changing for two-lane traffic flow", *Journal of advanced Transportation*, vol. 43, no. 3, pp. 245–273, 2009.

[15] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory", *transportation Reasearch B*, vol. 28B, no. 4, pp. 269–287, 1994.

[16] Frejo J. R. D. ; Camacho E. F. and Horowitz R., "A parameter identification algorithm for the METANET model with a limited number of loop detectors", in: *51st IEEE Conference on Decision and Control*, Maui, Hawaii, 2002.

[17] B. D. Greenshields, J. R. Bibbins, W. S. Channing and H. H. Miller, "A study of traffic capacity", in: *Proceedings of the Fourth Annual Meeting of the Highway Reseach Board*, Washington, 1934.

[18] H. Greenberg, "An analysis of traffic flow", *Oper. Res.*, vol. 7, no. 1, pp. 79–85, 1959.

[19] R. T. Underwood, *"Speed, Volume, and Density Relationships: Quality and Theory of Traffic Flow, Yale Bureau of Highway Traffic"*, New Haven, CT, USA: Yale Univ. Press, 1961.

[20] D. Wang, X. Ma, D. Ma and S. Jin, "A Novel Speed-Density Relationship Model Based on the Energy Conservation Concept", *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. PP, no. 99, pp. 1–11, 2016.

[21] R. J. Leveque, *Numerical Methods for Conservation Laws*, Zurich: Birkhauser Verlag AG, 1992.

[22] H.M.Zhang, "A mathematical theory of trac hysteresis", *Transportation Research Part B*, vol. 33, pp. 1–23, 1999.

[23] himao Fan, *Generic Second Order Models*, University of Illinois at Urbana-Champaign, [Online]. Available: http://publish.illinois.edu/shimao-fan/research/generic-second-order-models/. [Accessed 12 July 2017].

[24] C. Daganzo, "Requiem for second-order fluid approximations of traffic flow", *Transportation Research B*, vol. 29B, no. 4, pp. 277–286, 1995.

[25] A. Aw and M. rascle, "Resurrection of "second order" Models of Traffic Flow", *Society for Industrial and Applied Mathematics*, vol. 60, no. 3, pp. 916–938, 2000.

[26] H. Zhang, "A non-equilibrium traffic model devoid of gas-like behavior", *Transportation Research Part B*, vol. 36, pp. 275–290, 2002.

[27] A.K. Gupta and V.K. Katiyar, "A new anisotropic continuum model for traffic flow", *Physica A*, vol. 368, pp. 551–559, 2006.

[28] T.Q. Tang, H.J. Huang, C.Q. Mei and S.G. Zhao, "A dynamic model for traffic network flow", *Physica A*, vol. 387, pp. 2603–2610, 2008.

[29] Delis A.I., Nikolos I.K. and Papageorgiou M., "Macroscopic traffic flow modelling with adaptive cruise control: Development and numerical solution", *Computers and mathematics with Application*, vol. 70, pp. 1921–1947, 2015.

[30] Bellouquid A. and Delitata M., "Asymptotic limits of a discrete kinetic theory model of vehicular traffic", *Applied Mathematics Letters*, vol. 24, pp. 672–678, 2011.

[31] Tang T. -Q; Chen L.; Wu Y.-H and Caccetta L., "A macro traffic flow model accounting for real-time traffic state", *Physica A*, vol. 437, pp. 55–67, 2015.

[32] Nooshin Davoodi ; Ali R. Soheili ; and S. Mehdi Hashemi, "A macro-model for traffic flow with consideration of drivers reaction time and distance", *Nonlinear Dynamics*, vol. 83, pp. 1621–1628, 2016.

[33] F. Siebel and W. Mauser, "On the fundamental diagram of traffic flow", *Industrial and Applied Mathematics*, vol. 66, no. 4, pp. 1150–1162, 2006.

[34] N. Farhi and M. a. Q. J. Goursat, ”Derivation of the fundamental traffic diagram for two circular roads and a crossing using minplus algebra and Petri net modeling.”, in: *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05*. 44th IEEE Conference, 2005.

[35] Y. Kim, PhD Thesis: *Online traffic flow model applying dynamic flow-density relation*, Munich, 2002.

