1st European Meeting on Algorithmic Challenges of Big Data (ACBD 2014)

Warsaw, Poland, May 5-7, 2014

Hannah Bast (University of Freiburg)

Semantic Search on Big Data

Semantic search goes beyond standard full-text search in that it tries to understand the meaning of the query and of the searched documents. In this talk, I will provide a glimpse of the state of the art in this area and I will show various demos. The focus will be on the underlying algorithmic problems, and on the particular problems that arise when the data becomes really big.

Graham Cormode (University of Warwick)

Sketches, Streaming and Big Data

One natural way to deal with the challenge of Big Data is to make the data smaller. That is, to seek a compact (sublinear) representation of the data so that certain properties are (approximately) preserved. We can think of these as a generalization of sufficient statistics for properties of the data. The area of “streaming algorithms” seeks algorithms which can build such a summary as information is processed incrementally. An important class of streaming algorithms are sketches: carefully designed random projections of the input data that can be computed efficiently under the constraints of the streaming model. These have the attractive property that they can be easily computed in parallel over partitions of the input. They aim at optimizing a variety of properties: the size of the summary; the time required to compute the summary; the number of 'true' random bits required; and the accuracy guarantees that result.

Artur Czumaj (University of Warwick)

Testing Cluster Structure of Graphs

Cluster analysis is a fundamental task in data analysis that aims at partitioning a set of objects into a disjoint collection of objects with similar characteristics. In this talk, we will use the concept of conductance to measure the quality of cluster structure and will focus on a question of approximately recognizing cluster structure of a graph in sublinear time in the framework of property testing in the bounded degree model. We show how to test in $O^*(\sqrt{n})$ time whether a graph with n nodes can be partitioned into no more than k parts (clusters) such that the outer-conductance of each cluster is at most ϕ_{out} and the inner-conductance of the induced subgraph on each cluster is at least ϕ_{in}, for a large spectrum of parameters k, ϕ_{out}, ϕ_{in}. By the lower bound of $\Omega(\sqrt{n})$ for testing graph expansion, which corresponds to the case when $k = 1$ in our problem, our algorithm is asymptotically optimal up to polylogarithmic factors.

This is joint work with Pan Peng and Christian Sohler.

Ilias Diakonikolas (University of Edinburgh)

Learning Structured Distributions from a Constant Number of Samples

I will overview recent results on learning structured distributions. These results follow from a new approach to density estimation based on piecewise polynomial approximations. The key tool that enables our new approach is a computationally efficient general algorithm for learning probability distributions that are well approximated by piecewise polynomial density functions.

Based on joint works with S. Chan, R. Servedio and X. Sun.

Pierre Fraigniaud (LIAFA)

Distributed Decision and Verification

This talk will survey recent advances in the framework of distributed decision and verification. The former refers to the ability of system components (e.g., processors) to collectively check whether the system satisfies some prescribed property. The latter refers to the same task, when the system components are additionally provided with certificates. The talk with consider three general contexts in which distributed decision and verification were recently investigated: the LOCAL model, capturing the essence of local computation, the CONGEST model, capturing the essence of limited bandwidth in networks, and the WAIT-FREE model, capturing the essence of crash-prone asynchronous systems.

Fabrizio Grandoni (IDSIA)

Subcubic Equivalences between Graph Centrality Problems, APSP and Diameter

Measuring the importance of a node in a network is a major goal in the analysis of social networks, biological systems, transportation networks etc. Different centrality measures have been proposed to capture the notion of node importance. For example, the center of a graph is a node that minimizes the maximum distance to any other node (the latter distance is the radius of the graph). The median of a graph is a node that minimizes the sum of the distances to all other nodes. Informally, the betweenness centrality of a node w measures the fraction of shortest paths that have was an intermediate node. Finally, the reach centrality of a node w is the smallest distance r such that any s-t shortest path passing through $+$ has either s or t in the ball of radius r around w.

The fastest known algorithms to compute the center and the median of a graph, and to compute the betweenness or reach centrality even of a single node take roughly cubic time in the number n of nodes in the input graph. It is open whether these problems admit truly subcubic algorithms, i.e. algorithms with running time $\tilde{O}(n^{3-\delta})$ for some constant $\delta > 0$.

We relate the complexity of the mentioned centrality problems to two classical problems for which no truly subcubic algorithm is known, namely All Pairs Shortest Paths (APSP) and Diameter. It is easy to see that Diameter can be solved using an algorithm for APSP with a small overhead. However, no reduction is known in the other direction, and it is entirely possible that Diameter is a truly easier problem than APSP.

