

Research Cooperations in Focus

Thematic Fundamental Research Collaborations
Grounding **In**novation in
Informatics and **In**focommunications (**3IN**)



SZÉCHENYI  2020



HUNGARIAN
GOVERNMENT

European Union
European Social
Fund



INVESTING IN YOUR FUTURE

Research Cooperations in Focus

Thematic Fundamental Research Collaborations Grounding Innovation in
Informatics and Infocommunications (3IN)
2017-2020

EFOP-3.6.2-16-2017-00013



Project Manager: Imola Koszta
Proofreader: Gábor Gercsák
Design and Layout: Krisztina Irás

Photo credits: www.pexels.com (cover), www.freepik.com (DNA-helix, page 17)

ELTE Eötvös Loránd University, Faculty of Informatics
Budapest, 2020

Contents

Introduction	4
1. Programming Languages and Language Processing Tools	5
2. Software Technology for IoT Devices and Edge Computing	7
3. Efficient and Lightweight Cryptographic Mechanisms	9
4. Deeply Programmable Networks	11
5. Novel Scheduling Algorithms and Queueing Systems	13
6. Expert Systems for Health-Care	15
7. Cyber-Physical Systems with Human in the Loop	17
8. Multi-Modal Machine Learning	19
9. Communicating Digital Evolutionary Machines	21

Introduction

In 2017, the Faculty of Informatics at Eötvös Loránd University (ELTE) formed a consortium with the Faculty of Electrical Engineering and Informatics at the Budapest University of Technology and Economics (BME) and the Faculty of Information Technology and Bionics at Pázmány Péter Catholic University (PPKE) to implement the *3IN project*, supported by the European Union and co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00013, *Thematic Fundamental Research Collaborations Grounding Innovation in Informatics and Infocommunications*). The project has lasted for 39 months and was carried out in Martonvásár (ELTE), Balatonfüred (BME) and Esztergom (PPKE), which are towns in the convergence regions of Hungary. The aims of the project were

- to create a network of researchers in the field of ICT and computer science which could work effectively at European level,
- and to pursue fundamental research in order to develop capacities for successful participation in future research and innovation activities.

Our budget was 957,498,056 HUF (approximately 2.6 million EUR), of which Eötvös Loránd University has used 370,773,056 HUF. The funding at Eötvös Loránd University has allowed 23 young researchers and PhD students to start their academic career, and 20 senior researchers to strengthen their research groups. This booklet presents some of their research results achieved in the course of the 3IN project.



Kozsik Tamás

Dr. Tamás Kozsik
ELTE, project thematic lead

1. Programming Languages and Language Processing Tools

Programming language research is research carried out by computer scientists for the benefit of other computer scientists, and the development of tools by software engineers for the benefit of other software engineers. In the 3IN project we have investigated techniques to analyze and systematically transform the text of program code written in various programming languages. Furthermore, we studied methodologies to teach programming languages, and ways to define special-purpose programming languages. Our goal is to make accessible technologies and tools which can be used to reduce software development costs and to improve the quality of the produced software.

One of our research focuses has been the enhancement of the RefactorErl tool. It is a static source code analysis and refactoring tool, which has been developed at Eötvös Loránd University since 2006, supported by and in collaboration with Ericsson Hungary. RefactorErl allows software developers analyze source code written in the Erlang programming language, and collect information regarding the dependencies among source code entities. For example, the tool can help us find out where a given function is called, what values a variable can hold, or which processes communicate with each other. This functionality supports code comprehension, thereby facilitating teamwork, the introduction of the code to new developers, and debugging.

We have researched static analysis techniques with respect to other programming languages, namely Scala and C++ as well. C++ is said to be like a sharp knife: you can cut many things with it, including your fingers. Less experienced programmers may easily write C++ code which contains serious bugs and runtime errors. Even the best C++ compilers are unable to detect and prevent many of these problems. Static analysis tools may identify a wider range of potential errors. The Clang compiler infrastructure contains a tool, Static Analyzer, which is based on a particularly successful analysis approach, symbolic execution. We have proposed methods to make this tool even more effective. We have defined a domain-specific language, in which one can specify how to access “resources” in a program correctly – the correctness of resource accesses can be verified by the Clang Static Analyzer based on the code generated from the specifications. Execution time as well as memory and secondary storage consumption are key issues when static analysis is applied on real-world code. We have also evaluated the effect of certain parameters to the performance of Clang Static Analyzer. Finally, we have designed a technique to mitigate a well-known problem: information

loss between the analysis of different compilation units. Our approach supports the creation of “subprogram summaries” (descriptions omitting implementation details), which allow more precise modular analysis with affordable execution costs.

Refactoring tools enable programmers to safely change the structure of their code in order to make it look better, to turn it easier to maintain, or to prepare it for extension with new functionality. RefactorErl offers a range of refactoring transformations. It can even be used to turn a sequential computation into a parallel one (the implementation of this feature has commenced in the EU-FP7 project “ParaPhrase”). An important property of tool-supported, automated refactoring transformations is that they preserve the semantics of the code, i.e. they do not change what the code does. We have started to investigate how



1. PROGRAMMING LANGUAGES AND LANGUAGE PROCESSING TOOLS

a domain-specific language can advance the description of reliable refactoring transformations, and how a framework for producing machine-verifiable correctness proofs of refactoring transformations can be constructed. The above ideas have been adapted to other programming languages, such as Java and Scala. We have explored how the proving process becomes simpler if refactoring transformations are instantiated from provenly correct schemas.

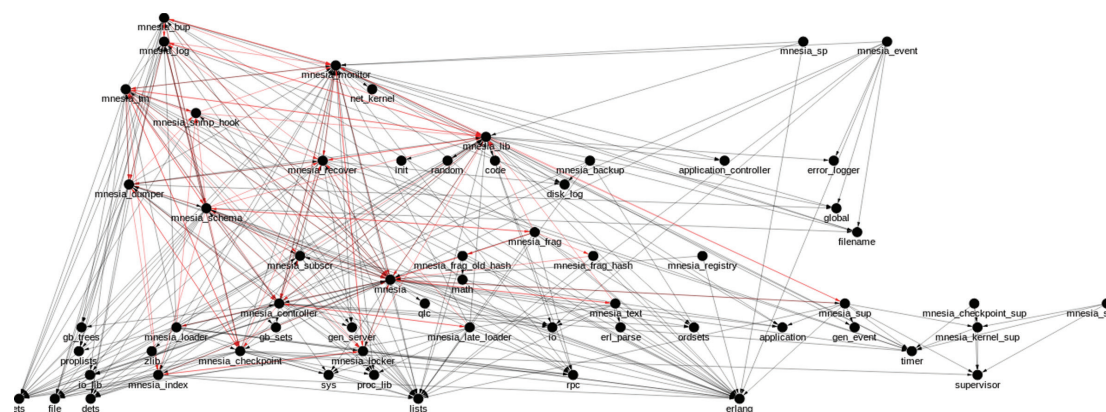
When proving properties of programming languages such as correctness of compiler optimizations or refactoring transformations, it is essential that the programming language has a formal specification. Due to the intricate details of language specifications, the only reliable way to check the validity of such arguments is to formalize them in a computer-based proof assistant such as Coq, Isabelle/HOL or Agda. An aspect of our research was developing techniques for more convenient and efficient formalization of programming languages and arguments about programming languages. Most of the computer proof assistants are based on a formal system called Martin-Löf type theory (MLTT). The syntax of a programming language is described in MLTT using inductive types: these allow the convenient definition of functions by pattern matching (e.g. the semantics of a language) and proofs about the language by structural induction. The definition of most programming languages involves mutual inductive types, that is, multiple inductive types referring to each other that are defined at the same time. We defined the first direct formal syntax for such mutual inductive types and proved that they can be translated to a single inductive type. When formalizing the equational theory of a language, one has to work with syntactic entities quotiented by the equations. It can be very cumbersome to work with explicit quotients as equality proofs have to be constructed for every case. We developed a technique based on shallow embedding which allows formalizing syntactic arguments without explicit equality proofs. Our technique is applicable whenever the equational theory of our target language is close to that of MLTT. The technique can be applied to MLTT itself, resulting in a simple self-description of MLTT. We applied this technique for formalizing important meta theoretic arguments about MLTT in a very abstract and concise way. Properties covered by our formalization include canonicity, the fact that every boolean term is equal to either true or false. Another such property is parametricity, which implies that polymorphic functions behave in a uniform way, regardless of the concrete types they are applied at.

