Home > Doctoral students > PhD thesis defenses > PhD thesis defenses 2014

2014

December

Benjamin ROTH
Effective Distant Supervision for End-to-End Knowledge Base Population Systems
(Advisor: Prof. Dietrich Klakow)
Mon, 22 Dec 2014, 11:00, building C7 4, room 117

The growing amounts of textual data require automatic methods for structuring relevant information so that it can be further processed by computers and systematically accessed by humans. The scenario dealt with in this dissertation is known as Knowledge Base Population (KBP), where relational information about entities is retrieved from a large text collection and stored in a database, structured according to a pre-specified schema. Most of the research in this dissertation is placed in the context of the KBP benchmark of the Text Analysis Conference (TAC KBP), which provides a test-bed to examine all steps in a complex end-to-end relation extraction setting. In this dissertation a new state of the art for the TAC KBP benchmark was achieved by focussing on the following research problems:
(1) The KBP task was broken down into a modular pipeline of sub-problems, and the most pressing issues were identified and quantified at all steps.
(2) The quality of semi-automatically generated training data was increased by developing noise-reduction methods, decreasing the influence of false-positive training examples.
(3) A focus was laid on fine-grained entity type modelling, entity expansion, entity matching and tagging, to maintain as much recall as possible on the relational argument level.
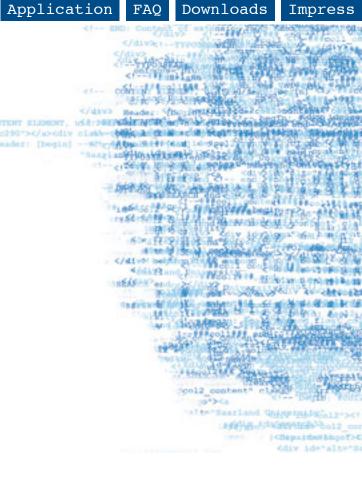(4) A new set of effective methods for generating training data, encoding features and training relational classifiers was developed and compared with previous state-of-the-art methods.

Foteini ALVANAKI
Mining interesting events on large and dynamic data
(Advisor: Prof. Sebastian Michel)
Mon, 22 Dec 2014, 9:30, building E1 4, room 0.24

Nowadays, almost every human interaction produces some form of data. These data are available either to every user, e.g. images uploaded on Flickr or to users with specific privileges, e.g. transactions in a bank. The huge amount of these produced data can easily overwhelm humans that try to make sense out of it. The need for methods that will analyse the content of the produced data, identify emerging topics in it and present the topics to the users has emerged. In this work, we focus on emerging topics identification over large and dynamic data. More specifically, we analyse two types of data: data published in social networks like Twitter, Flickr etc. and structured data stored in relationaldatabases that are updated through continuous insertion queries.
In social networks, users post text, images or videos and annotate each of them with a set of tags describing its content. We define sets of co-occurring tags to represent topics and track the correlations of co-occurring tags over time. We split the tags to multiple nodes and make each node responsible of computing the correlations of its assigned tags. We implemented our approach in Storm, a distributed processing engine, and conducted a user study to estimate the quality of our results.
In structured data stored in relational databases, top-k group-by queries are defined and an emerging topic is considered to be a change in the top-k results. We maintain the top-k result sets in the presence of updates minimizing the interaction with the underlying database. We implemented and experimentally tested our approach.

Karl BRINGMANN
Sampling from Discrete Distributions and Computing Fréchet Distances
(Advisor: Prof. Kurt Mehlhorn)
Wed, 17 Dec 2014, 12:30, building E1 4, room 0.24

In the first part of this thesis, we study the fundamental problem of sampling from a discrete probability distribution. Specifically, given non-negative numbers $p_1,...,p_n$ the task is to draw i with probability proportional to p_i. We extend the classic solution to this problem, Walker's alias method, in various directions: We improve its space requirements, we solve the special case of sorted input, we study sampling natural distributions on a bounded precision machine, and as an application we speed up sampling a model from physics.
The second part of this thesis belongs to the area of computational geometry and deals with algorithms for the Fréchet distance, which is a popular measure of similarity of two curves and can be computed in quadratic time (ignoring logarithmic factors). We provide the first conditional lower bound for this problem: No polynomial factor improvement over the quadratic running time is possible unless the Strong Exponential Time Hypothesis fails. We also present an improved approximation algorithm for realistic input curves.

Raphael REISCHUK
Declarative Design and Enforcement for Secure Cloud Applications

(Advisor: Prof. Michael Backes)
Tues, 16 Dec 2014, 15:00, building E1 7, room 0.01

The growing demands of users and industry have led to an increase in both size and complexity of deployed software in recent years. This tendency mainly stems from a growing number of interconnected mobile devices and from the huge amounts of data that is collected every day by a growing number of sensors and interfaces. Such increase in complexity imposes various challenges — not only in terms of software correctness, but also with respect to security.
In the talk, I will address three complementary approaches to cope with these challenges:
(i) appropriate high-level abstractions and verifiable translation methods to executable applications in order to provide guarantees towards flawless implementations, (ii) strong cryptographic mechanisms in order to realize the desired security goals, and (iii) convenient methods in order to incentivize the correct usage of existing techniques and tools.
In more detail, I will present a framework for the declarative specification of functionality and security, together with advanced compilers for the verifiable translation to executable applications. Moreover, I will present cryptographic primitives for the enforcement of cloud-based security properties: homomorphic message authentication codes ensure the correctness of evaluating functions over data outsourced to unreliable cloud servers; and efficiently verifiable non-interactive zero-knowledge proofs convince verifiers of computation results without the verifiers having access to the computation input.