We show that Radius, Median and Betweenness Centrality are equivalent under subcubic reductions to APSP, i.e. that a truly subcubic algorithm for any of these problems implies a truly subcubic algorithm for all of them. We then show that Reach Centrality is equivalent to Diameter under subcubic reductions. The same holds for the problem of approximating Betweenness Centrality within any constant factor. Thus the latter two centrality problems could potentially be solved in truly subcubic time, even if APSP required essentially cubic time. Indeed, our reductions already imply an algorithm for Reach Centrality in graphs with small integer weights that is faster than the best known algorithm for APSP in the same family of graphs.

This is joint work with Amir Abboud and Virginia Vassilevska Williams.

Martin Grohe (RWTH Aachen University)

Dimension Reduction via Colour Refinement

Colour refinement is a basic algorithmic routine for graph isomorphism testing, appearing as a subroutine in almost all practical isomorphism solvers. It partitions the vertices of a graph into "colour classes" in such a way that all vertices in the same colour class have the same number of neighbours in every colour class. There is a tight correspondence between colour refinement and fractional isomorphisms of graphs, which are solutions to the LP relaxation of a natural ILP formulation of graph isomorphism.

We introduce a version of colour refinement for matrices and extend existing quasilinear algorithms for computing the colour classes. Then we generalise the correspondence between colour refinement and fractional automorphisms and develop a theory of fractional automorphisms and isomorphisms of matrices.

We apply our results to reduce the dimensions of systems of linear equations and linear programs. Specifically, we show that any given LP L can efficiently be transformed into a (potentially) smaller LP L' whose number of variables and constraints is the number of colour classes of the colour refinement algorithm, applied to a matrix associated with the LP. The transformation is such that we can easily (by a linear mapping) map both feasible and optimal solutions back and forth between the two LPs. We demonstrate empirically that colour refinement can indeed greatly reduce the cost of solving linear programs. This work grew out of applications in machine learning and probabilistic inference, where inference tasks are modelled by linear programs with the type of regularities that make our method very effective.

This is joint work with Kristian Kersting, Martin Mladenov, and Erkal Selman.

Giuseppe F. Italiano (University of Rome “Tor Vergata”)

Strong Bridges and Strong Articulation Points of Directed Graphs

Given a directed graph G, an edge is a strong bridge if its removal increases the number of strongly connected components of G. Similarly, a vertex is a strong articulation point if its removal increases the number of strongly connected components of G. Strong articulation points and strong bridges are related to the notion of 2-vertex and 2-edge connectivity of directed graphs, which surprisingly seems to have been overlooked in the past. In this talk, I will cover some very recent work in this area, both from the theoretical and the practical viewpoint.

Stefano Leonardi (University of Rome "Sapienza")

Efficient Computation of the Weighted Clustering Coefficient

The clustering coefficient of an unweighted network has been extensively used to quantify how tightly connected is the neighbor around a node and it has been widely adopted as an important measure for assessing the quality of nodes in a social network. The computation of the clustering coefficient is a challenging computational task that requires to count the number of triangles in the graph. Several recent works proposed efficient sampling, streaming and MapReduce algorithms that allow to overcome this computational bottleneck.

As a matter of fact, the intensity of the interaction between nodes, that is usually represented with weights on the edges of the graph, is also an important measure of the statistical cohesiveness of a network.

Recently various notions of weighted clustering coefficient have been proposed but so far all those techniques are hard to implement on large-scale graphs.

In this work we first show how standard sampling techniques can be used to obtain efficient estimator for the most commonly used measures of weighted clustering coefficient. Furthermore we also propose a novel graph-theoretic notion of clustering coefficient in weighted networks. Based on the observation that edges with large weights are more likely to play a role in the social network, we give an interpretation of the weight of an edge as the probability of existence of the edge in the graph. We therefore define the weighted clustering coefficient as the expected clustering coefficient on a family of random graphs. We show that our notion of weighted clustering coefficient can be computed in polynomial time and that can be efficiently approximated with sampling algorithms. We finally show experimentally interesting properties of the weighted clustering coefficient and we prove the accuracy and efficiency of our estimators.

Joint work with Silvio Lattanzi (Google Research NY).