Thematic leader: Dr. Tamás KOZSIK (Tamas.Kozsik@elte.hu)

Selected publications:

- Eötvös Loránd University, Dept. Programming Languages and Compilers: The RefactorErl tool.
<https://plc.inf.elte.hu/erlang/>
- M. TÓTH, I. BOZÓ, T. KOZSIK: *Pattern Candidate Discovery and Parallelization Techniques*. In IFL 2017: Proceedings of the 29th Symposium on the Implementation and Application of Functional Programming Languages. ACM Press, (2017) pp. 1–26.
- A. KAPOSI, A. KOVÁCS, N. KRAUS: *Shallow Embedding of Type Theory is Morally Correct*. In Mathematics of Program Construction, LNCS 11825, Springer International Publishing, (2019) pp. 329–365. DOI: 10.1007/978-3-030-33636-3_12
- E. FÜLÖP, N. PATAKI: *Comprehensive Evaluation of Cross Translation Unit Symbolic Execution*. In Proc. 11th International Conference on Applied Informatics (ICAI 2020). CEUR Workshop Proceedings, Vol. 2650, (2020) pp. 105-113.

6



When proving properties of programming languages such as correctness of compiler optimizations or refactoring transformations, it is essential that the programming language has a formal specification. Due to the intricate details of language specifications, the only reliable way to check the validity of such arguments is to formalize them in a computer-based proof assistant such as Coq, Isabelle/HOL or Agda. An aspect of our research was developing techniques for more convenient and efficient formalization of programming languages and arguments about programming languages. Most of the computer proof assistants are based on a formal system called Martin-Löf type theory (MLTT). The syntax of a programming language is described in MLTT using inductive types: these allow the convenient definition of functions by pattern matching (e.g. the semantics of a language) and proofs about the language by structural induction. The definition of most programming languages involves mutual inductive types, that is, multiple inductive types referring to each other that are defined at the same time. We defined the first direct formal syntax for such mutual inductive types and proved that they can be translated to a single inductive type. When formalizing the equational theory of a language, one has to work with syntactic entities quotiented by the equations. It can be very cumbersome to work with explicit quotients as equality proofs have to be constructed for every case. We developed a technique based on shallow embedding which allows formalizing syntactic arguments without explicit equality proofs. Our technique is applicable whenever the equational theory of our target language is close to that of MLTT. The technique can be applied to MLTT itself, resulting in a simple self-description of MLTT. We applied this technique for formalizing important meta theoretic arguments about MLTT in a very abstract and concise way. Properties covered by our formalization include canonicity, the fact that every boolean term is equal to either true or false. Another such property is parametricity, which implies that polymorphic functions behave in a uniform way, regardless of the concrete types they are applied at.

Selected publications:

M. TÓTH, I. BOZÓ, T. KOZSIK: *Pattern Candidate Discovery and Parallelization Techniques*. In IFL 2017: Proceedings of the 29th Symposium on the Implementation and Application of Functional Programming Languages. ACM Press, (2017) pp. 1–26.

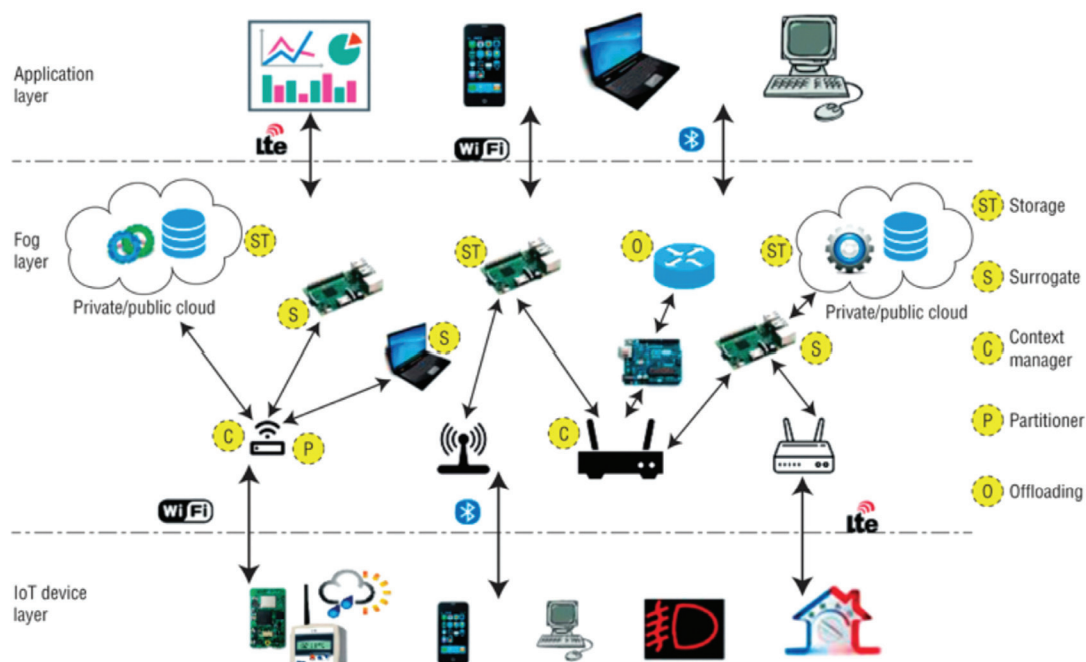
E. FÜLÖP, N. PATAKI: *Comprehensive Evaluation of Cross Translation Unit Symbolic Execution*. In Proc. 11th International Conference on Applied Informatics (ICAI 2020). CEUR Workshop Proceedings, Vol. 2650, (2020) pp. 105-113.

2. Software Technology for IoT Devices and Edge Computing

The key for the future Internet of Things (IoT) is the shift from a static architecture to a dynamically evolving, self-organizing one. Devices of extremely varying capabilities need to collaborate and access all information necessary to ensure their optimal work and keeping flexibility in the network configuration. The supporting infrastructure should be able to optimize resource consumption and to allow the connected device to interact with the most convenient node in the network without compromising the perceived Quality of the Service.