Sebastian GERLING
Plugging in Trust and Privacy – Three Systems to Improve Widely Used Ecosystems
(Advisor: Prof. Michael Backes)
Tues, 16 Dec 2014, 9:00, building E1 7, room 0.01

The era of touch-enabled mobile devices has fundamentally changed our communication habits. Their high usability and unlimited data plans provide the means to communicate any place, any time and lead people to publish more and more (sensitive) information. Moreover, the success of mobile devices also led to the introduction of new functionality that crucially relies on sensitive data (e.g., location-based services). With our today's mobile devices, the Internet has become the prime source for information (e.g., news) and people need to rely on the correctness of information provided on the Internet. However, most of the involved systems are neither prepared to provide robust privacy guarantees for the users, nor do they provide users with the means to verify and trust in delivered content.
This dissertation introduces three novel trust and privacy mechanisms that overcome the current situation by improving widely used ecosystems. With WebTrust we introduce a robust authenticity and integrity framework that provides users with the means to verify both the correctness and authorship of data transmitted via HTTP. X-pire! and X-pire 2.0 offer a digital expiration date for images in social networks to enforce post-publication privacy. AppGuard enables the enforcement of fine-grained privacy policies on third-party applications in Android to protect the users privacy.

November

Christian KURZ
Constrained Camera Motion Estimation and 3D Reconstruction
(Advisor: Prof. Hans-Peter Seidel)
Fri, 28 Nov 2014, 11:00, building E1 4, room 0.19

Although imaging devices have become ubiquitous, content creation from visual data is still a tedious task and requires a high amount of skill and expertise. To reduce the amount of time and effort required to create high-quality content, new flexible and reliable tools are needed.
Advances to the state of the art of computer vision are presented to further this agenda:
First, a generalized framework for constrained camera motion estimation is introduced, which allows the simultaneous introduction of a variety of constraints on the camera configuration and the scene geometry.
Second, a new framework for symmetry-aware template fitting is presented, which allows the creation of high-quality models from low-quality input 3D scans.

David SPIELER
Numerical Analysis of Long-Run Properties for Markov Population Models
(Advisor: Prof. Verena Wolf)
Fri, 28 Nov 2014, 11:45, building E1 7, room 001

One of the most versatile modeling formalism is the one given by Markov chains as used for the performance analysis of queuing systems or for cost benefit ratio optimizations in the financial sector.
In systems biology, chemical reaction networks have originally been studied using deterministic models. However, when it recently became apparent that only stochastic effects can explain certain phenomenons, Markov chains again turned out to be a suitable modeling formalism in the form of Markov population models. Those Markov chains possess a structured but potentially infinite state space where each state encodes the current counts of a fixed number of population types.
Due to the infinite state space, classical steady state analysis methods can not be used directly. Hence, this doctoral thesis presents a highly efficient method to compute a finite state space truncation entailing most of the steady state probability mass. Further, stochastic complementation is lifted to the infinite state space setting and is combined with truncation based reachability analysis and aggregation to derive state wise steady state bounds. This method achieves high performance even for stiff systems. Particular attention is paid on a system's ability to maintain stable oscillations and thus optimized analysis methods are developed alongside. In order to prove their applicability, all techniques are evaluated on a large variety of biological models.

Gerrit KAHL

Dual Reality Framework – Basistechnologien zum Monitoring und Steuern von Cyber-Physischen Umgebungen
(Advisor: Prof. Antonio Krüger)
Wed, 26 Nov 2014, 14:00, building D3 2 (DFKI), room -2.17 (Reuse)

Diese Promotionsarbeit untersucht die Thematik des Monitoring und der Steuerung von Cyber-Physischen Umgebungen (CPE). In diesem Zusammenhang wird das Konzept und die Umsetzung eines Dual Reality (DR) Frameworks präsentiert, welches sich aus zwei Kom- ponenten zusammensetzt: dem Dual Reality Management Dashboard (DURMAD) zur interaktiven dreidimensionalen Visualisierung von CPE und dem Event Broadcasting Service (EBS), einer modularen Kommunikationsinfrastruktur. Hierbei stellt DURMAD basierend auf dem DR-Konzept den aktuellen Status der Umgebung in einem 3D-Modell visuell dar. Gleichzeitig umfasst es weitere Auswertungs- und Darstellungsformen, welche verschiedene Formen der Entscheidungsunterstützung für Manager der Umgebung bieten. Speziell entwickelte Filtermechanismen für den EBS ermöglichen eine Vorverarbeitung der Informationen vor dem Versenden bzw. nach dem Empfangen von Events. Durch offene Strukturen können externe Applikationen an das DR-Framework angeschlossen werden. Dies wird anhand von Objektgedächtnissen, semantischen Beschreibungen und Prozessmodellen präsentiert. Basierend auf einer Formalisierung von Dual Reality wird der Begriff Erweiterte Dual Reality (DR++) definiert, welcher auch die Auswirkungen von Simulationen in DR-Applikationen umfasst. Durch eine Integration des DR-Frameworks in das Innovative Retail Laboratory werden die Potenziale der erarbeiteten Konzepte anhand einer beispielhaften Implementierung in der Einzelhandelsdomäne aufgezeigt.

Matthias DIETZEN
Modeling protein interactions in protein binding sites and oligomeric protein complexes
(Advisor: Prof. Thomas Lengauer)
Thu, 20 Nov 2014, 16:00, building E1 4, room 0.24

Three-dimensional structures of protein-ligand and protein-protein complexes can provide key insights into biochemical processes within living cells, yet, their experimental determination is often expensive, time-consuming, or can fail due to the heterogeneity in the complex composition and thus the binding affinities of different components. A computational prediction of these structures can overcome these problems in certain cases and is thus highly demanded in many areas of research.
In this work, we address two questions: first, can one predict conformational changes of the protein backbone upon ligand binding, using the energetically most favorable motions obtained from normal mode analysis of elastic network models, and second, can one computationally assemble large protein complexes, using the structures and stoichiometries of their monomers and the approximate interaction geometries.
For the first problem, using a diverse set of 433 pairs of bound and unbound protein conformations, we could show that the benefit from such motions is small: modeling ligand-induced conformational changes using normal modes is rather ineffective. To solve the second problem, we have developed a novel scoring function and an efficient algorithm for iterative complex assembly based on pairwise dockings, 3D-MOSAIC, that, on a diverse benchmark set of 308 complexes, can accurately and efficiently assemble protein complexes of up to 60 monomers and 15 protein types.