Aleksander Mądry (EPFL)

Towards Algorithmic Graph Theory in Nearly-linear Time

In recent years, the emergence of massive computing tasks that arise in the context of web applications and networks has made the need for efficient graph algorithms more pressing than ever. In particular, it leads us to focus on reducing the running time of the algorithms to make them as fast as possible, even if it comes at a cost of reducing the quality of the returned solution.

In this talk, I will discuss the above theme and sketch two concrete scenarios in which obtaining such close to linear-time approximation algorithms was already possible. Along the way, I will briefly hint on the broader algorithmic toolkit that underlies these results.

Alberto Marchetti Spaccamela (University of Rome "Sapienza")

Feasibility Analysis in the Sporadic DAG Task Model

Real-time scheduling, a central problem in critical embedded systems, consists in determining the sequence of execution of tasks with deadlines. The design must ensure that the timing constraints imposed by the surrounding physical system can be guaranteed. The major differences between the problems studied in traditional scheduling theory and real-time

scheduling theory is the focus that real-time scheduling theory places on recurrent workloads. A real-time system is usually modeled as a finite collection of independent recurring tasks, each of which generates a potentially infinite sequence of jobs. Every job is characterized by an arrival time, an execution requirement, and a deadline, and it is required that a job complete execution between its arrival time and its deadline. Different formal models for recurring tasks place different restrictions on the values of the parameters of jobs generated by each task.

The DAG model has been proposed for representing recurrent precedence-constrained tasks to be executed on multiprocessor platforms, where each recurrent task is modeled by a directed acyclic graph (DAG), a period, and a relative deadline. Each vertex of the DAG represents a sequential job, while the edges of the DAG represent precedence constraints between these jobs. All the jobs of the DAG are released simultaneously and have to be completed within some specified relative deadline. The task may release jobs in this manner an unbounded number of times, with successive releases occurring at least the specified period apart. The feasibility problem is to determine whether given a set of recurrent tasks all jobs of all tasks can be scheduled to always meet all deadlines on a specified number of dedicated processors.

We show that the Earliest Deadline First algorithm (EDF) has a speedup bound of $2 - 1/m$, where m is the number of processors, while the Deadline Monotonic algorithm (DM) has a speedup bound of $3 - 1/m$. Moreover, we present polynomial and pseudopolynomial time tests, of differing effectiveness, for determining whether a set of sporadic DAG tasks can be scheduled by EDF or DM to meet all deadlines on a specified number of processors.

Joint work with V.Bonifaci, S.Stiller and A.Wiese.

Claire Mathieu (École Normale Supérieure)

The Glass Ceiling Effect in Social Networks

The glass ceiling may be defined as “the unseen, yet unbreakable barrier that keeps minorities and women from rising to the upper rungs of the corporate ladder, regardless of their qualifications or achievements'”. Although undesirable, it is well documented that many societies and organizations exhibit a glass ceiling.

In this paper we formally define and study the glass ceiling effect in social networks and provide a natural mathematical model that (partially) explains it. We propose a biased preferential attachment model that has two type of nodes, and is based on three well known social phenomena: i) rich get richer (preferential attachment) ii) minority of females (or other group) in the network and iii) homophily (preference to bond with similar people). We prove that our model exhibits a strong glass ceiling effect and that all three conditions are necessary, i.e., removing any one of them, will cause the model not to exhibit a glass ceiling effect.

Additionally we present empirical evidence of student-mentor networks of researchers that exhibits all the above properties: female minority, preferential attachment, homophily and a glass ceiling.

Joint work with Chen Avin, Barbara Keller, Zvi Lotker, David Peleg, and Yvonne-Anne Pignolet.

Kurt Mehlhorn (Max Planck Institute for Computer Science)

The Cost of Virtual Address Translation

Modern computers are not random access machines (RAMs). They have a memory hierarchy, multiple cores, and a virtual memory. We address the computational cost of the address translation in the virtual memory.

Starting point for our work on virtual memory is the observation that the analysis of some simple algorithms (random scan of an array, binary search, heapsort) in either the RAM model or the EM model (external memory model) does not correctly predict growth rates of actual running times.

We propose the VAT model (virtual address translation) to account for the cost of address translations and analyze the algorithms mentioned above and others in the model.

The predictions agree with the measurements.

We also analyze the VAT-cost of cache-oblivious algorithms.

This is algorithm engineering work on large data. The full paper is going to appear in Journal of Experimental Algorithmics.

Joint work with Tomasz Jurkiewicz.

Uli Meyer (Goethe University Frankfurt am Main)

On DFG Priority Programme "Algorithms for Big Data"

We will give an overview of the DFG Priority Programme "Algorithms for Big Data."

