Reg. No: 2018/31/N/ST6/02310; Principal Investigator: mgr inż. Karol Krzysztof Gotfryd

Algorithmic aspects of processing and analyzing large datasets in distributed systems

The large-scale distributed information systems play currently an important role, still finding more applications in many areas, in which there exists a need for quick processing of huge amounts of data. The sources of such data are both complex decentralized monitoring systems, gathering and processing the results of continuously measured values of parameters of interests in observed environment, as well as dynamic networks consisting of interacting mobile devices, including smartphones or other wearables like smartwatches or fitness trackers. Complex distributed systems encompassing, among others, wireless networks or sensor networks comprising of small, resource constrained devices, find also their applications for detection of dangerous events, detection of motion, monitoring industrial infrastructure or in the systems implementing the concept of Internet of Things, like intelligent transport systems, “smart home” solutions or systems for remote health monitoring.

At the core of each of the aforementioned applications are gathering, exchanging and processing of huge amount of data, including performing various computations of different complexity. In light of still increasing requirements on efficiency and reliability of such systems, there is a need for designing effective methods, allowing some elementary computations (serving usually as key building blocks for more complex procedures) to be performed fast and precisely, on large volumes of distributed data, including data streams generated dynamically with high intensity.

In the research project we plan to consider such classes of problems related to selected topics from the area of decentralized processing of big datasets. They encompass the problems of data aggregation, estimation of their distribution, anomalies detection and alarming in wireless networks. The studied problems are one of the most significant in the field of theory of distributed computations. In case of a large amount of distributed data, the lightweight devices comprising the network are unable to store them in their limited memory. Thus, it is necessary to aggregate the data, calculating some statistics like average, median or the number of values satisfying certain conditions, relevant from the perspective of given applications. Computation of the exact results in decentralized distributed systems usually is a very complex problem. Fortunately, in many cases the knowledge of only some approximate values is sufficient (e.g. when using a system providing real-time traffic information, we are merely interested in obtaining some general information and the exact numeric values are irrelevant). Nevertheless, the precision of obtained results is an important question. For example, in a system for monitoring temperature or level of carbon monoxide, detection of values even slightly exceeding the threshold should immediately trigger an alarm, resulting in informing other devices about the observed anomaly as soon as possible (simultaneously minimizing the probability of rising false alarms).

In the planned project we aim at concentrating on algorithmic aspects of studied problems, looking for novel, fast and energy efficient protocols, adapted to still growing amount of processed data, including streamed data. We are also going to pay attention to the resistance of constructed algorithms to various disruptions inherent to wireless communication. The proposed methods will be the subject to detailed experimental analysis, involving a series of numerical simulations. The research objective will be also to formally prove correctness of analyzed algorithms as well as determine precision of returned results. We believe that this approach will allow to provide convincing arguments on efficiency and robustness of the protocols. In our theoretical research we will use various advanced analytical and mathematical techniques, including the methods from the theory of probability, graph theory, combinatorial analysis and statistics.