On the Application of the COMPRAM Methodology for an Understanding of Societal Problems

An Application to the Lack of Access to Good Quality Electricity in DR Congo

Kamiba Kabuya Isaac[1], Muhindo Kule Mutengi[2], Antoine Bagula[3], and Kyandoghere Kyamakya[4]

[1]Ecole Supérieure d'Informatique Salama, Lubumbashi, DR Congo

[2]Institut Supérieur de Techniques Appliquées (ISTA), Goma, DR Congo

[3]University of the Western Cape, Department of Computer Science, Cape Town, South Africa

[4]Institute of Smart Systems Technologies, Transportation Informatics Group (TIG), Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Abstract: Access to electricity is recognized as a primary and natural need, which is lacking in the Democratic Republic of the Congo (DRC), where more than 95% of the population live in the dark or have intermittent access to the electricity due to a very dysfunctional power grid. This creates a complex societal problem, which can be addressed by a clear understanding of the level of the problem on one hand, and the efficient analysis of the causes and societal phenomena resulting from it on the other hand. This paper builds around the COMPRAM methodology to define the conceptual model of the problem into a "natural language" and facilitates its comprehension. A causal analysis is used to shed more light on the problem's understanding by identifying its dynamic character through 30 variables linked by their causal links and polarity. The issue and significance of the resulting causal diagram reveals the structure of the problem and prepares for a future system dynamics analysis of the whole system.

1 Introduction

It is widely known that access to electricity is key to address many societal and technological issues and help the developing world to bridge the technology gap in areas such as cities' management, public health, and food security through applications such as smart cities [1, 2], electronic health [3, 4], smart agriculture [5, 6] and many other niche applications, which are still the preserve of the developing world. However, many citizens of developing countries are still either living in the dark or provided poor and/or intermittent access to a dysfunctional electricity grid.

The DR Congo, for example, is a country which has 85 million inhabitants distributed into about 15 million families [7]. Furthermore, it has the third hydroelectrical potential of the world. Nevertheless, it is a country where the lack of access to good quality electricity has become a complex societal problem. Until the end of 2014, DR Congo has had 12 hydroelectric power stations and 68 small thermal power stations distributed in the different provinces and supplying respectively 2417 MW and 28 MW [8] for an installed production capacity of 2,445 MW which is sold in the country at an average price of 76 USD per MWh [8]. The SUSTAINABLE ENERGY FOR ALL (SE4ALL) projection for the DR Congo in 2030 estimates that the electrical energy consumption per year and per inhabitant would increase from 0.08 toe (1394 Kwh / year / hbt) to 0.09 toe 15865 Kwh / year / hbt). Out of approximately 12 million households estimated in 2011, only 1.6 million households had access to electricity in urban areas and 1.2 million in rural areas (this corresponds to almost 1% of the total population of the country) [8]. The dynamic and transformational nature of this problem and its societal and technological implications require the use of a rigorous and efficient methodology for its analysis.

1.1 COMPRAM Methodology

While different methodologies can be used for this analysis, the COMPRAM methodology seems to be a ″natural fit″ for this type of problems as it can provide a clear understanding of the level of the problem on one hand, and the efficient analysis of its causes and societal implications on the other hand. COMplex PRoblem hAndling Methodology (COMPRAM) was invented by Dorien Detombe [9] as a theoretical methodology, which can be used to analyze complex societal problems. It is structured into seven layers dividable in phases and sub-phases. The first COMPRAM layer (Layer 1) can also be subdivided into 9 sub-layers or phases named P1 to P9. For each of these 9 phases, three methodological steps are used: i) the personal thoughts of the analyst (named M1), ii) the documentary research (named M2), and iii) the discussions with relevant selected experts and/or stakeholders (named M3). The focus of this paper lies on phase 2 (that is P2) of the first COMPRAM layer with an emphasis of analyzing the main causes and effects related to the access to electricity as a complex societal problem in DR Congo.

1.2 Contribution and Outline

The main question raised in a previous paper [10] was, whether the lack of access to ″good quality electricity″ and to ″broadband Internet″ can considered a complex societal problem for the DR Congo. In that paper, we defined and described the problem in a ″natural language″ according to a synoptic diagram as a first step in the process of considering the solution to the problem. This current paper takes a step ahead into the analysis but focussing only on the lack of access to electricity. As emphasized by all problem solving methods for different domains which suggest that one cannot implement the solution to a problem without studying the root causes of the latter, we efficiently analyse the causes of the problem and their implications for the actors and the society concerned.