Friedhelm Meyer auf der Heide (University of Paderborn)

Distributed Data Streams in Dynamic Environments

Consider a scenario in which n nodes of a mobile ad-hoc network continuously collect data. Their task is to permanently update aggregated information about the data, for example the maximum, the sum, or the full information about all data received by all nodes at a given time step. This aggregated information has to be disseminated to all nodes.

Assume that a node can broadcast information proportional to a constant number of data items per round. We propose the following performance measures for distributed algorithms for this kind of tasks: The delay is the maximum time needed until the aggregated information about the data collected at some time has arrived at all nodes. Note that a too large communication volume needed for producing an output can lead to the effect that the delay grows unboundedly over time. Therefore, we have to cope with the restriction that not for all, but only for a fraction of rounds outputs are computed. We refer to this fraction as the output rate of the algorithm.

In this talk, we will propose several research problems and preliminary results related to this scenario, among them trade-offs between delay and output rate for aggregation problems under different kinds of dynamics in mobile ad-hoc network.

Harald Räcke (Technische Universität München)

Computing Cut-Based Hierarchical Decompositions in Almost Linear Time

We present a fast construction algorithm for the hierarchical tree decompositions that lie at the heart of oblivious routing strategies and that form the basis for approximation and online algorithms for various cut problems in graphs.

Given an undirected graph $G = (V, E, c)$ with edge capacities, we compute a single tree $T = (V_T, E_T, c_T)$, where the leaf nodes of T correspond to nodes in G, such that the tree approximates the cut-structure of G up to a factor of $O(\log^4 n)$. The best existing construction by Harrelson, Hildrum, and Rao[HHR03] just guarantees a polynomial running time but offers a better approximation guarantee of $O(\log^2 n \ \log\log n)$.

Phrasing our results in terms of vertex sparsifiers, we obtain the following result. For a graph $G = (V, E)$ with a subset S of terminals, we compute a tree T with at most $2|S|$ vertices (and the leafs of T correspond to nodes in S) such that T is a flow-sparsifier for S in G with quality $O(\log^2 n \log^2 k)$, where $|V| = n$ and $|S| = k$.

The running time of our algorithm is $O(\text{polylog}\, n \cdot T(m, 1/\log^3 n))$ where $T(m, \varepsilon)$ is the time for computing an approximate maxflow. The latter is almost linear due to the recent results of Sherman [She13] and Kelner et al. [KOLS13].

This is joint work with Hanjo Täubig and Chintan Shah.

Piotr Sankowski (University of Warsaw)

Power Law Complexity

The aimof this work is to give a plausible theoretical explanationwhy many algorithms run faster on real-world data then predicted by algorithmic worst-cases. In particular, we are able to show that on power law networks many graph problems have much lower algorithmic complexity then implied by classical/book solutions. The problems that we are able to solve faster on such networks include classical P-time problems like: transitive closure, maximum matching, eigenvalue problem, page-rank, shortest paths, counting triangles, etc. Additionally, we observe that some NP-hard problem allow faster exponential time algorithms, e.g., maximum clique. Finally, we prove that there exist structure oblivious algorithms that run faster on power law networks without explicit knowledge of this fact.

This is joint work with Marek Cygan, Paweł Brach and Jakub Łącki.

Christian Sohler (TU Dortmund)

Algorithmic Challenges in Statistics

Many modern statistical techniques do not sufficiently scale to use them with Big Data. Statistical models are typically formulated over the reals and have difficulties to deal with data sets that are inherently discrete. In my talk I will give a few examples for algorithmic challenges in statistics and problems in algorithms that are inspired by statistical data modelling and also discuss some first results.

Mikkel Thorup (University of Copenhagen)

Bottom-k and Priority Sampling, Set Similarity and Subset Sums with Minimal Independence

We consider bottom-k sampling for a set X, picking a sample $S_k(X)$ consisting of the k elements that are smallest according to a given hash function h. With this sample we can estimate the relative size $f = |Y|/|X|$ of any subset Y as $|S_k(X) \cap Y|/k$. A standard application is the estimation of the Jaccard similarity $f = |A \cap B|/|A \cup B|$ between sets A and B. Given the bottom-k samples from A and B, we construct the bottom-k sample of their union as $S_k(A \cup B) = S_k(S_k(A) \cup S_k(B))$, and then the similarity is estimated as $|S_k(A \cup B) \cap S_k(A) \cap S_k(B)|/k$.

