STOCHASTIC NETWORK MODELS INSPIRED BY SZEMERÉDI'S REGULARITY LEMMA, WITH APPLICATIONS TO BIG DATA ANALYSIS AND COMPRESSION

Hannu Reittu, Technical Research Centre of Finland VTT, Finland, Hannu.Reittu@vtt.fi
Fülöp Bazsó, Wigner Research Centre for Physics, Budapest, Hungary, bazso@mail.kfki.hu
Ilkka Norros, VTT, Finland, ikka.noros@vtt.fi

Szemerédi's celebrated Regularity Lemma (SRL) is a cornerstone of modern graph theory and even beyond, like the proof of the famous Green-Tao theorem in number theory shows. Roughly speaking, SRL states that any large graph can be approximately substituted by a bounded number of pseudo-random bipartite graphs. Although SRL has numerous theoretical applications, there are almost no applications to real life networks. However, SRL indicates a structure that is very convenient and potentially useful for practice. In our suggested method we fix an integer k, the order of considered partitions of the node set, and try to find a partition that fits with a pseudo-random "regular structure". The objective is to minimize the overall description length of the data (MDL). We have created corresponding efficient algorithms and examined several real life networks, and in most cases the method works well and a reasonable regular structure can be found. The classes of nodes have usually some meaningful interpretation, like customers with characteristic behaviour etc. The algorithm scales well, since the structure can be found from a sample of the network, while the rest of the network is classified in linear time. This could open a way to use such a sorting method also for the analysis of a very big data in the form of matrices.