Reg. No: 2020/39/D/ST6/00419; Principal Investigator: dr inż. Krzysztof Szymon Turowski

Toward an Estimation of Information Content for Graph Structures

Krzysztof Turowski

These days graphs are widely used to describe many different structures: social networks, protein interactions, or even functional dependencies in the brain. These networks have often millions of nodes and contain various interactions. There arises a natural question concerning effective storage, access, and processing of such data.

Of course, in the case of real-world networks, we do not have any direct access to knowledge according to what the model it has evolved. But we still may use our broader knowledge e.g. biological, about the particular processes under investigation to infer what are the possible ranges of parameters which give the largest generation probability. For example, for protein interaction graphs there exists a wide agreement among scientists that the base mechanisms are gene duplication and mutation. These properties are well caught by the duplication random graph model, in which we build graph incrementally by choosing a vertex at random, adding its copy and then addition or removal of its edges according to some predefined rules. One of the aims of our research is:

(1) verification of these hypotheses with respect to different versions of duplication random graph model and comparison to other models often met in the literature.

The classic theory of information has introduced entropy as a natural measure of information content for probability distributions over an abstract set of objects. We may use this concept also to graphs and ask about the (labeled) graph entropy i.e. how much space is needed (on average) to store a labeled graph. We may ask also about the *structural* (unlabeled) graph entropy: how much space we can save if we do not need to store any IDs of vertices, but only relations between them. The knowledge of both parameters is useful for an assessment of proposed compression algorithms and it may be helpful for finding an optimal algorithm. Therefore, as the next topic of our research we single out

(2) estimation of the graph entropy and the structural graph entropy for duplication random graph models and fast compression algorithms with known good guarantees of approximation.

The question about graph structure is closely related to a search for symmetries and regularities in a graph. The first of these properties is captured by a notion of automorphisms group i.e. relabelings which leave all the connections between labels unchanged. The second property may be often expressed by such graph parameters as the degree distribution or the distributions of certain simple structures e.g. triangles. For example, if a graph has many nodes with exactly one and the same neighbor, then intuitively there should exist a compact description of such substructure. The starting point of these lines of research will be:

(3) research on the distribution of small structures in a random graph generated from duplication random graph models,

(4) analysis of the degree distribution for graphs generated from duplication random graph models,

(5) estimation of the distribution of symmetries for duplication random graph models.