Ilya RESHETOUSKI
Kaleidoscopic Imaging
(Advisor: Dr. habil. Ivo Ihrke)
Thu, 6 Nov 2014, 10:00, building E1 4, room 0.19

Kaleidoscopes have a great potential in computational photography as a tool for redistributing light rays. In time-of-flight imaging the concept of the kaleidoscope is also useful when dealing with the reconstruction of the geometry that causes multiple reflections. This work is a step towards opening new possibilities for the use of mirror systems as well as towards making their use more practical. The focus of this work is the analysis of planar kaleidoscope systems to enable their practical applicability in 3D imaging tasks.
We analyse important practical properties of mirror systems and develop a theoretical toolbox for dealing with planar kaleidoscopes. Based on this theoretical toolbox we explore the use of planar kaleidoscopes for multiview imaging and for the acquisition of 3D objects. The knowledge of the mirrors positions is crucial for these multi-view applications. On the other hand, the reconstruction of the geometry of a mirror room from time-offlight measurements is also an important problem. We therefore employ the developed tools for solving this problem using multiple observations of a single scene point.

October

Johannes GÜNTHER
Ray Tracing of Dynamic Scenes
(Advisor: Prof. Philipp Slusallek)
Fri, 24 Oct 2014, 10:15, building E1 4, room 0.19

In the last decade ray tracing performance reached interactive frame rates for non-trivial scenes, which roused the desire to also ray trace dynamic scenes. Changing the geometry of a scene, however, invalidates the precomputed auxiliary data-structures needed to accelerate ray tracing. In this thesis we review and discuss several approaches to deal with the challenge of ray tracing dynamic scenes.
In particular we present the motion decomposition approach that avoids the invalidation of acceleration structures due to changing geometry. To this end the animated scene is analyzed in a preprocessing step to split it into coherently moving parts. Because the relative movement of the primitives within each part is small it can be handled by special, pre-built kd-trees. Motion decomposition enables ray tracing of predefined animations and skinned meshed at interactive frame rates. Our second main contribution is the streamed binning approach. It approximates the

evaluation of the cost function that governs the construction of optimized kd-trees and BVHs. As a result construction speed especially for BVHs can be increased by one order of magnitude while still maintaining their high quality for raytracing.

Markus MAINBERGER
PDE-based image compression based on edges and optimal data
(Advisor: Prof. Joachim Weickert)
Mon, 6 Oct 2014, 14:00, building E1 7, room 001

This thesis investigates image compression with partial differential equations (PDEs) based on edges and optimal data.
It first presents a lossy compression method for cartoon-like images. Edges together with some adjacent pixel values are extracted and encoded. During decoding, information not covered by this data is reconstructed by PDE-based inpainting with homogeneous diffusion. The result is a compression codec based on the perceptual meaningful image features which is able to outperform JPEG and JPEG2000.
In contrast, the second part of the thesis focuses on the optimal selection of inpainting data. The proposed methods allow to recover a general image from only 4% of all pixels almost perfectly, even with homogeneous diffusion inpainting.
A simple conceptual encoding shows the potential of an optimal data selection for image compression: The results beat the quality of JPEG2000 when anisotropic diffusion is used for inpainting.
Finally, the thesis shows that the combination of the concepts allows for further improvements.

September

Ralf KARRENBERG
Automatic SIMD Vectorization of SSA-based Control Flow-Graphs
(Advisor: Prof. Sebastian Hack)
Tues, 30 Sept 2014, 11:00, building E1 7, room 001

Data-parallel applications such as particle simulations, stock option price estimation, or video decoding require the same computations to be performed on huge amounts of data. To exploit all available parallelism for best performance, the code must make use of all available cores as well as SIMD instruction sets: These allow to efficiently execute a single operation on multiple input values at once per core.
This thesis presents Whole-Function Vectorization (WFV), an approach that allows a compiler to automatically exploit SIMD instructions in data-parallel settings. Without WFV, one processor core executes a single instance of a data-parallel function. WFV transforms the function to execute multiple instances at once using SIMD instructions.
For simple, straight-line code, the transformation is easily applied and delivers drastically improved performance. However, problems such as particle simulations or shading in computer graphics exhibit more complex code. Our evaluation presents evidence that in such scenarios, a naive WFV approach will often not improve performance or even slow down execution.
We describe an advanced WFV algorithm that includes a variety of analyses and code generation techniques and show that it improves the performance of the generated code.

August

Markus LÖCHTEFELD
Spatial Interaction with Mobile Projected Displays
(Advisor: Prof. Antonio Krüger)
Thu, 14 August 2014, 15:15h, building D3 2 (DFKI), Reuse seminar room

The miniaturization of projection technology recently reached a point where battery powered pico-projectors can be carried in ones pocket. These projectors not only allow for mobile stand-alone devices but also the integration into other devices. These new class of devices allow to create ad-hoc mobile projected displays everywhere. Until today this development is mainly technology driven. While projector phones – smart- phones with integrated projector – are available, the fact that they are mainly hand-held devices, present a variety of challenges for application and interaction techniques, for the mobile projected display they provide. In this thesis we enable spatial interaction with mobile projected displays to overcome the current technological driven approaches. We contribute by investigating the following three different directions: Applications exploiting spatial features; Analysis of spatial alignment of mobile projected displays; and Interaction techniques exploiting spatial memory techniques. Through the development and evaluation of applications and interaction techniques that rely on spatial features we exploit the users spatial memory and allow mobile projected displays to become a valuable addition of our current mobile devices.

Pepijn CROUZEN
Modularity and Determinism in Compositional Markov Models
(Advisor: Prof. Holger Hermanns)
Thu, 14 August 2014, 14:00h, building E1 7, room 0.01

Markov chains are a versatile and widely used means to model an extensive variety of stochastic phenomena, but describing a complex system as a monolithic Markov chain is difficult and error-prone. In this thesis we show that we can construct such complex Markov chains in a sound manner through the composition of a number of simple input/output interactive Markov chains (I/O-IMCs), which arise as an orthogonal combination of continuous-time Markov chains and input/output automata.