Besides this introduction and the conclusion, the body of text of this article includes two sections, namely the research methodology and the results. In the research methodology, we explain in general terms the COMPRAM methodology and present a use-case associated with the current research focus. In the second section of this paper, we describe the results of our research in two parts: firstly, a presentation of the theoretical framework and secondly its causal diagram.

2 Research Methodology

In the COMPRAM methodology and the dynamics of complex systems, discovering the main causes of a problem also leads to the knowledge of their causality. It also gives a structure to the problem in its abstract dimensions, without, however, describing or presenting the events produced by the variables identified in the causal diagram. To find the main causes and the links to the actors, we remain in the COMPRAM context as described by Detombe [9]. COMPRAM considers three relevant societal dimensions: power, emotions and knowledge. The method relies on a 7-layered problem-solving process, whereby each of the 7 layers communicates with and complements each other. The first layer concerns the development of a conceptual model of the problem. It is a description and clear definition of its various dimensions before considering its change or resolution process [9]. Like any of the others, the COMPRAM first layer can also be divided into phases or sub-phases of understanding, analysis and treatment of the problem. Several graphical tools in diagram representation and several computer tools can be used to represent and understand the problem. The use of computer tools for dynamic modeling of the problem is very common and desired.

In this article, we discuss the second phase we have set ourselves at the level of the first layer concerning the problem of the lack of access to electricity in the DR Congo. In this phase are found the actors involved directly or indirectly, the effective causes of the problem, their effects and the links that bind all these factors or parameters, and culminate in the development of the causal diagram of the problem.

We consider the problem in its multidisciplinary, multidimensional, multi-sectorial and multi-level and multi-actor facets. We have used reflection and personal observations, discussions and documentary research to gather the data for our research. Our basic source of information has been the work done in [10]. After identifying and analyzing the efficient causes of the complexity of the lack of access to electricity in the DR Congo, the elaboration of the causal diagram using the Software Tool Vensim allowed us to show the related causalities in the structure of the problem. The use of an Ishikawa's style diagram seemed to us to be inefficient in this area of analysis of a complex societal problem. In this respect, the dynamics of the systems presented by COMPRAM was in our opinion, the best option.

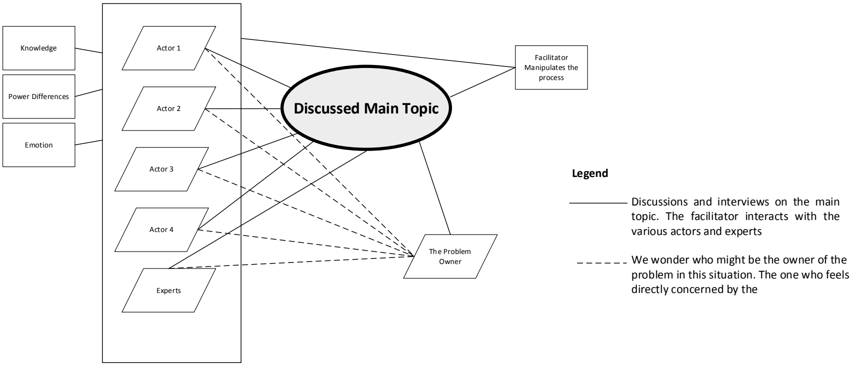

Knowledge
Power Differences
Emotion
Actor 1
Actor 2
Actor 3
Actor 4
Experts
Discussed Main Topic
Facilitator Manipulates the process
The Problem Owner
Legend
Discussions and interviews on the main topic. The facilitator interacts with the various actors and experts
We wonder who might be the owner of the problem in this situation. The one who feels directly concerned by the

Fig. 1: COMPRAM USE-CASE: A brief schematic description of the first two layers of COMPRAM and of the related involved players.

Figure 1 gives an idea of how the process applies the methodology in our current research framework.

This diagram (see Fig. 1) shows that we must first understand the different powers that the actors of the problems have, the emotions raised by the problems at hand and their knowledge of the problem in any process of discussion, observation and reflection on the problem [9].

