Reg. No: 2020/39/B/ST6/03288; Principal Investigator: prof. dr hab. Tomasz Arkadiusz Jurdziński

Efficient distributed and parallel algorithms on big and dynamic data

Ubiquitos presence of IT tools produces big amount of data collected in data centers or distributed among nodes of large networks. Technological and physical constraints make it difficult or infeasible to collect all data in one place, to facilitate fast processing. Time constraints on the other hand require parallel processing of such big data, i.e., computational tasks need to be split into subtasks executed in parallel on many computational units. Although problems of this type have appeared for years, recent developments have significantly changed the main assumptions and leading technologies for big data processing (e.g., Map Reduce and Spark frameworks) as well as for communication protocols/technologies for exchange of data (e.g., Internet of Everything, 5G). The goal of the project is to build new efficient algorithmic techniques for contemporary computing and communication paradigms adjusted to transmissions and processing of such big data. It covers design of algorithms for specific key computational problems as well as discovering structural properties of the studied computational environments.

For processing of bid data we are going mainly to develop algorithms and possibly establish impossibility results in the Massive Parallel Computing (MPC) model and closely connected model of distributed algorithms called Congested Clique (CC). These models reflect computational environment in contemporary big data centers, known in industry as e.g. MapReduce, Spark or Hadoop. Although many (surprising) results have been obtained for the MPC and CC models recently, the techniques of speeding up algorithms are on early stage, especially for scenarios with dynamic data changes and/or under the assumption that memory of a single machine/node is (significantly) limited. Given our experience e.g. on design of solutions for connectivity and spanning trees in Congested Clique, we are going to develop efficient solutions for other significatn algorithmic problems such scenarios or show that such fast algorithms do not exist.

The basic model designed to build algorithmic solutions for local wireless communication is called Multiple Access Channel, or MAC. It is assumed that many participants use a single wireless channel to communicate and, because of interferences, a transmitted message can be received only if no other message is transmitted at the same time by other node. Although MAC and its algorithmic primitives are studied over the years (see e.g. backoff protocols), most of results do not take into account key aspects of contemporary wireless technologies, as dependencies between devices and non-standard interference patterns, limited resources of tiny sensors or lack of opportunity to recharge batteries. We are going to develop techniques for models incorporating these aspects, especially design algorithmic foundations for basic communication and message delivery problems, e.g., packet or link scheduling in the scenario of dynamically arriving data or design of combinatorial primitives for periodic schedules. Generalizations of MAC found applications in other areas of computer science and other disciplines, in particular in *group testing* schemes. Based on our recent experience in design of group testing algorithms we are going build new ones, adjusted to constraints of applications in contemporary medical testing

The above described aspects of collection and processing of big data share many techniques of distributed and parallel algorithms and computing. Models of large area/scale distributed networks form a natural extension of these models, where connections between nodes of a network might be described by arbitrary graphs. In particular, it is not guaranteed that each pair of nodes of a network is connected by a direct link. In our research we will also consider these extensions, with specific focus on applications in data collection and processing using sensors with limited capabilities and big amount of data generated over time which cannot be gathered at one machine.

Prospective results of our project will consist of new algorithmic solutions for explored problems as well as discovery of computational difficulty of specific problems and limitations of the explored models of parallel/distributed computing. We hope to build new faster algorithms, better adjusted for contemporary communication and computation infrastructure. Although we focus on basic research, the goal is to build solutions with prospective applications, thanks to careful adjustment of the model and efficiency measures to contemporary technology, its key features and limitations.