July

Faraz Makari MANSHADI
Scalable Optimization Algorithms for Recommender Systems
(Advisor: Dr. Rainer Gemulla)

Tue, 15 July 2014, 12:15h, building E1 4, room 024

Recommender systems have now gained significant popularity and been widely used in many e-commerce applications. Predicting user preferences is a key step to providing high quality recommendations. In practice, however, suggestions made to users must not only consider user preferences in isolation; a good recommendation engine also needs to account for certain constraints. For instance, an online video rental that suggests multimedia items (e.g., DVDs) to its customers should consider the availability of DVDs in stock to minimize customer waiting times for accepted recommendations. Moreover, every user should receive a small but sufficient number of suggestions that the user is likely to be interested in.
This thesis aims to develop and implement scalable optimization algorithms that can be used (but are not restricted) to generate recommendations satisfying certain objectives and constraints like the ones above. State-of-the-art approaches lack efficiency and/or scalability in coping with large real-world instances, which may involve millions of users and items. First, we study large-scale matrix completion in the context of collaborative filtering in recommender systems. For such problems, we propose a set of novel shared-nothing algorithms which are designed to run on a small cluster of commodity nodes and outperform alternative approaches in terms of efficiency, scalability, and memory footprint. Next, we view our recommendation task as a generalized matching problem, and propose the first distributed solution for solving such problems at scale. Our algorithm is designed to run on a small cluster of commodity nodes (or in a MapReduce environment) and has strong approximation guarantees. Our matching algorithm relies on linear programming. To this end, we present an efficient distributed approximation algorithm for mixed packing-covering linear programs, a simple but expressive subclass of linear programs. Our approximation algorithm requires a poly-logarithmic number of passes over the input, is simple, and well-suited for parallel processing on GPUs, in shared-memory architectures, as well as on a small cluster of commodity nodes.

Chenglei WU
Inverse Rendering for Scene Reconstruction in General Environments
(Advisor: Prof. Christian Theobalt)
Thu, 10 July 2014, 14:00h, building E1 4, room 019

Techniques for capturing the real world, which are able to generate 3D models from captured images or videos, are a hot research topic in computer graphics and computer vision. Despite significant progress, many methods are still highly constrained and are confined to only studio environments. In this thesis, we proposes new scene reconstruction techniques that succeed in general environments, even using as few as two cameras. Contributions are made in terms of reducing the constraints of marker-less performance capture on lighting, background and the required number of cameras. The primary theoretical contribution lies in the investigation of light transport mechanisms for high-quality 3D reconstruction in general environments.

Bastian BEGGEL
Determining and Utilizing the Quasispecies of the Hepatitis B Virus in Clinical Applications
(Advisor: Prof. Thomas Lengauer)
Tue, 8 July 2014, 14:00h, building E1 4, room 021

Chronic hepatitis B caused by infection with the hepatitis B virus (HBV) affects about 240 million people worldwide and is one of the major causes of severe liver cirrhosis and liver cancer. The population of viruses within a host is called the viral quasispecies. This thesis provides statistical methods to infer relevant information about the viral quasispecies of HBV to support treatment decisions. We introduce a new genotyping methodology to identify dual infections, which can help to quantify the risk of interferon therapy failure. We present a method to infer short-range linkage information from Sanger sequencing chromatograms, a method to support treatment adjustment after the development of resistance to nucleos(t)ide analogs. Additionally, we provide the first fullgenome analysis of the G-to-A hypermutation patterns of the HBV genome. Hypermutated viral genomes form a subpopulation of the quasispecies caused by proteins of the human innate immune system editing the genome of exogenous viral agents. We show that hypermutation is associated with the natural progression of hepatitis B, but does not correlate with treatment response to interferon.

June

Christian KLEIN
Matrix Rounding, Evolutionary Algorithms, and Hole Detection
(Advisor: Prof. Benjamin Doerr)
Mon, 23 June 2014, 15:00h, building E1 4, room 024

In this thesis we study three different topics from the field of algorithms and data structures. First, we investigate a problem from statistics. We give two randomised algorithms that can round matrices of fractional values to integer-valued matrices. These matrices will exhibit only small rounding errors for sums of initial row or column entries. Both algorithms also round each entry up with probability equal to its fractional value. We give a derandomisation of both algorithms.
Next, we consider the analysis of evolutionary algorithms (EAs). First, we analyse an EA for the Single Source Shortest Path problem. We give tight upper and lower bounds on the optimisation time of the EA. For this, we develop some new techniques for such analyses. We also analyse an EA for the All-Pairs Shortest Path problem. We show that adding crossover to this algorithm provably decreases its optimisation time. This is the first time that the usefulness of crossover has been shown for a non-constructed combinatorial problem.
Finally, we examine how to retrieve the implicit geometric information hidden in the communications graph of a wireless sensor network. We give an algorithm that is able to identify wireless nodes close to a hole in this network based on the connectivity information alone. If the input fulfils several preconditions, then the algorithm finds a node near each boundary point and never misclassifies a node.

