Vladimir Batagelj

Large networks

1 Individual Project's contribution to the CRP

1.1 Aims and Objectives

In the last 15 years the large networks became one of the central fields of research and a basis for development of several other fields. Our objective is to develop efficient algorithms for analysis of large networks with the corresponding theoretical background, and software which can deal with very large networks (with some billions (10^9) of vertices and lines). In this project we shall concentrate on the following topics:

Analysis of the structure of large networks and algorithms for very large networks. In large networks analysis a fundamental task is discovery of global structure, identification of important elements and parts of network, and analysis of the position of selected element or group of elements in the network. In previous years we developed some efficient algorithms (cores, generalized cores, islands [4, 5, 30]) for these tasks. We will improve existing algorithms for network analysis and develop new ones that function well on very large networks. Selected new algorithms will be included in program Pajek [6, 24].

We also intend to adapt (development of a library) several existing and newly developed algorithms for 64-bit machines which support very large computer memories (several tens of Giga bytes) thus enabling analysis of very large networks.

Another direction of work is the extension of clustering algorithms with relational constraints (Ferligoj, Batagelj, 1983) [11, 12] to large networks. The basic idea is to compute the dissimilarities only between the linked units (vertices). This should allow us to analyze networks that were previously beyond our reach.

An approach to get insight into the structure of large network is also to reduce it to its *skeleton* by removing less important lines and/or vertices. Two such methods preserving the connectivity for weighted networks are the minimum spanning tree and Pathfinder algorithms. Determining a minimum spanning tree is a standard task. There exists an $O(m + n \log n)$ algorithm for determining MST (Cormen et al., 2001: 561-579[9]).

The Pathfinder algorithm was proposed in eighties (Schvaneveldt, 1990) [28] for simplification of weighted networks. It removes from the network all lines that do not satisfy the triangle inequality – if for a line a shorter path exists connecting its endpoints then the line is removed. The original matrix based Pathfinder algorithm has the complexity $O(qn^3)$, q is the neighborhood size parameter. The first improvement based on fast power computation was proposed by Guerrero-Bote et al. (2006)[15] and reduced complexity to $O(n^3 \log q)$. Additional improvements for some special cases were given by Quirin et al. (2008)[26]. For sparse networks in general case there is still some space for improvements – for each vertex we have to compute only the values of shortest paths to all its neighbors. We shall develop an efficient algorithm along these lines.

There are also some network structure related problems we intend to investigate, but we currently still don't know how to approach them:

- Decomposition of acyclic networks to 'basins'.
- Blockmodeling in large networks – maybe by using iterative colorings.
- Efficient algorithms were developed also for *modular decomposition* (Habib and Paul, 2009; Papadopoulos and Voglis, 2006)[16, 25] and *split decomposition* (Joeris et al., 2009[19]). Can these decompositions be used in large network analysis?

XYZ decompositions. The research on this topic began at Dagstuhl Seminar No. 08191 in 2008 [7]. The motivation was an atempt to provide a general framework to network vizualization approaches based on combination of link and matrix representations [1, 17, 18].

Given two graph classes $\mathcal{X}$ and $\mathcal{Y}$, a graph $G = (V, E)$ is an *$\mathcal{X}$-graph of $\mathcal{Y}$-graphs* (or *$(\mathcal{X}, \mathcal{Y})$-graph*, for short) if a family $V_1, V_2, \ldots, V_h$ of disjoint subsets of V, called *clusters*, can be identified, such that:

1. every cluster induces a graph belonging to class $\mathcal{Y}$, and
2. the *reduced graph* G^* obtained from G by collapsing each cluster into a single vertex and replacing multiple edges with a single one is a graph of class $\mathcal{X}$.

If subset $V_1, V_2, \ldots V_h$ are requested to be a partition of V, that is, if we add the constraint that $V = V_1 \cup V_2 \cup \cdots \cup V_h$, then we call G a *strong* $(\mathcal{X}, \mathcal{Y})$-graph, otherwise we call G a *weak* $(\mathcal{X}, \mathcal{Y})$-graph or, simply, an *$(\mathcal{X}, \mathcal{Y})$-graph.* The strong model of $\mathcal{X}$-graph of $\mathcal{Y}$-graphs, also known as *two level clustered graphs*, was introduced in [8].