3 Efficient Causes' Analysis Results

3.1 Context Description

In general, electricity is needed in almost all human daily activities. Houses and buildings, companies' work places and other public places (hospitals, schools, churches, roads, markets, etc.) are electricity powered, provided a good electrical power supply is available. Another worthwhile impact is the revolution that telecommunications and computer science bring to modern society. These two technological building blocks cannot work without electricity. A telephone in its diverse forms has become an instrument, which reduces distances and reaches beyond country borders, enables contacts, accelerates various human activities programs, and assists in organizing the daily activities of a human being. The computer in its various forms (desktop or laptop) brings a huge set of services and useful tools to both individuals and corporations. Providing air conditioning inside buildings and vehicles is another application of the electric power supply. Besides these selected common applications of the electric power, the following paragraphs will cover some other illustrative cases.

Understanding how to address the root causes of the lack of access to electricity in DR Congo requires an analysis of its potential and capacity of supplying the electric power.

Despite its hydro-electrical potential, which can even enable the country to export electricity to other African countries, DR Congo has more than 90% of its population without access to electricity supply from a public power grid. Furthermore, the 10% of the population which have an access to the electricity grid usually experience effective power supply availability between 5% to 10% of the time, i. e. an average of less than 2 hours per day. This poor electricity supply makes this big and rich country look like a cosmic big "dark hole".

In Sub-Saharan Africa, the deficiency of the electric power capacity is a restraint to the development, despite of the energetic potential of the areas [11]. This is a serious concern in DR Congo, where access to electricity is significantly low in large parts of the country. As revealed by the indicators of the World Bank and those of the CAID, there is a big gap between the access rates to electricity in the urban when compared to the rural areas (the rate varies in the margins of 0.5% up to 44% between provinces) [12]. Despite its enormous hydroelectric potential estimated to be around 100 Gigawatts, after Russia and China, the national electrification rate in DR Congo is extremely poor: 9% in 2010, from which 1% for rural population access[1].

It is worth noting that those actors who benefit from the bad situation (no electric grid, low availability in areas where is some electric grid) usually apply a business model, which is essentially based on filling the gap created by the missing access to good quality electricity (GQE) by providing alternative products and services: kerosene, diesel generators, etc. However, many of these alternatives are generally very costly, are neither sustainable nor scalable solutions of the problem and do also severely pollute the environment. In DR Congo, it is widely accepted that the prosperity of such businesses based on alternative solutions is usually an indication of the worsening of electricity problem. To the best of our knowledge, there have been efforts from all layers of the population to denounce the situation, but these efforts seem to have not yet produced the expected effects and sufficient pressure on the relevant deciders and policy makers. A solution to this widely known problem is still to be found.

The daily victims of the situation include citizens in all the layers of the DR Congo's society. They include the following players: families and common people, the public in general, the public health sector, the education institutions, and many others, who are aware that electricity is an enabler for social welfare and other various forms of welfare including psychological, spiritual, intellectual, professional, etc. Missing access to electricity is therefore being deprived from a core basic life right.

According to article 48 of the DR Congo's constitution of February 18th, 2006, electricity is similar to drinkable water and hence a fundamental need for life. Therefore, every person living in DR Congo has the right to have a decent access

[1] Tuesday, April 2nd, 2013, during the reconditioning of the Group 2 of the hydroelectric power plant Inga 1, which had been inert since 16 years, the Minister of Hydraulic Ressources and Energy, Mr Bruno Kapandji, has drawn a non-shining inventory of the electricity service in DRC: "The rate of electricity access in our country is estimated at 9%, this means only 7 millions of privileged from a total of 75 million inhabitants. Moreover, there exists a big disparity between the provinces, this rate varies from 0.5% in the Occidental Kasaï province up to about 45% in the Kinshasa province [13].

to electricity. Unfortunately, the implementation of this article of the constitution is still very far from the reality found on the ground.

