Mesoscopic Structure of Complex Networks

Ph. D. thesis booklet

Gergely Tibély

Supervisor: Dr. János Kertész

Budapest University of Technology and Economics

Department of Theoretical Physics

2011

Preliminaries

In the recent years, it became increasingly widespread to describe systems of very different origin, e.g. from social, biological or computer science contexts, in terms of networks. Analysing the pattern of connections for systems consisting of several units made advancement possible in problems such as modelling the spreading of epidemics, understanding the interaction of genes, or the description of human relations on the societal level. These investigations frequently applied models and methods originating from statistical physics.

The most well-known graph models, although their global structure are different, have no structure at mesoscopic levels (larger than 1 node, but smaller than the whole graph), neglecting statistical fluctuations. It is a very natural question to ask whether it is also true for real-world networks or they contain regions denser than their surroundings. Common sense suggests that these subgraphs - if exist - carry relevant information about the constituting nodes, e. g. common interest in social networks or common function in networks of cell biology.

In this light, it is not surprising that one of the most active fields within the research of complex networks is the search for such locally dense subgraphs (“community detection”). Of course, the problem is much older than the activity of the last 10 years (roughly) -- even the ancient Greeks clustered objects into classes, but biology, sociology or computer science also know similar problems. Even statistical physics has preliminaries in this field, due to the superparamagnetic clustering method. Correspondingly, evolution of the current methods started from the existing methods of other fields, and from the standard toolbox of statistical physics (random walks, Potts model).

Aside of the results of the previous years, there are still important problems. First of all, no generally accepted precise definition of ``locally dense subgraph'' has appeared yet. Consequently, several different proposals were made, raising the second question: testing the proposed methods. Although a lot of improvements were made in this question recently, we are still far from a “case closed” state. Decreasing the running time of the methods is another constant problem.

Aims

In the recent years, the emphasis of community detection was on developing new methods, which resulted in a proliferation of community detection methods. In parallel, in several cases our knowledge about these new methods barely increased, especially regarding their behaviour on real-world networks. Accordingly, one aim of the dissertation was to analyse existing methods, especially their characteristics on real networks.

Although several methods were developed for detecting dense subgraphs, a generally accepted precise definition of dense subgraphs is still lacking. On the one hand, most of the methods contain a technical definition, around which the method is built -- these are precise, but more or less ad hoc. On the other hand, there is the generally accepted but far from precise phrase “nodes having more edges among themselves than to the rest of the graph”. In this situations, the most important aim of the dissertation is to get closer to a precise and well-established definition, and to demonstrate the results in a new method.

Finally, the practical application of the existing knowledge about community detection was also an aim.

New scientific results

1. I showed that the label propagation method is equivalent to finding the local minima of energy of a simple zero-temperature kinetic Potts model. I found that 1) on the tested empirical networks, the number of local minima is unexpectedly large, 2) the algorithm is highly unstable, different runs resulting in different local minima [2].

2. For the matrix of the absolute values of correlation coefficients of stock returns, which is equivalent to a weighted fully connected graph, I found that contrary to previous suggestions in the literature, the high ranking eigenvectors are inappropriate for the identification of clusters, at least without further a priori information, because the eigenvectors are frequently localised on combinations of clusters instead of single clusters. Using diffusion-related matrices instead of the correlation matrix does not

enhance the situation.

I also found that the deviations of the first eigenvector's components from the first-order perturbation theory are strongly correlated with the square of the node strengths. Comparing it to a proper 0-model -- an ensemble of matrices lacking cluster structure but otherwise having similar features -- I found that this feature is absent in the 0-model, consequently it can be attributed to the presence of cluster structure. Checking the presence of clusters in correlation matrices can be done quicker in this way than by analysing the whole spectrum [1].

3. I defined a measure, tiling imperfection, for quantifying hierarchical relationship of partitions and covers. I found that on a multimillion-node, phone call data-based network, the results of different community detection methods -- Infomap, Clique Percolation and modularity optimisation by the Louvain method -- have low tiling imperfections, in other words, they are quite close to each other, up to subset-superset relations [3].

4. I pointed out that the definition of communities should include not only the property characterising the subgraph's isolation from its surroundings, which I termed separation and was considered previously, but also cohesion -- resistance against splitting into two --, thus I proposed a 2-dimensional system of requirements instead of the former 1-dimensional criterion. I explicitly declared the required properties of communities.
   I created a simple test for the two criterions. I proposed a measure for quantifying the cohesion of a subgraph, based on the second eigenvalue of the corresponding Laplacian matrix of the subgraph. I introduced a new community detection method, taking into account both separation and cohesion. The method is capable of handling strongly overlapping communities and multi-level structures. Tests were conducted on the state-of-the-art benchmark (LFR), as well as on some real-world networks [4].

5. For optimising telecommunication network planning, I elaborated a method, which partitions the network into clusters, assigns a hub-node to each

cluster and routes all traffic via the corresponding hubs. Under favourable conditions, the costs of the test networks were reduced by 5-18\%, using the current network planning tool of Nokia Siemens Networks. Redundant links, and not very compact networks are required for success.

Scientific publications in connection with the thesis points

1. T. Heimo, G. Tibély, J. Saramäki, K. Kaski, J. Kertész,
   Spectral methods and cluster structure in correlation-based networks,
   Physica A **387**, 5930 (2008).

2. G. Tibély, J. Kertész,
   On the equivalence of the label propagation method of community detection and a Potts model approach,
   Physica A **387**, 4982 (2008).

3. G. Tibély, M. Karsai, L. Kovanen, K. Kaski, J. Kertész, J. Saramäki,
   Communities and beyond: mesoscopic analysis of a large social network with complementary methods,
   Phys. Rev. E **83**, 056125 (2011).

4. G. Tibély,
   Criterions for locally dense subgraphs,
   arXiv:1103.3397, elfogadva (*Physica A*).

Further scientific publications

5. D. Nagy, G. Tibély, J. Kertész,
   The effect of disorder on the hierarchical modularity in complex systems,
   Fractals **14**, 101 (2006).

6. G. Tibély, J.-P. Onnela, J. Saramäki, K. Kaski, J. Kertész,
   Spectrum, intensity and coherence in weighted networks of a financial market,
   Physica A **370**, 145 (2006).