The work as a whole is an effort to summarize the process to build an offloading framework for arbitrary tasks of IoT devices. Another main contribution is the collection of requirements for these new IoT networks and computing systems, in particular the need for edge computing, offloading paradigms and an underlying infrastructure that could be supported using 5th Generation mobile networks standards, Clustering, and the use of Artificial Intelligence. We propose and analyse a global view of the problem and identify some specific possible technical solutions.

Offloading resource-intensive tasks from connected mobile and IoT devices and executing them not only in the cloud but also in the fog is a promising direction (edge computing). In the field of Fog computing, because of the highly dynamic nature of Mobile Networks, the problem of automating task distribution and code migration has not yet been solved by any resource orchestrator. We propose a solution to the problem in the form of an orchestrator that can build overlay computational networks from a volatile set of nodes. Each node of this overlay can deploy specific services and is able to satisfy the requested task from any participant node. We also demonstrate the infrastructure purpose by suggesting an existing Federated Learning Application.



2. SOFTWARE TECHNOLOGY FOR IOT DEVICES AND EDGE COMPUTING

The problem of service availability assurance and optimal performance in 5G edge computing is solved by employing static or rudimentary dynamic scheduling techniques. In the cases where Kubernetes orchestration is used, the default scheduler manages the allocation of services with methods designed to suit general cloud application requirements, not specific to the requirements of edge computing (i.e. evaluating basic usage devices parameters). This design has its shortcomings, especially due to its slow reactions to environmental changes. In such circumstances, it would be preferred to take the physical, operational, and network parameters into consideration along with the software states and orchestrate the applications dynamically.

We propose an improved design that integrates real-time information about the edge devices in the scheduler decision algorithm enabling advanced orchestration and management along with improved fault tolerance capabilities of the network edge. We validate the utility of the design by comparing it to the default Kubernetes scheduler.

Distributed Computing in 5G Mobile Networks is a potential requirement for certain applications that depend on low latency and information sharing through or with data information sources. Such applications may be observed as a distributed application.

We present a tool and method to optimize the deployment of distributed applications, dividing it into Modules, in a 5G Mobile Network environment. To do so we apply an approximation algorithm for the Path Computation and Function Placement Problem. We show that under certain circumstances it is beneficial to deploy parts of such applications in a Cloud Computing environment with Distributed Cloud resources at the Mobile Network Edge. We verify our findings with an example, an Augmented Reality application.

Thematic leader:

Dr. Zoltán ISTENES (istenes@inf.elte.hu)

Dr. Zoltán HORVÁTH (hz@inf.elte.hu)

Selected publications:

P. KISS, A. REALE, Ch. J. FERRARI, Z. ISTENES: *Deployment of IoT applications on 5G Edge*. In 2018 IEEE International Conference on Future IoT Technologies (Future IoT), 18-19 Jan. 2018, Eger, Hungary, DOI: 10.1109/FIOT.2018.8325595

A. REALE, P. KISS, Ch. J. FERRARI, B. KOVÁCS, L. SZILÁGYI, M. TÓTH: *Application Functions Placement Optimization in a Mobile Distributed Cloud Environment*. In Studia Universitatis Babeş-Bolyai Informatica 63(2):37-52, June 2018, DOI: 10.24193/subbi.2018.2.03

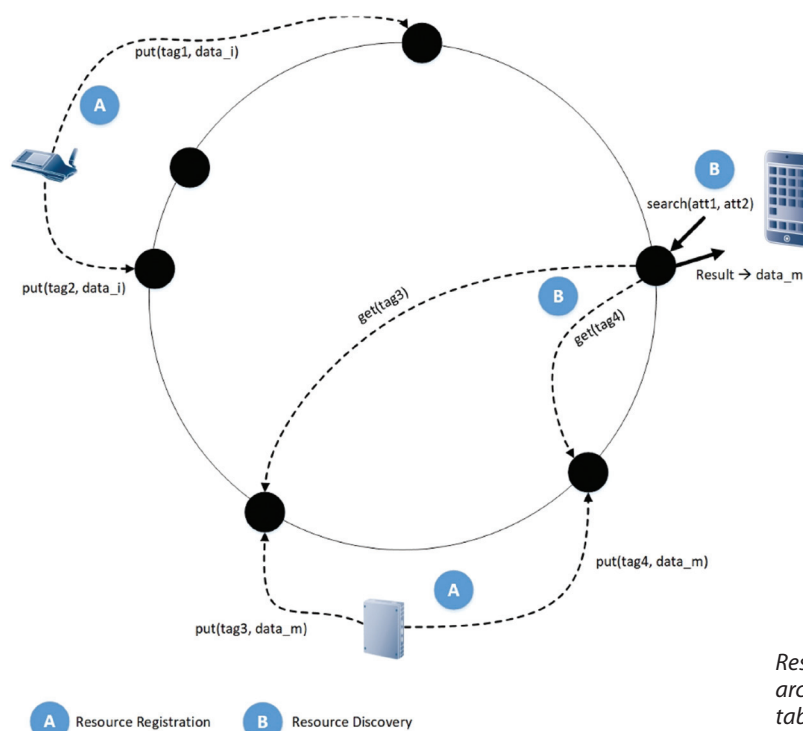
A. REALE, P. KISS, M. TÓTH, Z. HORVÁTH: *Designing a Decentralized Container Based Fog Computing Framework for Task Distribution and Management*. In International Journal of Computers and Communications, Volume 13, 2019

3. Efficient and Lightweight Cryptographic Mechanisms

Though a significant part of the accumulators of medical sensors is very complicated to charge (like in-body sensors), the stored/transmitted data are sensitive, claiming that the protection of the communication is a must. Traditional cryptography algorithms have significant computational and power costs, hence cost-effective mechanisms are preferable during the development. We are focusing on two main related topics: lightweight cryptographic protocols and efficient differential privacy and homomorphic encryptions for special purposes.

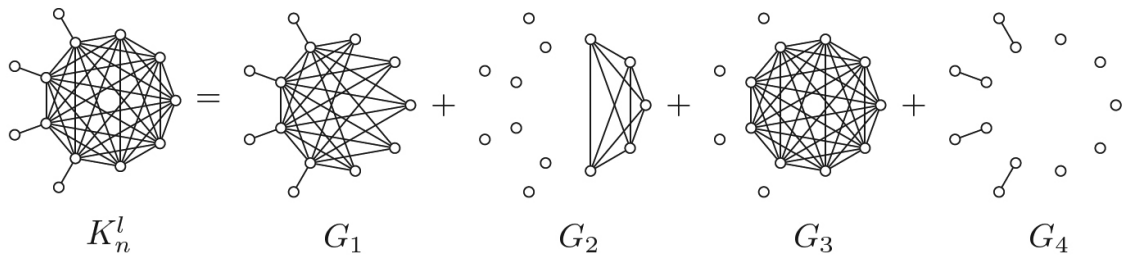
The main research goal of the first topic is development and implementation of cryptographic methods applicable in restricted environments. Various systems with reduced resources can be considered: computational or communication complexity or the necessary storage capacity can be a bottleneck. The general solutions for differential privacy and homomorphic encryption have significant computational overhead. However, for particular applications there exist special implementations achieving security and efficiency together. Using such cryptographic methods some algorithmic steps – like linear data-processing operations – can be computed securely in the cloud.