Rayna DIMITROVA

Synthesis and Control of Infinite-State Systems with Partial Observability
(Advisor: Prof. Bernd Finkbeiner)
Thu, 12 June 2014, 10:30h, building E1 7, room 0.01

Synthesis methods automatically construct a system, or an individual component within a system, such that the result satisfies a given specification. The synthesis procedure must take into account the component's interface and deliver implementations that comply with its limitations. For example, what a component can observe about its environment may be restricted by imprecise sensors or inaccessible communication channels. In addition, sufficiently precise models of a component's environment are typically infinite-state, for example due to modeling real time or unbounded communication buffers. This thesis presents novel synthesis methods for infinite-state systems with partial observability. The contributions are structured into three parts, corresponding to a classification of such systems according to the interface between the synthesized component and its environment. First, we obtain decidability results for a class of systems where the synthesized component has a given finite set of possible observations and a finite set of possible actions, which is the case, for example, for lossy channel systems. We provide symbolic synthesis algorithms for lossy channel systems with partial observability for safety and reachability specifications.Our second contribution is a counterexample-guided abstraction refinement scheme for synthesis of systems in which the actions available to the component are still finitely many, but no finite set of possible observations is fixed a priori. The developed method is applicable to infinite-state systems such as mutex protocols, which are out of the scope of other available techniques. Finally, we study systems in which, in addition to having a possibly infinite set of observations, the component can choose between infinitely many possible actions. This is the case, for example, in timed controller synthesis with partial observability. We extend the abstraction-refinement procedure to develop the first systematic method for automatic generation of observation predicates for timed control with safety requirements.

Yassen ASSENOV
Identification and prioritization of genomic loci with disease-specific DNA methylation
(Advisor: Prof. Thomas Lengauer)
Tue, 10 June 2014, 14:00h, building E1 4, room 021

DNA methylation patterns and other epigenetic mechanisms are an indispensable cell apparatus in development; they respond to environmental stimuli and are dysregulated in cancer and other diseases.
The abundance and growing amount of methylation and other epigenetic datasets act as a valuable resource in the search for answers to biological and medical questions, but they also pose great challenges due to their heterogeneity and high dimensionality. My thesis introduces computational techniques for handling DNA methylation data with a focus on disease-oriented studies. It addresses the questions of quality control and normalization, inter- and intra-group variability, identification of differentially methylated loci, prioritization of biomarker candidates, as well as prediction of cancer type and other phenotypes.

Konstantin HALACHEV
Exploratory visualizations and statistical analysis of large, heterogeneous epigenetic datasets
(Advisor: Prof. Thomas Lengauer)
Tue, 10 June 2014, 12:00h, building E1 4, room 024

Epigenetic marks, such as DNA methylation and histone modifications, are important regulatory mechanisms that allow a single genomic sequence to give rise to a complex multicellular organism. The increasing quantity of epigenetic data highlights a major bottleneck in bioinformatic research, namely the lack of bioinformatic tools for analyzing these data. I present methodology and software toolkit (EpiExplorer and EpiGRAPH) to explore epigenomic data interactively that facilitates trained biologists to recognize interesting aspects and pursue them further. EpiExplorer applies state-of-the-art information retrieval methods and indexing structures to offer instant interactive exploration of large epigenetic datasets. EpiGRAPH helps identify and model significant biological associations among epigenetic and genetic properties for sets of regions. Using EpiExplorer and EpiGRAPH, independently or in a pipeline, provides the bioinformatic community with access to large databases of annotations, allows for exploratory visualizations or statistical analysis and facilitates reproduction and sharing of results.

Arnaud FIETZKE
Labelled Superposition
(Advisor: Prof. Christoph Weidenbach)
Thu, 5 June 2014, 9:00h, building E1 5, room 002

The work presented in this thesis consists of two parts: The first part presents a formalization of the splitting rule for case analysis in superposition and a proof of its correctness, as well as the integration into splitting of a novel approach to lemma learning, which allows the derivation of non-ground lemmas. The second part deals with the application of hierarchic superposition, and in particular superposition modulo linear arithmetic SUP(LA), to the verification of timed and probabilistic timed systems. It contains the following three main contributions: Firstly, a SUP(LA) decision procedure for reachability in timed automata, which is among the first decision procedures for free first-order logic modulo linear arithmetic; secondly, an automatic induction rule for SUP(LA) based on loop detection, whose application allows a generalization of the decidability result for timed automata to timed automata with unbounded integer variables; and thirdly, a formalism for modelling probabilistic timed systems with first-order background theories, as well as a novel approach for the analysis of such systems based on a reduction to model checking using SUP(LA) proof enumeration.

May

Hagen BLANKENBURG
Computational Methods for Integrating and Analyzing Human Systems

Biology Data
(Advisor: Prof. Mario Albrecht)
Wed, 28 May 2014, 14:30h, building E1 7, room 008

The combination of heterogeneous biological datasets is a key requirement for modern molecular systems biology. Of particular importance for our understanding of complex biological systems like the human cell are data about the interactions of proteins with other molecules. In this thesis, we develop and apply methods to improve the availability and the quality of such interaction data. We also demonstrate how these data can be used in interdisciplinary studies to discover new biological results.
First, we develop technical systems for the instant integration of interaction data that are stored and maintained in separate online repositories. Second, we implement a computational framework for the application of multiple scoring algorithms to qualitatively assess different aspects of interaction data. Our methods are based on distributed client-server systems, ensuring that the services can be updated permanently. This promotes access to interaction data by allowing researchers to expand the client-server systems with their own service.
Third, we focus our application studies on integrative network-based analyses of human host factors for viral infections. Our applications provide new biological insights into the life cycle of the hepatitis C virus and identify new potential candidates for antiviral drug therapy.

Richard RÖTTGER
Active Transitivity Clustering of Large-Scale Biomedical Datasets
(Advisor: Prof. Jan Baumbach)
Wed, 28 May 2014, 11:00h, building E1 1, room 4.07

Clustering is a popular computational approach for partitioning data sets into groups of objects that share common traits. Due to recent advances in wet-lab technology, the amount of available biological data grows exponentially and increasingly poses problems in terms of computational complexity for current clustering approaches. In this thesis, we introduce two novel approaches, TransClustMV and ActiveTrans-Clust that enable the handling of large scale datasets by reducing the amount of required information drastically by means of exploiting missing values.
Furthermore, there exists a plethora of different clustering tools and standards making it very difficult for researchers to choose the correct methods for a given problem. In order to clarify this multifarious field, we developed ClustEval which streamlines the clustering process and enables practitioners conducting large-scale cluster analyses in a standardized and bias-free manner.
We conclude the thesis by demonstrating the power of clustering tools and the need for the previously developed methods by conducting real-world analyses. We transferred the regulatory network of E. coli K-12 to pathogenic EHEC organisms based on evolutionary conservation therefore avoiding tedious and potentially dangerous wet-lab experiments. In another example, we identify pathogenicity specific core genomes of actinobacteria in order to identify potential drug targets.

