The German-Israeli Workshop/Winter School on Algorithms for Big Data

Tel Aviv University

November 13-15, 2017

Organized by
Guy Even and Ulrich Meyer

Let a thousand filters Bloom

Martin Farach-Colton, Rutgers University

Bloom filters and other approximate membership query data structures (AMQs), are one of the great successes of theoretical computer science, with uses throughout databases, file systems, networks and beyond.

An AMQ maintains a set under insertions, sometimes deletions, and queries with one-sided error: if a queried element is in the set, the AMQ returns `present`, if it is not, the AMQ returns `present` with probability at most ϵ. An optimal AMQ uses $n \log \epsilon^{-1}$ bits of space when the represented set has n elements.

Yet there is a gap between the theoretical bounds provided by AMQs and their requirements in the field.

In this talk, I will describe how AMQs are used in practice and how this changes (for the better) our theoretical understanding of these data structures.

Instance Optimality and Query Complexity: Definitions, Results and Conjectures

Moni Naor, Weizmann Institute of Science

Instance optimality is a measure of goodness of an algorithm in which the performance of one algorithm is compared to others per input. This is in sharp contrast to worst-case and average-case complexity measures, where the performance is compared either on the worst input or on an average one, respectively.

In this talk I will explore the study of instance optimality in the query model (a.k.a. the decision tree model). In this model, instance optimality of a function is formalized as follows: the complexity of an algorithm for the function is at most a constant factor larger than the complexity of an algorithm for the function that knows the given input, on every input separately. That is, we say that a function is instance optimizable if there is a decision tree for the function that performs on every input roughly as well as the best decision tree that is given the input as a certificate but is correct on every input. We will consider deterministic and randomized decision trees as well a model we call the anonymous certificate complexity where only information about the structure of the graph is known. Joint work with Ilan Komargodski.

Simple Distributed Graph Clustering using Modularity and Map Equation

Michael Hamann, KIT

We study large-scale, distributed graph clustering. Given an undirected, weighted graph, our objective is to partition the nodes into disjoint sets called clusters. Each cluster should contain many internal edges. Further, there should only be few edges between clusters. We study two established formalizations of this internally-dense-externally-sparse principle: modularity and map equation. We present two versions of a simple distributed algorithm to optimize both measures. They are based on Thrill, a distributed big data processing framework that implements an extended MapReduce model. The algorithms for the two measures, DSLM-Mod and DSLM-Map, differ only slightly. Adapting them for similar quality measures is easy. In an extensive experimental study, we demonstrate the excellent performance of our algorithms on real-world and synthetic graph clustering benchmark graphs.

Communication Efficient Algorithms

Lorenz Hbschle-Schneider, KIT

As the number of machines involved in distributed computations increases, communication rapidly becomes the bottleneck. To ensure scalability, algorithms with low communication volume and few messages are required. Oftentimes, these algorithms will be probabilistic in nature, enabling large improvements in running time. We are going to look at two examples in this talk: verifying correctness of a sum aggregation operation with very low overhead, and selecting the k most frequent objects from a large distributed input.

Scalable Kernelization for Maximum Independent Sets

Demian Hespe, KIT

The most efficient algorithms for finding maximum independent sets in both theory and practice use reduction rules to obtain a much smaller problem instance called a kernel. The kernel can then be solved quickly using exact or heuristic algorithms - or by repeatedly kernelizing recursively in the branch-and-reduce paradigm. It is of critical importance for these algorithms that kernelization is fast and returns a small kernel. Current algorithms are either slow but produce a small kernel, or fast and give a large kernel. We attempt to accomplish both of these goals simultaneously, by giving an efficient parallel kernelization algorithm based on graph partitioning and parallel bipartite maximum matching. We combine our parallelization techniques with two techniques to accelerate kernelization further: dependency checking that prunes reductions that cannot be applied, and reduction tracking that allows us to stop kernelization when reductions become less fruitful. Our algorithm produces kernels that are orders of magnitude smaller than the fastest kernelization methods, while having a similar execution time. Furthermore, our algorithm is able to compute kernels with size comparable to the smallest known kernels, but up to two orders of magnitude faster than previously possible. Finally, we show that our kernelization algorithm can be used to accelerate existing state-of-the-art heuristic algorithms, allowing us to find larger independent sets faster on large real-world networks and synthetic instances.