We show here that even if the hash function is only 2-independent, the expected relative error is $O(1/\sqrt{fk})$. For $fk = \Omega(1)$ this is within a constant factor of the expected relative error with truly random hashing.

For comparison, consider the classic approach of k ×min-wise where we use k hash independent functions $h_1, \dots, h_k$, storing the smallest element with each hash function. For k ×min-wise there is an at least constant bias with constant independence, and it is not reduced with larger k. Recently Feigenblat et al. showed that bottom-k circumvents the bias if the hash function is 8-independent and k is sufficiently large. We get down to 2-independence for any k. Our result is based on a simply union bound, transferring generic concentration bounds for the hashing scheme to the bottom-k sample, e.g., getting stronger probability error bounds with higher independence.

For weighted sets, we consider priority sampling which adapts efficiently to the concrete input weights, e.g., benefiting strongly from heavy-tailed input. This time, the analysis is much more involved, but again we show that generic concentration bounds can be applied.

Monday

10:45 Welcome coffee + Welcome speech

11:15 **Mikkel Thorup** (University of Copenhagen)

Bottom-k and Priority Sampling, Set Similarity and Subset Sums with Minimal Independence

11:45 **Claire Mathieu** (École Normale Supérieure)

The Glass Ceiling Effect in Social Networks

12:15 Lunch (provided)

14:00 **Fabrizio Grandoni** (IDSIA)

Subcubic Equivalences between Graph Centrality Problems, APSP and Diameter

14:30 **Giuseppe F. Italiano** (University of Rome "Tor Vergata")

Strong Bridges and Strong Articulation Points of Directed Graphs

15:00 **Marek Zdanowski** (COST-Action representative for Poland)

COST-Action: Q&A

15:45 Coffee break

16:30 **Hannah Bast** (University of Freiburg)

Semantic Search on Big Data

17:00 **Graham Cormode** (University of Warwick)

Sketches, Streaming and Big Data

19:00 Dinner at **Dawne Smaki** (to be confirmed; not covered by the registration fee)

Address: Dawne Smaki, Nowy Świat 49, Warsaw

Tel: 22 465 83 20

URL: http://www.dawnesmaki.pl/en/

Tuesday

9:30 **Alberto Marchetti Spaccamela** (University of Rome “Sapienza”)

Feasibility Analysis in the Sporadic DAG Task Model

10:00 **Kurt Mehlhorn** (Max Planck Institute for Computer Science)

The Cost of Virtual Address Translation

10:30 Coffee break

11:15 **Stefano Leonardi** (University of Rome “Sapienza”)

Efficient Computation of the Weighted Clustering Coefficient

11:45 **Artur Czumaj** (University of Warwick)

Testing Cluster Structure of Graphs

12:15 Lunch (provided)

14:00 **Uli Meyer** (Goethe University Frankfurt am Main)

On DFG Priority Programme "Algorithms for Big Data"

14:30 Discussion about COST Action Proposal and follow up actions

15:30 Coffee break

16:15 **Friedhelm Meyer auf der Heide** (University of Paderborn)

Distributed Data Streams in Dynamic Environments

16:45 **Pierre Fraigniaud** (LIAFA)

Distributed Decision and Verification

19:00 Dinner at **Brasserie Warszawska** (to be confirmed; not covered by the registration fee)

Address: Brasserie Warszawska, ul. Górnośląska 24, Warsaw

Tel: 22 628 94 23

URL: http://www.brasseriewarszawska.pl/en_index.html

Wednesday

9:30 **Ilias Diakonikolas** (University of Edinburgh)

Learning Structured Distributions from a Constant Number of Samples

10:00 **Martin Grohe** (RWTH Aachen University)

Dimension Reduction via Colour Refinement

10:30 Coffee break

11:15 **Christian Sohler** (TU Dortmund)

Algorithmic Challenges in Statistics

11:45 Discussion on the objectives and future plans of the Interest Group on Algorithmic Foundations of Information Technology

12:30 Lunch (provided)

14:00 **Harald Räcke** (Technische Universität München)

Computing Cut-Based Hierarchical Decompositions in Almost Linear Time

14:30 **Aleksander Mądry** (EPFL)

Towards Algorithmic Graph Theory in Nearly-linear Time

15:00 Coffee break

15:30 **Piotr Sankowski** (University of Warsaw)

Power Law Complexity

16:00 Final discussion

16:30 End of the workshop