One possible idea is to delegate a part of the costly computation or stored data to other users of the system. To protect the sensitive data, the goal was to develop secure distributed protocols for various scenarios. Another interesting research direction is the application of secret



Resource registration and discovery architecture for distributed hash table based search in Internet of Things networks

3. EFFICIENT AND LIGHTWEIGHT CRYPTOGRAPHIC MECHANISMS



Decomposition construction for a graph-based secret sharing scheme.

sharing methods, where the sensitive data is shared between the participants in a way that only some pre-defined qualified subset can recover it. The goal was to develop novel efficient systems for realistic scenarios. Last but not least, the analysis, development and implementation of efficient homomorphic and differential privacy algorithms were in focus. Here, the goal was to analyze methods based on simple operations.

We proposed novel optimal graph-based secret sharing schemes for a large family of graphs by sophisticated weighted constructions and information theoretic arguments. Additionally, we suggested several hierarchical secret sharing schemes based on finite geometry constructions. We designed DHT-based distributed communication protocols, especially address and resource discovery methods. Some of the solutions can be applied in restricted IoT environment as well.

We investigated how privacy-preserving data aggregation mechanisms can be built from alternative security assumptions, especially numeration systems, which could serve as a post-quantum alternative to factoring or DLOG-based methods. We deduced new complexity theoretic results on the difficulty of the low-order assumption. Our results indicate that for a large part of RSA-moduli, the security of protocols built on the low-order assumption attains the same level as the security of RSA itself.

Thematic leaders:

Dr. Péter BURCSI (bupe@inf.elte.hu)

Dr. Péter LIGETI (ligetipeter@inf.elte.hu)

Selected publications:

P. BURCSI, I. A. SERES: *A Note on Low Order Assumptions in RSA groups*. IACR Cryptol. ePrint Arch. 2020, 402.

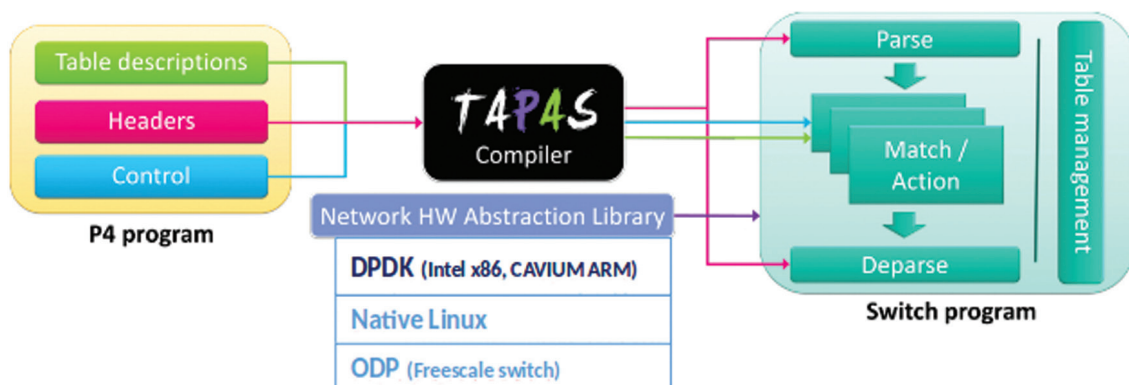
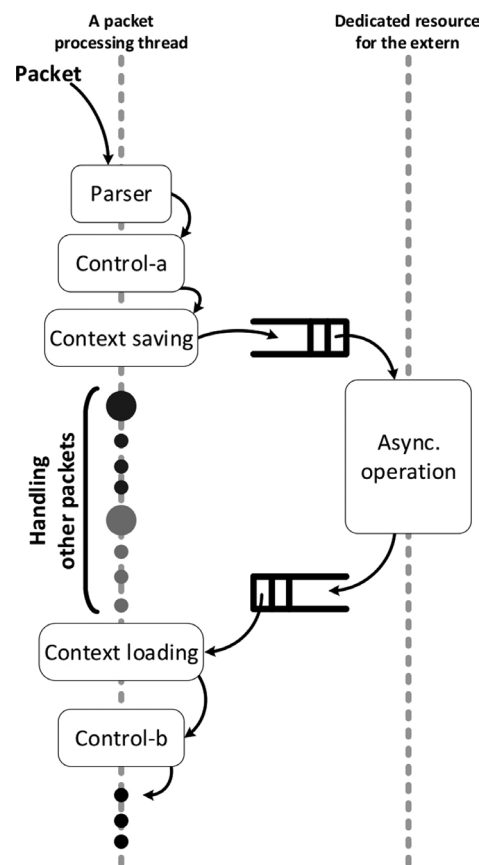
M. GYARMATI, P. LIGETI: *Smallest graphs achieving the Stinson bound*. IEEE Transactions on Information Theory, 66 (7) 2020 pp. 4609–4612 .

M. B. M. KAMEL, Á. NAGY, P. LIGETI: *An Address Propagation Model in P2P and F2F Networks*. Studia Universitatis Babeş-Bolyai Informatica, 64 (1) 2019 pp. 91–101.

4. Deeply Programmable Networks

Software-Defined Networks (SDN) separates networking processes into a control plane and a data plane, decoupling network control and forwarding functions. This concept results in a more flexible network enabling a directly programmable network control. Deeply Programmable Networking (DPN) is an extension to SDN. It enables not only a programmable control plane, but a (deeply) programmable data plane resulting in a much more flexible protocol management. This allows of handling substantially new protocols easy.

In our research, we study how the DPN concept is adaptable to real-world (industrial) problems. We have focused on the P4 language as a possible solution. P4 is a high-level, domain specific language for programming protocol-independent packet processors developed by a consortium of some noted universities (including ELTE) and numerous industrial partners. We have investigated the applicability of the language and different architecture models (used by the language) defining and implementing case studies in P4. We have also developed a P4 compiler, T4P4S, which separates the hardware-specific and hardware-independent program functionalities. The compiler is open-source and available in github (<https://github.com/P4ELTE/t4p4s>). We have also investigated some extension possibilities of the compiler. For example, we have analyzed how we can adapt the compiler to different architecture models and how we can



4. DEEPLY PROGRAMMABLE NETWORKS

use different execution models on it. We have also studied the options of adding external functionalities to P4 language and (because the default execution of external functions is synchronous) developed a solution for the application of asynchronous external functions. We have investigated the possibilities of analyzing functional and non-functional properties of P4 programs. As part of this, we have defined cost model for the different components of P4 programs, namely for the parser, the match/action table lookups and the deparser. The model uses hardware descriptions and some probabilities of the different execution paths as parameters and it is usable for predicting the execution cost of P4 programs.

Thematic leaders:

Sándor LAKI (lakis@elte.hu)

Máté TEJFEL (matej@elte.hu)

Selected publications:

P. VÖRÖS, D. HORPÁCSI, R. KITLEI, D. LESKÓ, M. TEJFEL, S. LAKI: *T4P4S: A Target-independent Compiler for Protocol-independent Packet Processors*. 2018 IEEE 19th International Conference on High Performance Switching and Routing (HPSR), Bucharest, Romania, 2018, pp. 1–8, doi: 10.1109/HPSR.2018.8850752.