Maximillian DYLLA
Efficient Querying and Learning in Probabilistic and Temporal Databases
(Advisor: Prof. Gerhard Weikum)
Fri, 9 May 2014, 11:00h, building E1 4, room 0.24

Probabilistic databases store, query, and manage large amounts of uncertain information. This thesis advances the state-of-the-art in probabilistic databases in three different ways:
1. We present a closed and complete data model for temporal probabilistic databases and analyze its complexity. Queries are posed via temporal deduction rules which induce lineage formulas capturing both time and uncertainty.
2. We devise a methodology for computing the top-k most probable query answers. It is based on first-order lineage formulas representing sets of answer candidates. Theoretically derived probability bounds on these formulas enable pruning low-probability answers.
3. We introduce the problem of learning tuple probabilities which allows updating and cleaning of probabilistic databases. We study its complexity, characterize its solutions, cast it into an optimization problem, and devise an approximation algorithm based on stochastic gradient descent.
All of the above contributions support consistency constraints and are evaluated experimentally.

Sarath Kumar KONDREDDI
Human Computing and Crowdsourcing Methods for Knowledge Acquisition

(Advisor: Prof. Gerhard Weikum)
Tue, 6 May 2014, 12:15h, building E1 4, room 0.24

Ambiguity, complexity, and diversity in natural language textual expressions are major hindrances to automated knowledge extraction. As a result state-of-the-art methods for extracting entities and relationships from unstructured data make incorrect extractions or produce noise. With the advent of human computing, computationally hard tasks have been addressed through human inputs. While text-based knowledge acquisition can benefit from this approach, humans alone cannot bear the burden of extracting knowledge from the vast textual resources that exist today. Even making payments for crowdsourced acquisition can quickly become prohibitively expensive.
In this thesis we present principled methods that effectively garner human computing inputs for improving the extraction of knowledge-base facts from natural language texts. Our methods complement automatic extraction techniques with human computing to reap the benefits of both while overcoming each other's limitations. We present the architecture and implementation of HIGGINS, a system that combines an information extraction (IE) engine with a human computing (HC) engine to produce high quality facts. The IE engine combines statistics derived from large Web corpora with semantic resources like WordNet and ConceptNet to construct a large dictionary of entity and relational phrases. It employs specifically designed statistical language models for phrase relatedness to come up with questions and relevant candidate answers that are presented to human workers. Through extensive experiments we establish the superiority of this approach in extracting relation-

centric facts from text. In our experiments we extract facts about fictitious characters in narrative text, where the issues of diversity and complexity in expressing relations are far more pronounced. Finally, we also demonstrate how interesting human computing games can be designed for knowledge acquisition tasks.

April

Bastian BLANKENBURG
Coalition Formation among Rational Agents in Uncertain and Untrustworthy Environments
(Advisor: Dr. Matthias Klusch)
Wed, 9 April 2014, 16:00h, building D3 4, VisCenter -1.63

Multiagent coalition formation based on cooperative game theory is a means to let rational software agents cooperate when trying to solve complex tasks on behalf of their users, increasing the agents' benefits. A number of such approaches have been proposed in the literature. But when applied in open environments, such as the internet, additional problems arise that are not well coped with by the traditional approaches. In this theses, solutions are proposed for some of these problems, so that coalition formation is better suited to be applied in realistic settings:
- Uncertainty: in open environments, agents often do not have complete knowledge. It is shown how to form coalitions efficiently by modeling uncertainty as fuzzy numbers. Simulation results are provided. Additionally, a method for coalition formation is proposed which is shown to form stable coalitions with guaranteed risk bounds.
- Defrauding or unreliable agents: it might be expected in open environments that some agents try to unsolicitedly increase their own profits at the cost of others by deception, or are generally unreliable. In this thesis, we combine a trust measure based approach with cryptographic techniques to obtain payment and communication protocols that are shown to hamper successful deception.
- Privacy preservation: most approaches require agents to reveal a considerable amount of information to each other, such as individual costs and valuations of certain outcomes. This might be unacceptable when personal, financial or other delicate data is concerned. To this end, the (to our best knowledge) first privacy preserving coalition formation algorithm is proposed.

March

Steffen METZGER
User-centric Knowledge Extraction and Maintenance
(Advisor: Prof. Ralf Schenkel)
Fri, 28 March 2014, 9:00h, building E1 4, room 0.24

An ontology is a machine readable knowledge collection. There is an abundance of information available for human consumption. Thus, large general knowledge ontologies are typically generated tapping into this information source using imperfect automatic extraction approaches that translate human readable text into machine readable semantic knowledge. This thesis provides methods for user-driven ontology generation and maintenance. In particular, this work consists of three main contributions:
1. An interactive human-supported extraction tool: LUKe. The system extends an automatic extraction framework to integrate human feedback on extraction decisions and extracted information on multiple levels.
2. A document retrieval approach based on semantic statements: S3K. While one application is the retrieval of documents that support extracted information to verify the correctness of the piece of information, another application in combination with an extraction system is a fact based indexing of a document corpus allowing statement based document retrieval.
3. A method for similarity based ontology navigation: QBEES. The approach enables search by example. That is, given a set of semantic entities, it provides the most similar entities with respect to their semantic properties considering different aspects.
All three components are integrated into a modular architecture that also provides an interface for third-party components.