In generalized blockmodeling [10] also the structure of bipartite subgraphs induced by the sets (V_i, V_j) is important. This adds the Z term to the decompositions with the additional requirement:

3. every pair of clusters induces a bipartite graph belonging to class $\mathcal{Z}$.

Most of blockmodeling problems seem NP-hard [13, 27].

Regarding the weak model, we explored some cases where dense $\mathcal{Y}$-graph are involved. In particular, the identification of cliques in large graphs may lead to a very effective strategy for information visualization. In fact, when we know that a subset of vertices is a clique, its internal edges are understood and do not need to be explicitly displayed. Unfortunately, recognizing $(planar, K_5)$-graphs is NP-hard. This result parallels the analogous result for the strong model [20].

In real life problems we are often searching for almost XYZ-decompositions – by formulating the corresponding optimization problems and developing methods for their solution. We intend to continue our research on this topic by considering other classes of graphs.

Evolution of networks. Nowadays, communities of researchers have been studied from many different angles for many years (e.g., Freeman, Barabasi, Newman [2, 14, 22, 23]). This subject is very attractive to researchers, as it is relatively simple to obtain data for the analysis [3]. Furthermore the collaboration of researchers can be improved based on the understanding of the underlying patterns and the relations in scientific communities.

However, most of the analysis focus on a single type of network (e.g., only on co-authorship or only on citation networks). In order to get a complete picture of a research community, all information contained in these different networks should be combined.

We will work on detailed development of methodological and algorithmic support of the approach to the analysis of 'Dynamics and Evolution of Research Communities' proposed at 08319 Dagstuhl seminar [29]. This approach is based on Web 2.0 and combines three levels:

- description: tags, properties;
- homophily networks (McPherson, Smith-Lovin, Cook, 2001) [21]: induced by the similarities by descriptions;
- social networks (based on homophily networks) and (explicit or implicit) groups (communities) in them.

All these levels are analyzed through time. A typical example of data are data about selected discipline from Web of Science. The approach should provide answers to the questions such as: what are the main topics in the disciplines, what are the main research groups, their leaders, are these groups growing, changing topic, what are the trends, groups dynamics (decay, growth, stability; merging, splitting). We will try to develop special algorithms for different problems in the elaboration / operationalization of the scheme, such as: identification of the groups, tracing the groups through time, identification of the kernel (leaders) of the group, measurement of the stability of groups, prediction of evolution of a group, etc.

1.2 Methodologies

We have studied similar problems in our previous projects. Therefore we know the mentioned areas well. The research will be done in two main steps:

1. Development and improvements of algorithms, study of their properties, implementations, development and refinement of the Dagstuhl schemes.
2. Testing of algorithms and applications in analysis of real data sets, preparation of presentations and papers.

References

[1] Abello J., van Ham F., Krishnan N.: ASK-GraphView : A Large Scale Graph Visualization System. IEEE Transactions on Visualization and Computer Graphics, Vol. 12, No. 5, September/October (2006).

[2] Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientic collaborations. Physica **311** (2002) 590–614.

[3] Batagelj, V.: Wos2pajek – networks from Web of Science (2007).

[4] Batagelj, V., Zaveršnik, M.: An $O(m)$ algorithm for cores decomposition of networks. CoRR, cs.DS/0310049, 2003.

[5] Batagelj, V., Zaveršnik, M.: *Generalized Cores*, 2002. (arXiv: cs.DS/0202039).

[6] Batagelj, V., Mrvar, A.: Pajek - Analysis and visualization of large networks. Graph drawing, 2265 (2002), 477-478.

[7] Batagelj, V., Brandenburg, F.J., Didimo, W., Liotta, G., Patrignani, M.: X-graphs of Y-graphs and their representations. In S. P. Borgatti, S. Kobourov, O. Kohlbacher, and P. Mutzel, editors, Graph Drawing with Applications to Bioinformatics and Social Sciences (Dagstuhl Seminar 08191), Dagstuhl Seminar Proceedings, 2008. (Working Group Report). Dagstuhl, May 4-9, 2008.