The Complexity of String Problems

Moshe Lewenstein, Bar-Ilan Univ

The time and space complexity of numerous core string problems has been stagnant for some time now. We explore one approach to explain this phenomena, via conditional lower bounds.

Selection from heaps, row-sorted matrices and X + Y using soft heaps

László Kozma, TU Eindhoven

We use soft heaps to obtain simpler optimal algorithms for selecting the k-th smallest item, and the set of k smallest items, from a heap-ordered tree, from a collection of sorted lists, and from X + Y, where X and Y are two unsorted sets. Our results match, and in some ways extend and improve, classical results of Frederickson (1993) and Frederickson and Johnson (1982).

Testing graph properties very efficiently

Artur Czumaj, Warwick Univ.

In this talk, we will survey recent advances on the problem of testing graph properties. We will consider a generic problem that for a given input graph G=(V,E) and a given graph property P (e.g., P may mean bipartiteness, 3-colorability, or planarity), we would like to determine if G satisfies property P or not. While the exact problem as defined here is often known to be computationally very hard (e.g., NP-hard, or even undecidable), we will focus on a simpler task, and we will want to distinguish between the input graphs that satisfy property P from the graphs that are far away from satisfying property P. Being far away means that one has to modify the input graph in more than, say, 1% of its representation to obtain a graph satisfying property P. We will survey recent results in this area and show that for many basic properties, one can test them in this framework very efficiently, often in sublinear-time, and sometimes even in constant time.

Testing for Forbidden Order Patterns in an Array

Ilan Newman, University of Haifa

We study testing of sequence' properties (e.g., real time series) that are defined by forbidden order patterns.

An Order pattern of length k is defined by a permutation $\pi = (\pi(1), \ldots, \pi(k))$ of k elements. A sequence $f : \{1, \ldots, n\} \to R$ of length n contains a pattern π if there is a subsequence of f in which the order of the elements in it is identical to the order defined by π. Namely, there are indices $i_1 < i_2 < \cdots < i_k$, such that $f(i_x) > f(i_y)$ whenever $\pi(x) > \pi(y)$. If f does not contain π, we say that f is π-free. For example, for $\pi = (2, 1)$, the property of being π-free is equivalent to being non-decreasing, i.e. monotone.

For a fixed order pattern π, of constant length k, and input sequence f stored in an array, we consider the property testing problem of distinguishing the case that f is π-free from the case that f differs in more than ϵn places from any π-free sequence. We show the following results:

- Being π-free can be tested in slightly sublinear number of queries for every fixed length π.

- There is a clear dichotomy between the monotone patterns and the non-monotone ones: For monotone patterns of length k, i.e., $(k, k-1, \ldots, 1)$ and $(1, 2, \ldots, k)$, we design non-adaptive, one-sided error ϵ-tests of $\mathrm{poly}(\log n)$ query complexity.

  For non-monotone patterns of size k we show that any non-adaptive one-sided error test requires at least $\Omega(\sqrt{n})$ queries. This general lower bound can be further strengthened for specific non-monotone k-length patterns to $\Omega(n^{1-2/(k+1)})$.

- We show that adaptivity can make a big difference in testing non-monotone patterns. We develop an adaptive algorithm that for any length 3 pattern π tests π-freeness by making $\mathrm{poly}(\log n)$ queries.

For all algorithms presented here, the running times are linear in their query complexity.

This is a joint work with Deepak Rajendraprasad, Christian Sohler and Yuri Rabinovich.

Lower Bounds under Bandwidth Limitations

Keren Censor-Hillel, Technion

I will describe known and new techniques for lower bounds for distributed graph computations in settings of limited bandwidth.