Marcus ROHRBACH
Combining Visual Recognition and Computational Linguistics
(Advisor: Prof. Bernt Schiele)
Wed, 26 March 2014, 18:00h, building E1 4, room 0.24

Extensive efforts are being made to improve visual recognition and semantic Understanding of language. However, surprisingly little has been done to exploit the mutual benefits of combining both fields. In this thesis we show how the different fields of research can profit from each other.
First, we scale recognition to 200 unseen object classes and show how to extract robust semantic relatedness from linguistic resources. Our novel approach extends zero-shot to few shot recognition and exploits unlabeled data by adopting label propagation for transfer learning.
Second, we capture the high variability but low availability of composite activity videos by extracting the essential information from text descriptions. For this we recorded and annotated a corpus for fine-grained activity recognition. We show improvements in a supervised case but we are also able to recognize unseen composite activities.
Third, we present a corpus of videos and aligned descriptions. We use it for grounding activity descriptions and for learning how to automatically generate natural language descriptions for a video. We show that our proposed approach is also applicable to image description and that it outperforms baselines and related work.

In summary, this thesis presents a novel approach for automatic video description and shows the benefits of extracting linguistic knowledge for object and Activity recognition as well as the advantage of visual recognition for understanding activity descriptions.

Christoph BAUMANN
Ownership-Based Order Reduction and Simulation in Shared-Memory

Concurrent Computer Systems
(Advisor: Prof. Wolfgang Paul)
Wed, 19 March 2014, 16:00h, building E1 7, room 001

The highest level of confidence in the correct functionality of system software can be gained from a pervasive formal verification approach, where the high-level language application layer is connected to the gate-level hardware layer through a stack of semantic layers coupled by simulation theorems. While such semantic stacks exist for sequential systems, the foundational theory of semantic stacks for concurrent systems is still incomplete. This thesis contributes to close this gap.
First we prove a general order reduction theorem establishing a model where processes are executing blocks of steps, being only interleaved at selectable interleaving-points. An ownership-based memory access policy is imposed to prove commutativity properties for non-synchronizing steps, enabling the desired reordering. In contrast to existing work, we only assume properties on the order-reduced level, thus providing a complete abstraction.
We then apply sequential simulation theorems on top of the block schedules and prove a general simulation theorem between two abstract concurrent systems including the transfer of safety properties.
Finally we instantiate our frameworks with a MIPS instruction set architecture, a macro assembler (MASM) semantics, and an intermediate language semantics for C. Applying the concurrent simulation theorem, we justify the concurrent semantics of MASM and C against their ISA implementation.

Arjun JAIN
Data-driven Methods for Interactive Visual Content Creation and Manipulation
(Advisor: Prof. Hans-Peter Seidel)
Wed, 19 March 2014, 10:00h, building E1 4 (MPII), room 0.19

Software tools for creating and manipulating visual content - be it for images, video or 3D models - are often difficult to use and involve a lot of manual interaction at several stages of the process. Coupled with long processing and acquisition times, content production is rather costly and poses a potential barrier to many applications. Although cameras now allow anyone to easily capture photos and video, tools for manipulating such media demand both artistic talent and technical expertise. However, at the same time, vast corpuses with existing visual content such as Flickr, Youtube or Google 3D Warehouse are today available and easily accessible.
This thesis proposes a data-driven approach to tackle the above mentioned problems encountered in content generation. To this end, statistical models trained on semantic knowledge harvested from existing visual content corpuses are created. Using these models, we then develop tools which are easy to learn and use even by novice users but still produce high-quality content. These tools have intuitive interfaces, and enable the user to have a precise and flexible control. Specifically, we apply our models to create tools to simplify the tasks of video manipulation, 3D modeling and material assignment to 3D objects.

Megha KHOSLA
Multiple Choice Allocations with Small Maximum Loads
(Advisor: Prof. Kurt Mehlhorn)
Tue, 4 March 2014, 11:00h, building E1 4 (MPII), room 0.24

The idea of using multiple choices to improve allocation schemes is now well understood and is often illustrated by the following example. Suppose n balls are allocated to n bins with each ball choosing a bin independently and uniformly at random. The maximum load, or the number of balls in the most loaded bin, will then be approximately $\log n \over \log \log n$ with high probability. Suppose now the balls are allocated sequentially by placing a ball in the least loaded bin among the $k \ge 2$ bins chosen independently and uniformly at random. Azar, Broder, Karlin, and Upfal showed that in this scenario, the maximum load drops to ${\log \log n \over \log k} + \Theta(1)$, with high probability, which is an exponential improvement over the previous case.
In this thesis we investigate multiple choice allocations from a slightly different perspective. Instead of minimizing the maximum load, we fix the bin capacities and focus on maximizing the number of balls that can be allocated without overloading any bin. In the process that we consider we have $m = \lfloor cn \rfloor$ balls and n bins. Each ball chooses k bins independently and uniformly at random. Is it possible to assign each ball to one of its choices such that the no bin receives more than ℓ balls? For all $k \ge 3$ and $\ell \ge 2$ we give a critical value, $c_{k,\ell}^*$, such that when $c < c_{k,\ell}^*$ an allocation is possible with high probability and when $c > c_{k,\ell}^*$ this is not the case.
In case such an allocation exists, how quickly can we find it? Previous work on total allocation time for case $k \ge 3$ and $\ell = 1$ has analyzed a breadth first strategy which is shown to be linear only in expectation. We give a simple and efficient algorithm which we also call local search allocation (LSA) to find an allocation for all $k \ge 3$ and $\ell = 1$. Provided the number of balls are below (but arbitrarily close to) the theoretical achievable load threshold, we give a inear bound for the total allocation time that holds with high probability.
We demonstrate, through simulations, an order of magnitude improvement for total and maximum allocation times when compared to the state of the art method.
Our results find applications in many areas including hashing, load balancing, data management, orientability of random hypergraphs and maximum matchings in a special class of bipartite graphs.