[8] Brandenburg, F.J.: Graph clustering I: Cycles of cliques. In Di Battista, G., ed.: Graph Drawing (Proc. GD '97). Volume 1353 of Lecture Notes Comput. Sci., Springer-Verlag (1997) 158–168.

[9] Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C. (2001) *Introduction to algorithms.* Cambridge (Mass.): MIT Press.

[10] Doreian P., Batagelj V., Ferligoj A.: Generalized Blockmodeling, Cambridge University Press, (2005).

[11] Ferligoj, A., Batagelj, V.: Clustering with relational constraint. Psychometrika, 47 (1982) 4, 413-426.

[12] Ferligoj, A., Batagelj, V.: Some types of clustering with relational constraints. Psychometrika, 48 (1983) 4, 541-552.

[13] Fialaa J., Paulusmaa D.: A complete complexity classification of the role assignment problem. Theoretical Computer Science 349(2005) 1, pp. 67-81.

[14] Freeman, L.C.: The impact of computer based communication of the social structure of an emerging scientific speciality. Social Networks **3** (1984) 201–221.

[15] Guerrero-Bote, V.P., Zapico-Alonso, F., Espinosa-Calvo, M.E., Crisóstomo, R.G. and de Moya-Anegón, F. (2006) *Binary Pathfinder: An improvement to the Pathfinder algorithm.* Information Processing and Management, 42(6): 1484–1490.

[16] Habib, M. and Paul, C. (2009) *A survey on algorithmic aspects of modular decomposition.* arXiv: cs.DM/0912.1457v2, 8 Dec 2009.

[17] Henry, N., Fekete, J.D., McGuffin, M.J.: NodeTrix: A hybrid visualization of social networks. IEEE Trans. Visual. and Comp. Graphics **13**(6) (2007) 1302–1309.

[18] Henry, N.: Exploring Large Social Networks with Matrix-Based Representations. Ph.D. Thesis, Cotutelle Universit Paris-Sud (France) and University of Sydney (Australia), July 2008.

[19] Joeris, B.L., Lundberg, S. and McConnell, R.M. (2009) *O(m log n) split decomposition of strongly-connected graphs.* Proceedings *Graph Theory, Computational Intelligence and Thought*, Haifa, September 2008, LNCS 5420, Berlin: Springer, pp. 158–171.

[20] Kratochvìl, J.: String graphs II: Recognizing string graphs is NP-hard. J. of Combinatorial Theory, Series B **52** (1991) 67–78.

[21] Mcpherson, M., Lovin, L.S., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology **27** (2001), 415–444.

[22] Newman, M.E.: The structure of scientific collaboration communities. Proceedings of the National Academy of Science (PNAS) **98** (2001) 404–409.

[23] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69** (2004).

[24] Pajek's Wiki: http://pajek.imfm.si

[25] Papadopoulos, C. and Voglis, C. (2006) *Drawing graphs using modular decomposition.* LNCS (GD'05) 3842, Berlin: Springer, pp. 343-354.

[26] Quirin, A., Cordón, O., Santamaria, J., Vargas-Quesada, B. and Moya-Anegón, F. (2008a) *A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time.* Information Processing and Management: an International Journal archive, 44(4): 1611–1623.

[27] Roberts F.S., Sheng L.: How hard is it to determine if a graph has a 2-role assignment? Networks 37 (2001) 2, pp. 6773.

[28] Schvaneveldt, R.W. (Ed.) (1990) *Pathfinder Associative Networks: Studies in Knowledge Organization.* Norwood, NJ: Ablex.

[29] Stumme, G., Müller, C., Batagelj, V., Hoser, B., Staab, S.: Dynamics and Evolution of Research Communities. Report from 08319 Dagstuhl seminar, September 2008, http://kathrin.dagstuhl.de/08391/Materials2/

[30] Zaveršnik, M., Batagelj, V.: Islands. In: XXIV International Sunbelt Social Network Conference, Portorož, Slovenia (2004).