D. LUKÁCS, M. TEJFEL, G. PONGRÁCZ: *Keeping P4 Switches Fast and Fault-free through Automatic Verification*. Acta Cybernetica, 24(1), 2019, 61–81. <https://doi.org/10.14232/actacyb.24.1.2019.6>

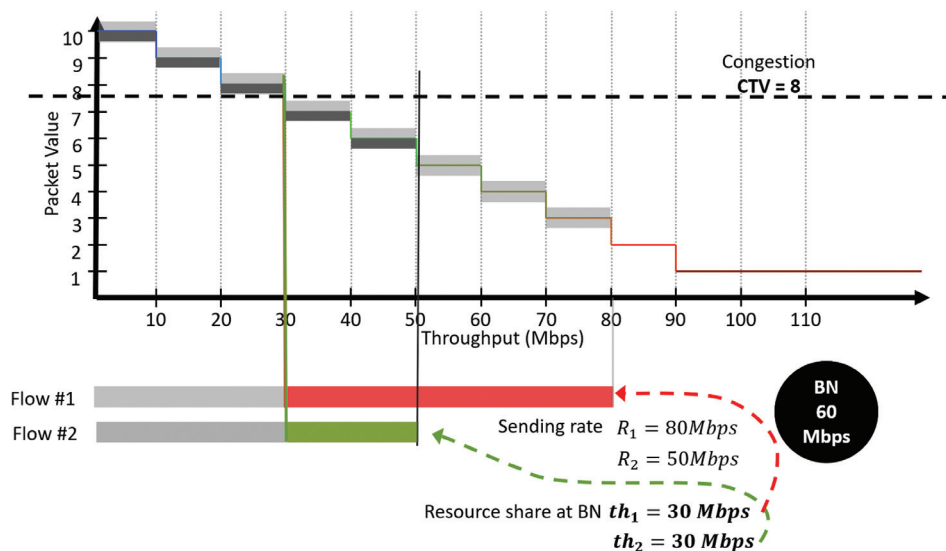
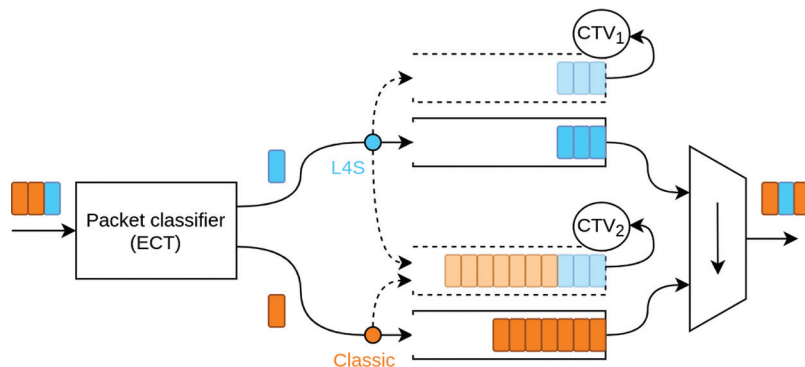
D. LUKÁCS, G. PONGRÁCZ, M. TEJFEL: *Performance guarantees for P4 through cost analysis*. 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, 2019, pp. 000305–000310, doi: 10.1109/Informatics47936.2019.9119260.

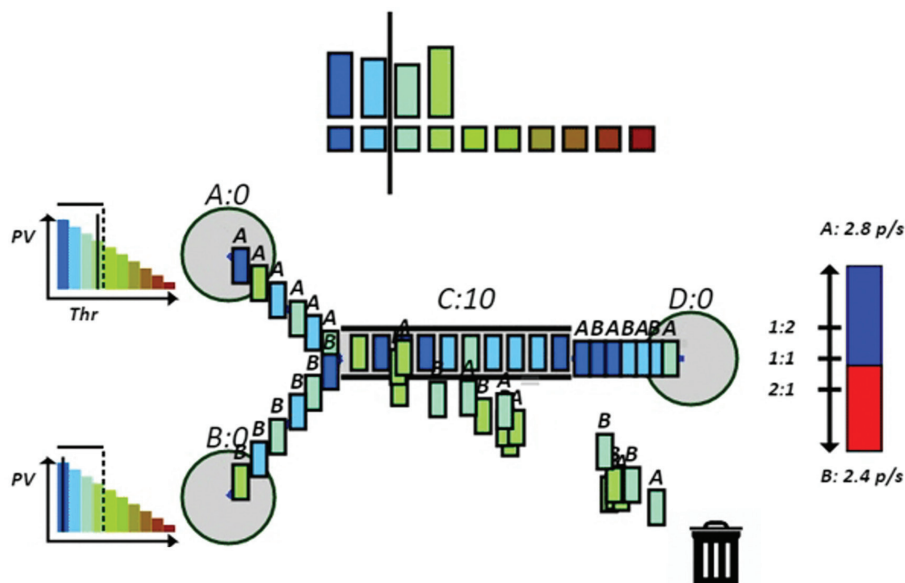
S. LAKI et al.: *The Price for Asynchronous Execution of Extern Functions in Programmable Software Data Planes*. 2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), Paris, France, 2020, pp. 23–28, doi: 10.1109/ICIN48450.2020.9059357.

5. Novel Scheduling Algorithms and Queueing Systems

With the advent of 5G, high-speed access networks become general making the appropriate overprovisioning of backhaul networks difficult and costly. This issue may be handled with robust and scalable Quality of Service solutions that have the potential to support congestion management and SLA enforcement/insurance among traffic aggregates at various levels (per-flow, per-user, per-application, etc.). Similar problems arise in multi-tenant data center and residential access network environments as well.

Providing good QoS to end users requires congestion management among traffic aggregates in overloaded situations. There is a rough consensus that any solution that keeps flow states inside the network will not scale with current demands. In addition, many novel applications cannot tolerate large delays and delay fluctuations, thus a comprehensive QoS framework also needs mechanisms for ensuring controlled buffering delay. To overcome the scalability issues





of stateful traditional QoS approaches, a number of core-stateless proposals have emerged in the past, but none of them can be viewed as a complete QoS solution focusing on either providing fair share among users or providing bounded delay.

In this project, we developed new resource sharing framework called Per Packet Value (PPV: <http://ppv.elte.hu>) that relies on and extends the capabilities of core-stateless forwarding concept. Our approaches consider both delay and throughput requirements independently. In contrast to prior methods, our approaches can provide ultra-low (sub-millisecond) latency for throughput-hungry applications without starving others. The developed algorithms are lightweight, and they can be implemented on programmable switches.