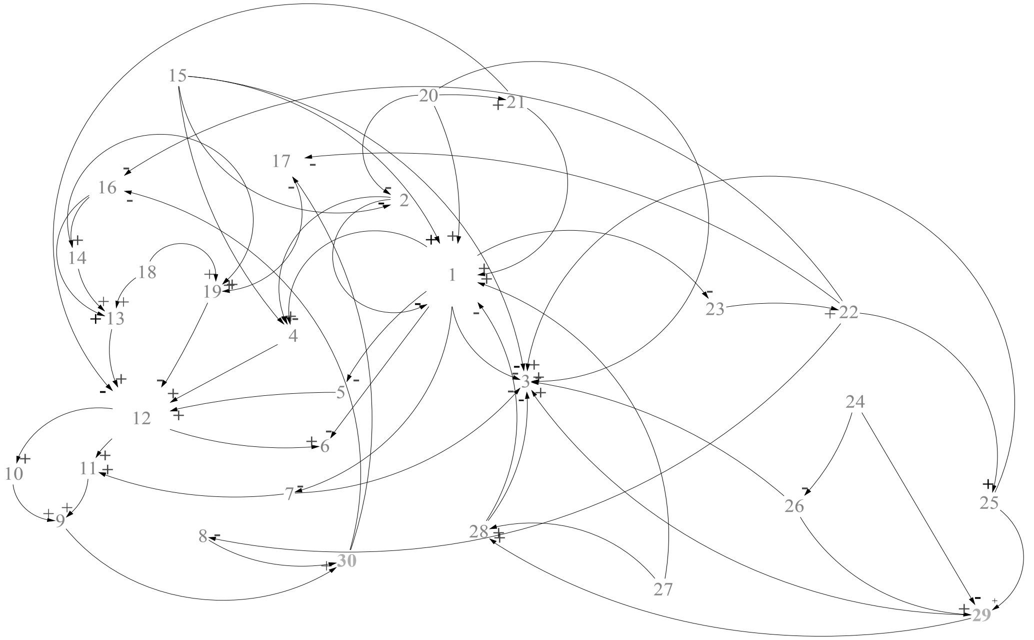
It is a fact that more than 98% of the villages/cities/towns of more than 1,000 inhabitants do crucially face this problem. Though it looks futile, the denunciation efforts have been, somehow, publicly demonstrated by the media and the churches through diverse alert messages.

Based on an understanding of the context in which the DR Congo finds itself in terms of electrical energy and the application of the COMPRAM methodology approach presented in Figure 1, we were able to trace the causal diagram of the problem.

3.2 Causal Loop Diagram (CLD)

Figure 2 shows the Causal Loop diagram (CLD) of the problem of access to good quality electricity through two scenarios: (a) lack of access even though an electricity grid exists nearby and; (b) lack of access due to the total absence of the electrical grid in the environment. This was our starting hypothesis for the construction of this diagram. According to the analysis of the data collected, we have collected 30 variables explaining the double lack of access to electricity in DR Congo. These variables include the following: the primary and secondary causes, the actors, and the relationships that bind them. Indeed, as pointed out in [14], the quality of electric energy used has considerable direct implications for the development of a society. There is a direct relationship between the development of a society and its state of the electricity grid [14]. In today's DR Congo, the lack of access to electricity directly impacts negatively the country's development factors. Indeed, DR Congo, like other Sub-Saharan African countries after independence, has neglected the internal development of its energy infrastructures and has concentrated more on the extensive export infrastructure of its natural resources [15].

The naming of the causal variables, their connections and the polarity of the arrows explain the causal relationships between the actors, the efficient causes and the relevant effects of the problem. This causal diagram shows just the structure of the complexity of the problem, it has a high degree of abstraction of the events that stem from each of the variables. It does not yet explain its very dynamic aspect of the events of the problem; this will be done as a result of the dynamic analysis of the system. The causality and polarity links that bind the variables do not describe the behavior of the variables; they only describe what might happen to the variable if there is a change in sign [16].



1. Companies involved in production, transport and distribution of electric energy (SNEL is the biggest state-owned company of that kind in DR Congo);
2. Non-democratic political class/system
3. Companies selling either diesel/gas power generators or solar panels or firewood, charcoal sellers, the oil men...
4. Bad govement economy policy
5. No Planning for Netwok extension and Upgrades
6. Poor electric network maintenance
7. No network extension/upgrades
8. Badly Operated Distribution Network
9. Low Availability of Electric Energy for end Customers
10. Insufficiency producted electric energy
11. Insufficiency transported electric energy
12. Low inverstiment of SNEL In electric network Infrastructures (Gird) and In electric energy Production Systems (Power Plants)
13. Low income Public Electricity Grid Operator from private Customers
14. Non-payement of electricity fees From some Public Electricity Grid Operator Customers
15. Law and policy makers (e.g. parliament, government, etc.); the common people; the press, the economical operators, etc
16. Number of Private Customers
17. Number of Industrial / Corporate Customer
18. Low price of the Kwh
19. Low income Public Electricity Grid Operator from industrial Customers
20. Policy (by political deciders) related to basic rights to access electricity
21. Planning (strategic long-term Planning by Ministry of Planning and Energy)
22. Poverty (of the population – see macro economy indicators)
23. Bad economic and management govement policy
24. Presence of Public Electricity Grid Operator from industrial in a given Area
25. No Privately Producted Electric Energy
26. No public producted energy
27. The General public opinion
28. Awareness (by population of their right to access electricity)
29. « No electric grid » in a given area where people leave
30. « Poor quality electric grid » in a given area where people leave

Fig. 2: Causal model related to the "lack of access to good quality electricity".