Blockchain Scalability

Roger Wattenhofer, ETH Zurich

Depending whether the audience is up to it, I will give a short introduction to Bitcoin, explaining some of the basics such as transactions and the blockchain. Then, I will discuss some interesting technical aspects in more detail. In particular, I will discuss how to improve scalability and throughput with duplex micropayment channels and other mechanisms. Apart from scalability, these channels also guarantee end-to-end security and instant transfers, laying the foundation of a network of payment service providers. Since all this is not a particularly good match for neither algorithms nor big data, I will begin with another result: How to compute a small balanced vertex separator of a graph in pretty much linear time.

Glass Ceiling and Power Inequality in Social Networks

David Peleg, Weizmann Institute

The talk will discuss the social effects of power inequality and glass ceiling in a society with two populations (e.g., men and women). We will introduce a model based on a bi-populated social network, define measures for power inequality and glass ceiling, and analyze the conditions for their occurrence in terms of three societal parameters, the relative size of the two populations, the level of homophily, and the extent of the ”leaky pipeline” phenomenon.

Dynamic Algorithms for Big Graphs Inspired by Distributed Computing

Leonid Barenboim, Open University of Israel

We consider dynamic graphs in the fully-dynamic centralized setting. In this setting the vertex set of size n of a graph G is fixed, and the edge set changes step-by-step, such that each step either adds or removes an edge. The goal in this setting is maintaining a solution to a certain problem (e.g., maximal matching, edge coloring) after each step, such that each step is executed efficiently. The running time of a step is called update-time. One can think of this setting as a dynamic network that is monitored by a central processor that is responsible for maintaining the solution. Currently, for several central problems, the best known deterministic algorithms for general graphs are the naive ones which have update-time $O(n)$. The existence of sublinear-in-n update-time deterministic algorithms for dense graphs and general graphs is an intriguing question. We address this question for maximal matching in the wide family of *graphs with* *bounded neighborhood independence.* This family includes unit-disc graphs, unit-ball graphs, line-graphs, graphs of bounded diversity, and many other graphs. We also show a solution for $O(\Delta)$-edge-coloring in sublinear time for *general graphs.*

In order to obtain our results we employ a novel approach that adapts certain distributed algorithms of the LOCAL setting to the centralized fully-dynamic setting. This is achieved by optimizing the *work* each processors performs, and efficiently simulating a distributed algorithm in a centralized setting. The simulation is efficient thanks to a careful selection of the network parts that the algorithm is invoked on, and by deducing the solution from the additional information that is present in the centralized setting, but not in the distributed one. We will present our experimental results of various network topologies and scenarios to demonstrate that our algorithms are highly-efficient in practice. Based on a joint work with Tzalik Maimon.

Locality

John Iacono, Université Libre de Bruxelles and New York University

Modern computers a strong preference for locality of preference whereby accessing data that is close to something that is recently accessed is much faster than accessing randomly located data. In this talk I will describe several data structures for fundamental problems whose design principle is to maximize locality of reference. These structures are usually analyzed in the cache-oblivious model, but I will also try to draw direct connections between these structures and locality, as well as between the cache-oblivious model and locality.

$2, 3, \ldots, k$: From approximating the number of edges to approximating the number of k-cliques (with a sublinear number of queries)

Dana Ron, Tel Aviv University

In this talk I will present an algorithms for approximating the number of k-cliques in a graph when given query access to the graph. This problem was previously studied for the cases of $k = 2$ (edges) and $k = 3$ (triangles). We give an algorithm that works for any $k \geq 3$ (and is actually conceptually simpler than that $k = 3$ algorithm).

We consider the standard query model for general graphs via (1) degree queries, (2) neighbor queries and (3) pair queries. Let n denote the number of vertices in the graph, m the number of edges, and C_k the number of k-cliques. We design an algorithm that outputs a $(1+\epsilon)$-approximation (with high probability) for C_k, whose expected query complexity and running time are $O\left(\frac{n}{C_k^{1/k}}+\frac{m^{k/2}}{C_k}\right)\text{poly}(\log n, 1/\epsilon, k)$.

Hence, the complexity of the algorithm is sublinear in the size of the graph for $C_k = \omega(m^{k/2-1})$. Furthermore, we prove a lower bound showing that the query complexity of our algorithm is essentially optimal (up to the dependence on $\log n$, $1/\epsilon$ and k).