Thematic leader:

Sándor LAKI (lakis@inf.elte.hu)

Selected publications:

S. LAKI et al.: *Core-Stateless Forwarding with QoS Revisited: Decoupling Delay and Bandwidth Requirements*.

Accepted for publication in IEEE/ACM Transactions on Networking, IF: 4.8, 2020

F. FEJES, G. GOMBOS, S. LAKI, Sz. NÁDAS: *Who will Save the Internet from the Congestion Control Revolution?*

Workshop on Buffer Sizing, 2–3 December, 2019 – Stanford University, CA, USA

Sz. NÁDAS, G. GOMBOS, F. FEJES, S. LAKI: *Towards Core-Stateless Fairness on Multiple Timescales*. IRTF/ACM Applied Networking Research Workshop 2019 (ANRW 2019), 22 July, 2019, Canada

Sz. NÁDAS, Z. TURÁNYI, G. GOMBOS, S. LAKI: *Stateless Resource Sharing in Networks with Multi-Layer Virtualization*. IEEE International Conference on Communications 2019 (ICC 2019) – Next-Generation Networking and Internet Symposium, pp. 1–7, 20–24 May 2019, Shanghai, China

Sz. NÁDAS, G. GOMBOS, P. HUDOBA, S. LAKI: *Towards a Congestion Control-Independent Core-Stateless AQM*. IRTF/ACM Applied Networking Research Workshop 2018 (ANRW 2018), 16 July, 2018, Canada

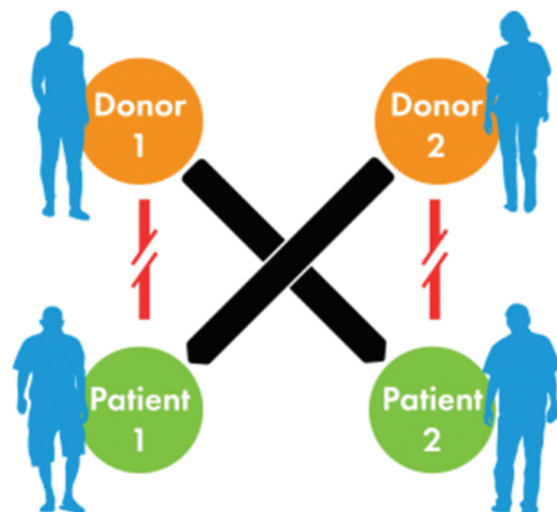
6. Expert Systems for Health-Care

While in the last centuries many research results have not become known and developed because of the lack of wide-ranging publication, huge amounts of scientific articles are now available to all researchers. However, human ability sets a limit to their processing. The project carried out research activities that automate the processing of thousands of scientific results (articles) related to specific topics using web mining, data mining, text mining technologies, and build mathematical and IT models of hidden information for specific health purposes. The project developed models for the following two specific topics.

Expert System for Selection of Ancestral Cell Donor

The goal was to develop an expert model capable of providing a much higher level of decision support in the process of umbilical cord transplantation than currently used international registers, taking into account findings in the literature and current medical expertise.

Actually, like the blood supply system, samples of community umbilical cord blood banks are transplanted in significant quantities: alien specimens are implanted in the treatment of diseases involving bone marrow transplantation. In case of bone marrow transplantation the entire genetic match is desired (e.g. 6/6 V HLA), while immature umbilical cord blood stem cells can be transplanted with a partial match (4/6) just as well. Around 4,000 foreign umbilical cord blood is transplanted worldwide each year, limited mainly by the availability of about 750,000 continuously growing frozen units. If the samples searched for in the international registries for the patient are equivalent in terms of transplantation and are equally partially consistent with the patient's genetics (HLA), further considerations are not taken into account currently in the selection of samples. The genetic characteristics (HLA) analyzed for matching may encode a predisposition or resistance in autoimmune, infectious, or indirectly tumorous diseases. If a physician had to consider whether to implant a patient's predisposed or resistant specimen, they would have to summarize the contents



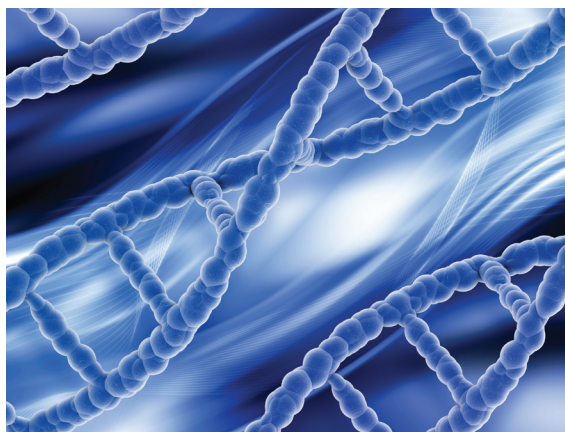
Source: transplantliving.org

6. EXPERT SYSTEMS FOR HEALTH-CARE

of 2,500 scientific papers. However, analyzing the literature using automated data mining methods allows the clinician to make quick decisions and lead to an immediate conclusion rather than a random transplant.

Analysis of HLA-encoded Genetic Properties and Bacterial Flora Relationship

The normal flora is the combination of bacteria that are on the body surface and in the body cavities permanently or temporarily, and do not cause disease in a healthy person. These are apathogenic microbes, which are not harmful to the human body. The importance of normal flora is increasingly recognized by medicine for the protection and disease of the body. More than 5,000 species of bacteria can be found in the human flora, some thousands of which can be directly cultivated, only the traces of the others. A unique feature for each person is that which bacterial species are found in their gut flora, for example. The genetic qualities carried by the HLA codes describe this uniqueness best. A very important goal in this topic was to provide comprehensive information about what genetic codes are present in the organism. The next step was to find out how the bacterial effect of each bacterium is related to genetic characteristics. Related research results have been published in thousands of scientific articles, however, the comprehensive information describing the overall HLA bacteria relationship has not yet been elaborated. Each publication contains results for some specific bacteria. This is a very active research area, as more and more illnesses can be traced back to some form of bacterial flora disorder and individual sensitivity. Our expert system supports the deeper examination of the relationship between bacteria and genetic properties.



Thematic leader:

Dr. Gyöngyi KOCSISNÉ SZILÁGYI (szilagyi@aszt.inf.elte.hu)

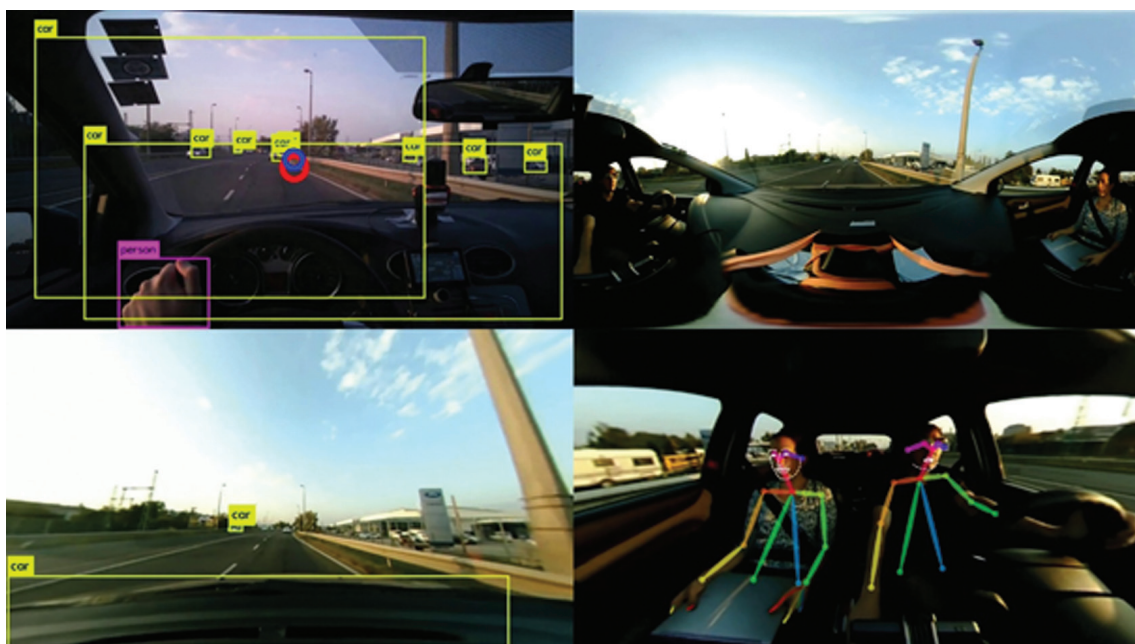
7 ● Cyber-Physical Systems with Human in the Loop

Cyber-Physical Systems (CPS) are complex applications in autonomous driving, medical monitoring, industrial control systems and in many other fields. The focus of R+D+I is on the machines and performance of complex structures. The role of human participants has become relevant in unexpected, possibly critical situations. As these are considered as stressful and dangerous cases, the precise analysis and prediction of human actions and behavior is crucial to avoid disasters.