It can be seen that the increasing number of victims who lack access, including families, the man in the street, health institutions and many others depend very much on the number of actions taken by the decision-makers, those who denounce the situation, and those who are aware of the situation.

4 From CLD (Causal Loop Diagram) to SFD (Stock and Flow Diagram) and Related Differential Equations

Starting from the CLD, it is possible to generate SFD. This is the next step of this modelling work in progress; this step will be the focus of our next paper. The SFD will enable a system dynamics modelling of the complex system at hand [16].

The SFD is a analysis consisting of stock, flows and auxiliary variables. The SFD enables the study of behaviors, the time-evolution of the interactions between the variables of the system [9].

A series of software tools can be used for creating the SFD and do support the related system-thinking process: Vensim, Stella, IThink, etc.

The following steps are followed in the global process of converting a LCD into a SFD [17–20]:

1. Writing the variables' names
2. Identifying the flows
3. Selecting the software tool to be used for the implementation of the System Dynamics simulation model (e.g: Vensim, Stella, IThink, etc.)
4. Constructing the system dynamics (SD) diagram
5. Adding the auxiliary variables to the SD diagram
6. Fixing or selecting the dimensions (i.e. unit) for each of the variables
7. Writing or generating the differential equations corresponding to the SD diagram
8. Conducting simulations
9. Analysis of the SD model especially all feedback loops, which do contribute to a good understanding of the system [21]

10. Generating graphs presenting simulation results and their subsequent analysis

It is worth noticing that the basic components of a SFD diagram are: the stock, the flow, the auxiliary variables, and the connectors. Fig. 3 does provide an illustrative example of a very simple SFD case.

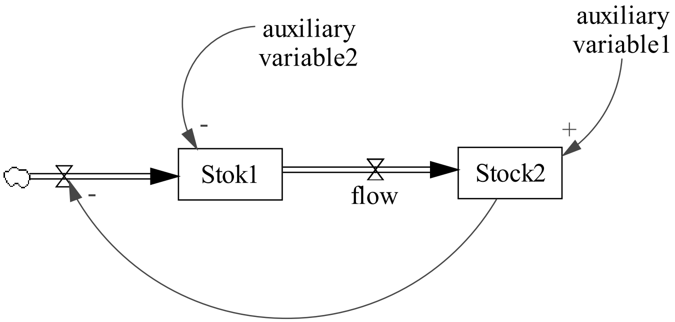


Fig. 3: Illustrative example of a SD "Stock and Flow Diagram", which is the basis brick of SD models.

5 Discussion and Future Work

The access to electricity is a primary and natural need that is heavily lacking in the DR Congo. There are a number of factors and causes that interact and explain the structure of the complexity of this problem. The lack of access to good quality electricity has a dual dimension: one where the grid exists but the access is strictly restricted and/or is of poor quality (esp. poor availability), and another where the grid does not exist at all and the access is not conceivable. This hypothesis is demonstrated to be true by the causal nature of the factors of the problem. Thus, without an in-depth dynamic analysis of the graph, it can be said that the infrastructure for production, transmission and distribution of energy no longer corresponds to the country's reality on the ground w.r.t. geography, demography, and industry, urban and rural contexts.

The demographic growth imposes a similar growth in the households. The current offer of electrical energy does no longer meets the expectations and needs of DR Congo as a country. Even in areas where the electric power grid exists, the of lack of access victims are very numerous, while the presence of those who can benefit from the situation increase.

The inadequacy of the electricity produced regarding both quality and quantity is a serious concern in DR Congo. This does results in a critically poor access to electricity; in fact the public electricity grid operator called SNEL does provide a very poor quality of service.

The lack of sufficient investments in network extension projects over several decades, the non-respect of the right of access to electricity by the authorities, the lack of awareness of the value of this energy to human development, and the absence of adequate investments in electrical energy production are all factors that limit the access. These same factors do not favor the installation of an electricity network in areas where it does not exist. The dynamic nature of all these variables analyzed through the lenses of systems dynamics does well reveal us what the situation will/may be in a couple of decades if no comprehensive solution is found.