Jörg HERTER
Timing-Predictable Memory Allocation in Hard Real-Time Systems
(Advisor: Prof. Reinhard Wilhelm)
Mon, 3 March 2014, 14:15, building E1 1, room 4.07

For hard real-time applications, tight provable bounds on the application's worst-case execution time must be derivable. Employing dynamic memory allocation, in general, significantly decreases an application's timing predictability. In consequence, current hard real-time applications rely on static memory management. This thesis studies how the predictability issues of dynamic memory allocation can be overcome and dynamic memory allocation be enabled for hard real-time applications. We give a detailed analysis of the predictability challenges

imposed on current state-of-the-art timing analyses by dynamic memory allocation. We propose two approaches to overcome these issues and enable dynamic memory allocation for hard real-time systems: automatically transforming dynamic into static allocation and using a novel, cache-aware and predictable memory allocator. Statically transforming dynamic into static memory allocation allows for very precise WCET bounds as all accessed memory addresses are completely known. However, this approach requires much information about the application's allocation behavior to be available statically. For programs where a static precomputation of a suitable allocation scheme is not applicable, we investigate approaches to construct predictable dynamic memory allocators to replace the standard, general-purpose allocators in real-time applications.

We present evaluations of the proposed approaches to evidence their practical applicability.

February

Sven GEHRING
Interaction with Media Facades – The design of interactive systems for large-scale urban screens
(Advisor: Prof. Antonio Krüger)
Fri, 21 February 2014, 15:00h, building D3 2 (DFKI), Reuse seminar room

Media facades are a prominent example of the digital augmentation of urban spaces. They denote the concept of turning the surface of a building into a large-scale urban screen. Due to their enormous size, they require interaction at a distance and they have a high level of visibility. Additionally, they are situated in a highly dynamic urban environment with rapidly changing conditions, which results in settings that are neither comparable, nor reproducible. Altogether, this makes the development of interactive media facade installations a challenging task.
The thesis investigates the design of interactive installations for media facades holistically. A theoretical analysis of the design space for interactive installations for media facades is conducted to derive taxonomies to put media facade installations into context. Along with this, a set of observations and guidelines is provided to derive properties of the interaction from the technical characteristics of an interactive media facade installation. This thesis further provides three novel interaction techniques addressing the form factor and resolution of the facade, without the need for additionally instrumenting the space around the facades. The thesis contributes to the design of interactive media facade installations by providing a generalized media facade toolkit for rapid prototyping and simulating interactive media facade installations, independent of the media facade's size, form factor, technology and underlying hardware.

Sven-Eric SCHELLHORN
Going Viral
(Advisor: Prof. Thomas Lengauer)
Tue, 4 February 2014, 14:15h, building E1 4 (MPI-Inf), room 0.24

Viruses are of considerable interest for several fields of bioscience research. The genomic richness of these entities, their environmental abundance, as well as their high adaptability and, potentially, pathogenicity make treatment of viral diseases challenging and antiviral research medically relevant. Especially if seen in a context of clinical reality that is dominated by chronic disease, drug resistance, and side effects, it becomes clear that an integrated view on virological data analysis spanning aspects from basic research to drug development and clinical applications is required in order to facilitate medical progress. Such an integrated view should, at the least, encompass early detection of emergent human pathogenic viruses (virus discovery), support antiviral drug discovery (drug target identification), and aid in developing and optimizing clinical treatment of infectious diseases in order to to curb viral drug resistance and increase therapy success (treatment optimization). This treatise aims to provide contributions to all three of these areas. Specifically, the thesis proposes three novel contributions to antiviral research that each concern analysis procedures of highthroughput experimental genomics data. First, a sensitive approach for detecting viral genomes and transcripts in sequencing data of human cancers is presented that improves upon prior approaches by allowing detection of viral nucleotide sequences that consist of human-viral homologs or are diverged from known reference sequences. Second, a computational method for inferring physical protein contacts from experimental protein complex purification assays is put forward that allows statistically meaningful integration of multiple data sets and is able to infer protein contacts of transiently binding protein classes such as kinases and molecular chaperones. Third, an investigation of minute changes in viral genomic populations upon treatment of patients with the mutagen ribavirin is presented that first characterizes the mutagenic effect of this drug on the hepatitis C virus based on deep sequencing data.

January

Lars FEUERBACH
Evolutionary Epigenetics – Identifying Functional Genome Elements by Epigenetic Footprints in the DNA
(Advisor: Prof. Thomas Lengauer)
Thu, 16 January 2014, 11:00h, building E1 4 (MPI-Inf), room 0.21

Over the last decade advances in genome sequencing have substantially increased the amount of available genomic DNA sequences. While these rich resources improved our understanding of genome function, research of the epigenome as a transient but heritable memory system of the cell could only profit from this development indirectly. Although epigenetic information in form of DNA methylation is not directly encoded in the genomic nucleotide sequence, it increases the mutation rate of cytosine-guanine dinucleotides by the CpG decay effect, and thus establishes epigenetic footprints in the DNA.
This thesis proposes four approaches to facilitate this information for research. For largely uncharacterized genomes, CgiHunter presents an exhaustive algorithm for an unbiased DNA sequence-based annotation of CpG islands, as regions that are protected from CpG decay. For species with well characterized point mutation frequencies, EqiScore identifies regions that evolve under distinct DNA methylation

levels. Furthermore, the derived equilibrium distributions for methylated and unmethylated genome region predict the evolutionary robustness of transcription-factor binding site motifs against the CpG decay effect. The AluJudge annotation and the underlying L-score provide a method to identify putative active copies of CpG-rich transposable elements within genomes. Additionally, epigenetic footprints in these sequences are applied to predict the germline epigenome of their loci. Moreover, AluJudge provides support for a targeted removal of epigenetically silenced repeat copies from CpG island annotations, which are subjected to a methylation-induced erosion process. Finally, the an adapted version of Felsensteins tree pruning algorithm enables the prediction of the germline methylome from multiple alignments of evolutionary related genome loci.

In a number of case studies on the human genome, I demonstrate how this evolutionary epigenomics tool kit can be applied to enhance the epigenomic characterization of the large quantity of currently sequenced vertebrate genomes. Furthermore, these studies show how to improve the identification of novel epigenetic functional genome regions in already well characterized species. Finally, the tool kit opens new avenues for computer-based research of the evolution of genome-wide DNA methylation.