One of our investigated topics was event detection and human action recognition inside a vehicle (driver is drinking/eating/talking/having a seizure, etc.) and outside, and in the traffic environment (cyclist in the pedestrian crossing, unexpected appearance of pedestrians, etc.).

To accomplish our goal, we investigated the fusion opportunities of (a) the outcomes of deep neural networks and (b) ontologies, which are prerequisites for human understanding. Combining the learning algorithms' knowledge and humans' common sense could lead to more effective situation analysis in real-life scenarios.

We have designed and implemented simplified systems focusing on car drivers' conditions in traffic scenarios. Moreover, we have developed a game for measuring users' divided attention skills and organized a series of experiments with specifically crafted experimental settings to measure the effect of increasing cognitive load and certain unusual input methods. Another outcome was a self-collected and post-annotated dataset of real-life traffic scenarios for measuring the driver's and the passenger's mental and physical state. An architecture was

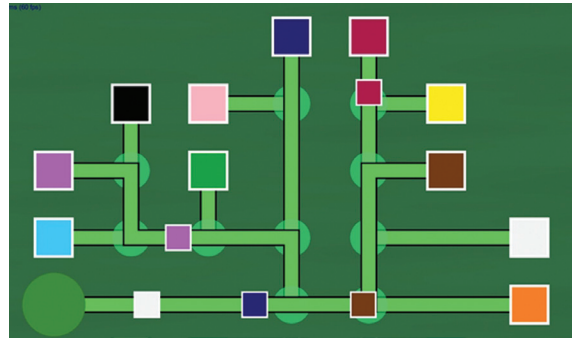


Overlays of state-of-the-art pose, face and object detections on the "CPS driving" self-collected dataset from a head-mounted smart glasses' and a dashboard camera's view

7. CYBER-PHYSICAL SYSTEMS WITH HUMAN IN THE LOOP

demonstrated in this “CPS driving” database through integrating traditional AI and deep learning methods with deterministic models and reasoning tools. Regarding the variability of traffic scenarios’ environment, we analyzed the predictability of unknown objects’ collision based on motion information from 2D videos.

Experimental methods applied: Tobii eye tracking device, novel tools for object recognition, human face and body-pose detection were used for understanding common events in traffic scenarios and inside the car environment. In our divided attention game, carefully designed scenarios and experimental protocols were used in human–computer interaction experiments, overall logging, mouse click recording, face recording and gaze tracking for the evaluation of the user performance and learning capabilities.



Divided attention game: The small squares need to be directed to their color-corresponding destinations

Thematic leaders:

András LŐRINCZ (lorincz@inf.elte.hu)

Selected publications:

Z. MILACSKI, K. B. FARAGÓ, A. FÓTHI, V. VARGA, A. LŐRINCZ: *Declarative description: The meeting point of artificial intelligence deep neural networks and human intelligence*. In: XAI 2018, Proceedings of the 2nd Workshop on Explainable Artificial Intelligence (pp. 97–103), 2018.

R. A. RILL, K. B. FARAGÓ, A. LŐRINCZ: *Strategic predictors of performance in a divided attention task*. PloS one, 13(4), e0195131, 2018.

R. A. RILL, A. LŐRINCZ: *Cognitive Modeling Approach for Dealing with Challenges in Cyber-Physical Systems*. Universitas Babeş–Bolyai Informatica, 51, 2019.

R. A. RILL, K. B. Faragó: *Gaze-based Cursor Control Impairs Performance in Divided Attention*. Acta Cybernetica, 23, pp. 1071–1087. 2018.

L. KOPÁCSI, Á. FÓTHI, Á. FODOR, E. SOMFAI, A. LŐRINCZ: *Common Fate Based Episodic Segmentation by Combining Supervoxels with Deep Neural Networks*. In: 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1–7). IEEE. 2019.

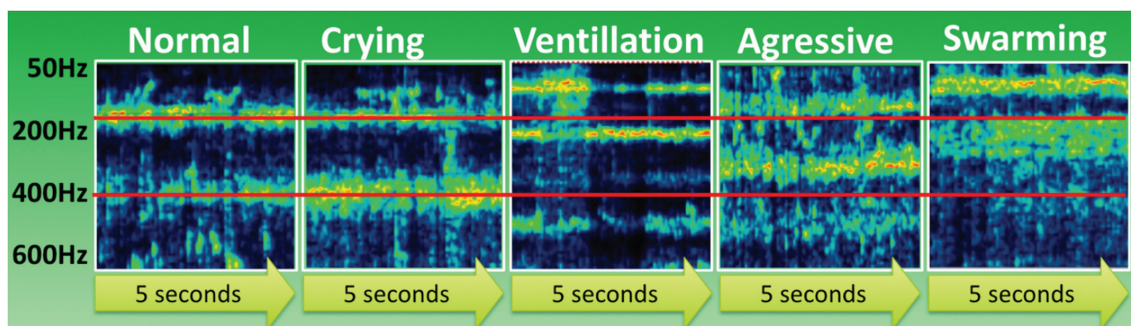
8. Multi-Modal Machine Learning

Machine learning (ML) on multi-modal data aims at deriving useful prediction models utilizing various types of data, supplemental to standard numerical measurements, from which text, image and time series are the most considerable. Just imagine humans making better decisions possessing not only numerical but also audio, visual and/or textual information. Even if the intuition behind multi-modal ML is reasonable, joint learning from image and text data is still a challenging task. Existing ML solutions developed for multi-modal data rely on ML models for unimodal data and, thus, there is a huge space for research and development.

For such, unimodal ML models were developed within the project, which will form the basis for multi-modal ML approaches developed in our ongoing research. We have been choosing various domains for our research, such that agriculture, finance and education. The main purpose was to develop models generic enough so that these could be deployed in various domains.

Regarding the education domain, we were focusing on automated essay evaluation (AEE), which plays an important part of intelligent tutoring systems (ITS). While there are many approaches to evaluate various types of tests, there is not much research devoted to evaluating textual essays of students. Evaluation of essays is a tedious task for a teacher since there is not always one good solution but, rather, many subjective criteria such as writing style or creativity, just to name a few. The developed ML models for AEE were integrated to an ITS, developed by us, and presented at various scientific conferences with great success: One of the developed models received the best student paper award at the International Conference on Computer Supported Education (CSEDU), while the developed ITS was nominated for the best short paper award at the International Conference on Intelligent Tutoring Systems (ITS).

Within the agricultural domain, we were working on image and audio data processing with applications in viticulture and apiculture, to be more exact, recognition of vine types based on the images of leaves and detecting anomalies in beehive sounds. Here, we com-



8. MULTI-MODAL MACHINE LEARNING

combined image processing and time-series mining techniques by representing a vine leaf as time-series and utilizing ML models for time-series classification. For better processing sounds, we were investigating various time-series pre-processing techniques and their utility for time-series classification.