References

[1] A. BAGULA, L. CASTELLI, et M. ZENNARO, ″On the design of smart parking networks in the smart cities: An optimal sensor placement model. Sensors″, 15(7):15443–15467, 2015.

[2] E. M. KARBAD, D. DJENOURI, et A. BAGULA, ″Car park management with networked wireless sensors and active rfid, In: *IEEE International Conference on Electro/Information Technology*. IEEE, 2015.

[3] A. BAGULA et al., ″Cloud based patient prioritization as service in public health care″, *Proceedings of the ITU Kaleidoscope 2016*, Bangkok, Thailand.

[4] M. MANDAVA, C. LUBAMBA, A. ISMAIL, H. BAGULA, et A. BAGULA, ″Cyberhealthcare for public healthcare in the developing world″, *Proceedings of the 2016 IEEE Symposium on Computers and Communication (ISCC)*, pp. 14–19, 2016.

[5] M. MASINDE et A. BAGULA, ″A framework for predicting droughts in developing countries using sensor networks and mobile phones″, *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists. ACM*, pp. 390–393, 2010.

[6] M. MASINDE et A. BAGULA, ″The role of ICTs in downscaling and upscaling integrated weather forecasts for farmers in Sub-Saharan Africa″, *ACM*, pp. 122–129, 2012.

[7] PNUD, République Démocratique du Congo. *Plan National de Développement Sanitaire PNUDS 2011-2015*, Kinshasa, 2015.

[8] PNUD, République Démocratique du Congo. *Rapport National. Energie durable pour tous à l'horizon 2030*, Kinshasa, 2013.

[9] D. DETOMBE, *Handling Societal Complexity. A study of the Theory of the Methodology of Complexity and the COMPRAM Methodology*. New York: Springer, 2015.

[10] I. KAMIBA, K. KYAMAKYA, et J. MEERA, "Complex Societal Problem related to the Internet Access and Electricity access in DRC", *IST-Africa 2017 Conference Proceedings*.

[11] C. PIEROU, "Les producteurs privés d'électricité?: une solution pour l'Afrique??", *Proparco N118*, nov-2013.

[12] I MUTALA, "Accès à l'énergie pour tous en RDC, une éclaircie en vue", 02-nov-2015.

[13] PRIMATURE, "Faible taux d'accès à l'électricité en RDC?: sortir des sentiers battus", Avril-2013.

[14] T. ABDULLAH et al., "Modeling the effectiveness of electric energy quality of knowledge management strategy by the systematic approach", Journal of Scientific Research and Development, pp. 198–203, 2015.

[15] S. BOTTTON, "L'accès à l'eau et à l'électricité dans les pays en développements. Comment penser la demande?", *IDDRI N109*, 2006.

[16] J. D. STERMAN, *Business Dnyamics. Systems Thinking and Modeling for a Complex Word*. McGraw-Hill, 2000.

[17] V. H. HÖRDUR, B. SALIM, et U. S. HARALD, "Causal Loop Diagrams-promoting deep learning of complex systems in engineering education", *LTHs 4:e Pedagogiska inspirationskonferensen*, Juin-2006.

[18] B. THOMAS, V. ANDREAS, B. SALIM, V. H. HÖRDUR, et S. MATS, *"Developping System Dynamics Models from Causal Loop"*.

[19] K. TEKNOMO, "System Dynamics Tutorial", evoledu. [En ligne]. `http://people.revoledu.com/kardi/tutorial/SystemDynamic/`.

[20] D. ARONSON et D. ANGELAKIS, "Step-By-Step Stoks and Flows Converting from Causal Loop Diagrams", *Syst. Think.*, vol. 10, no. 6, pp. 6–8, août 1999.

[21] PHILIPPE, "Diagramme de boucles causales", Décembre-2014. [En ligne]. `https://elegkhos.wordpress.com/2014/12/16/diagramme-de-boucles-causales/`.