This is joint work with Talya Eden and C. Seshadhri.

High Quality Hypergraph Partitioning

Sebastian Schlag, KIT

Hypergraphs are a generalization of graphs, where each edge can connect more than two vertices. Given an undirected hypergraph, the *k-way hypergraph partitioning problem* is to partition the vertex set into k disjoint blocks of bounded size, such that an objective function involving the cut edges is minimized. The problem has many important applications in practice such as scientific computing and VLSI design.

Since hypergraph partitioning is NP-hard and since it is even NP-hard to find good approximate solutions for graphs, heuristic *multilevel* algorithms are used in practice. These algorithms consist of three phases: In the *coarsening phase*, the hypergraph is coarsened to obtain a hierarchy of smaller hypergraphs that reflect the basic structure of the input. After applying an *initial partitioning* algorithm to the smallest hypergraph in the second phase, coarsening is undone and, at each level, a *local search* method is used to improve the partition induced by the coarser level.

In this talk, we briefly introduce the hypergraph partitioning problem along with the multilevel framework and present the partitioning framework KaHyPar (Karlsruhe Hypergraph Partitioning). It is the method of choice for a wide range of hypergraph partitioning tasks, computing better solutions than the widely used general purpose tools hMetis and PaToH.

Efficient Semantic Search on Big Data

Niklas Schnelle, Uni Freiburg

This talk takes a high-level look at the algorithms and principles behind our natural language question answering system Aqqu. Starting with the idea of storing knowledge as a graph of facts we will see how a machine can give meaningful answers to questions ranging from movies to geography.

Generation of Random Hyperbolic Graphs

Manuel Penschuck, Uni Frankfurt

Random graph models, originally conceived to study the structure of networks and the emergence of their properties, have become an indispensable tool for experimental algorithmics. Amongst them, hyperbolic random graphs form a well-accepted family. Based on a geometric embedding, they yield realistic complex networks while being both mathematically and algorithmically tractable. We present recent results and on-going research towards a fast, memory-efficient and scalable distributed sampling of massive hyperbolic random graphs.

Data-Aware Network Design: Some Results and Open Questions

Chen Avin, Ben-Gurion University of the Negev

We currently witness the emergence of interesting new network topologies optimized towards the data they need to serve, such as demand-aware datacenter interconnects (e.g., ProjecToR) and demand-aware peer-to-peer overlay networks (e.g., SplayNets). This talk will introduces a formal framework and approach to reason about and design such topologies.

In particular, we establish a connection between the communication request distribution and the expected path length in the network and show that this relationship depends on entropy measures of the communication matrix or the empirical entropy of the communication sequence.

We derive a general lower bounds and present asymptotically optimal network-aware design algorithms for important distribution families (such as sparse distributions and distributions of locally bounded doubling dimensions, among others.)

Scalable real-time semantic interpretation of text

Deepak Ajwani, Nokia Bell Labs

We are living in an era of information overload. Massive amount of content, in the form of social media, blogs, news, emails, research articles is continously vying for our attention. The limited ability of humans to absorb this content is fast becoming a bottleneck in the advancement of our species and there is an increasing need for automated tools to help us ingest information faster. Such automated tools should read the content on our behalf, understand their meaning, link with other relevant information and provide us a holistic summary in our context. A fundamental challenge in creating such tools is to understand the semantics of text in real-time. This talk will cover some fundamental primitives in the semantic analysis of text – named entity disambiguation, topic labeling and extending knowledge bases – and describe learning solutions for these primitives that support real-time interpretation and have scalable training.

Scalable, Transparent and Post-quantum Secure Computational Integrity, with applications to Crypto-currencies

Prof. Eli Ben-Sasson, Technion

Scalable Zero Knowledge (ZK) proofs are currently used to enhance privacy and fungibility in the Zcash cryptocurrency, and could potentially be used to solve Bitcoin's scalability problems.

This talk describes recent progress towards, and applications of, scalable and transparent zero knowledge proofs, whose setup requires only a public random string.

Joint work with Iddo Bentov, Ynon Horesh and Michael Riabzev.