Regarding the financial domains, the results of image processing techniques, namely optical character recognition, were combined with text and network processing techniques in order to recognize information from invoices. During the development, we have been proposing various, novel features for the task at hand which helped to improve the annotation accuracy.



Thematic leader:

Tomáš HORVÁTH (horvathtamas@inf.elte.hu)

Selected publications:

T. M. TASHU, T. HORVÁTH: *Semantic-Based Feedback Recommendation for Automatic Essay Evaluation*. In: Intelligent Systems and Applications. Advances in Intelligent Systems and Computing, Vol. 1038, Springer (2020), pp. 334-346.

T. M. TASHU, T. HORVÁTH: *A Layered Approach to Automatic Essay Evaluation Using Word-Embedding*. In: Communications in Computer and Information Science, Vol. 1022, Springer, Cham (2020), pp. 77-94.

A. BALAWI, A. AI ZOABI, J. L. SEIXAS JUNIOR, T. HORVÁTH: *Classification of a Small Imbalanced Dataset of Vine Leaves Images using Deep Learning Techniques*. ITAT 2020 Information Technologies - Applications and Theory, Hotel Tyrapol, Oravská Lesná, Slovakia, 2020: CEUR Workshop Proceedings Vol-2718, ISSN 1613-0073, 2019, pp.: 108-114.

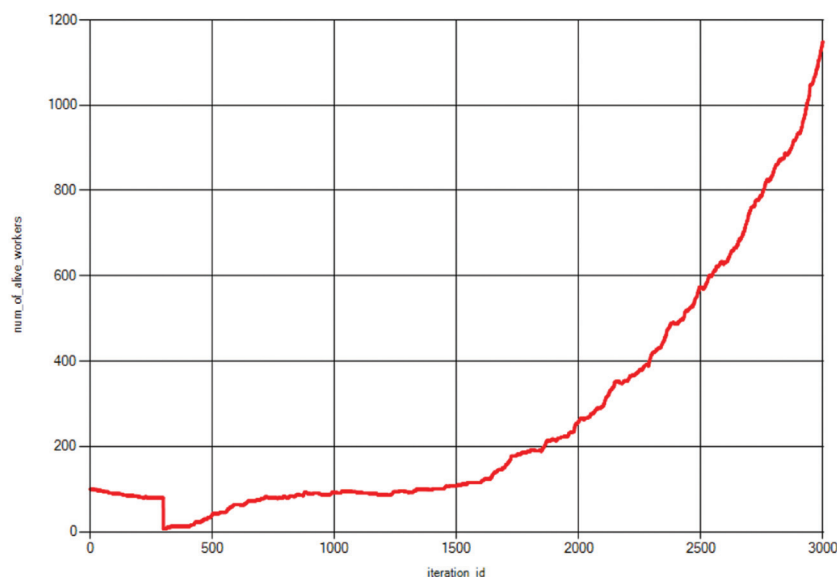
J. L. SEIXAS JUNIOR, T. HORVÁTH: *KNN Algorithm with DTW Distance for Signature Classification of Wine Leaves*. ITAT 2020 Information Technologies - Applications and Theory, Hotel Tyrapol, Oravská Lesná, Slovakia, 2020: CEUR Workshop Proceedings Vol-2718, ISSN 1613-0073, 2019, pp: 130-136.

9. Communicating Digital Evolutionary Machines

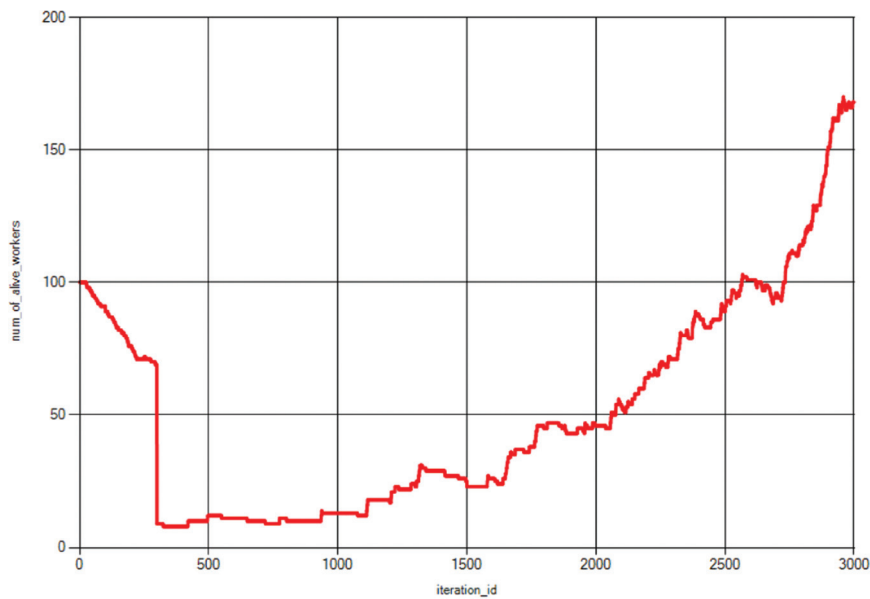
The spontaneous emergence of intelligence is a long-standing problem in artificial intelligence research. So far, it has only occurred in living environments, of which the evolution of life is the only example. The keynote speakers in plenary sessions of AI conferences claim that this issue is currently unresolved. We have created an environment in which we can investigate the spontaneous development of intelligent behaviors in an artificial environment. Our aim was to continue a previously started research, from which two books have been written.

It has also been concluded from previous research that initially completely unexperienced artificial entities (beings) are able to apply their experiences for their own survival. This working mechanism can be considered a very simple but spontaneously developed intelligent mode of operation.

So far, research has focused on non-communicating beings. Recently, we have begun to examine what happens if the creatures communicate with each other by exchanging knowledge when they meet, i.e. they are in one place at the same time. To achieve this, we introduced the concept of time and the ability to manage coincidences of multitudes if we launch



*The knowledge exchange has increased the chances of the population surviving.
The initial population of 100 individuals, after 3,000 iteration steps, changed to nearly 1,200.*



*Development of the population without knowledge exchange. After 3,000 iteration steps, 170 individuals are alive.
The starting number of individuals was 100.*

many hundreds or many thousands of beings at once. We also transformed the living space of the creatures from a chessboard world into a torus. This, world is finite, but there are no borders in it, hence creatures never collide with boundaries, and they can even perceive their world as infinite.

The results of the research are extremely convincing. The communication of beings, the exchange of knowledge, has dramatically increased their chances of survival even in cases where, without communication, there was degradation or extinction in the population.

Thematic leader:

István ELEK (elek@map.elte.hu)

Selected publication:

I. ELEK: *The Spontaneous Appearance of Intelligence, The Digital Evolution Machine*. Eötvös Publishing House, Budapest 2015.

I. ELEK: *Emergence of Intelligence*. NOVA Science Publishers, New York, 2018.

I. ELEK, T. HEGER, D. SINDELY, D. LINGER, Z. BLAZSIK: *Communicating digital evolutionary machines*. Computing Conference 2021, 15-16 July, London, UK. (accepted)