Parameter-driven Handover Interfaces of Safety-related Design Toolchains

Daniel Koß

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract:
Several steps within the development process of safety-related systems make use of totally different sets of computer-aided engineering tools. Specifications, for example, are usually built by behavioural models or linked text fragments like requirements or use cases, whereas hardware is usually designed with a hardware description language or wiring diagrams. Software is described by a sequential list of instructions based on the available hardware. At first sight, all of these steps seem independent of each other, but use can be made of the handover interfaces from one design step to the next to improve resource usage of the resulting system and, therefore, minimize unused parts and erroneous influences of unnecessary components. It will be shown, how a holistic view from specification via software and hardware design to the whole system can make use of inherent synergies to achieve a resource-adequate system that exactly fits the needs of the specified characteristics.

High-frequency Trading as a Real-time Application

René Schwantuschke

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract:
High-frequency trading (HFT) at stock exchanges is considered as an application in the financial sector for which real-time requirements hold. First, the term HFT is defined an its particularities are mentioned. Then, the ICT infrastructure deployed for HFT at the German stock exchange in Frankfurt is presented and discussed with respect to its suitability. It turns out that this platform was not designed with real-time capability in mind, and that the transport protocol TCP used so far is inadequate for HFT. Finally, proposals are made to achieve real-time capability by employing real-time operating systems, dedicated hardware components and real-time transmission networks.

Index of Authors

Online-Buchshop für Ingenieure

■■ VDI nachrichten

BUCHSHOP

Online-Shops

Fachliteratur und mehr - jetzt bequem online recherchieren & bestellen unter: www.vdi-nachrichten.com/ Der-Shop-im-Ueberblick

Fahrzeugkonzepte für das 2. Jahrhundert Automobiltechnik
Vehicle Concepts for the 2nd Century of Automotive Technology
VDI-Berichte 1653

Täglich aktualisiert: Neuerscheinungen VDI-Schriftenreihen

Fortschritt-Berichte VDI
Weiterentwicklung der Endfunktionsprüfung von Verbrennungskraftmaschinen in der Großserienproduktion

Im Buchshop von vdi-nachrichten.com finden Ingenieure und Techniker ein speziell auf sie zugeschnittenes, umfassendes Literaturangebot.

Mit der komfortablen Schnellsuche werden Sie in den VDI-Schriftenreihen und im Verzeichnis lieferbarer Bücher unter 1.000.000 Titeln garantiert fündig.

Im Buchshop stehen für Sie bereit:

VDI-Berichte und die Reihe **Kunststofftechnik**:

Berichte nationaler und internationaler technischer Fachtagungen der VDI-Fachgliederungen

Fortschritt-Berichte VDI:

Dissertationen, Habilitationen und Forschungsberichte aus sämtlichen ingenieurwissenschaftlichen Fachrichtungen

Newsletter „Neuerscheinungen":

Kostenfreie Infos zu aktuellen Titeln der VDI-Schriftenreihen bequem per E-Mail

Autoren-Service:

Umfassende Betreuung bei der Veröffentlichung Ihrer Arbeit in der Reihe Fortschritt-Berichte VDI

Buch- und Medien-Service:

Beschaffung aller am Markt verfügbaren Zeitschriften, Zeitungen, Fortsetzungsreihen, Handbücher, Technische Regelwerke, elektronische Medien und vieles mehr – einzeln oder im Abo und mit weltweitem Lieferservice

VDI nachrichten | **BUCHSHOP** **www.vdi-nachrichten.com/Der-Shop-im-Ueberblick**

Die Reihen der Fortschritt-Berichte VDI:

1 Konstruktionstechnik/Maschinenelemente
2 Fertigungstechnik
3 Verfahrenstechnik
4 Bauingenieurwesen
5 Grund- und Werkstoffe/Kunststoffe
6 Energietechnik
7 Strömungstechnik
8 Mess-, Steuerungs- und Regelungstechnik
9 Elektronik/Mikro- und Nanotechnik
10 Informatik/Kommunikation
11 Schwingungstechnik
12 Verkehrstechnik/Fahrzeugtechnik
13 Fördertechnik/Logistik
14 Landtechnik/Lebensmitteltechnik
15 Umwelttechnik
16 Technik und Wirtschaft
17 Biotechnik/Medizintechnik
18 Mechanik/Bruchmechanik
19 Wärmetechnik/Kältetechnik
20 Rechnerunterstützte Verfahren (CAD, CAM, CAE CAQ, CIM ...)
21 Elektrotechnik
22 Mensch-Maschine-Systeme
23 Technische Gebäudeausrüstung

ISBN 978-3-18-3**85710**-4