Annales Univ. Sci. Budapest., Sect. Comp. 42 (2014) 37–54

CONJECTURES ON PHASE TRANSITION AT CORRELATION CLUSTERING OF RANDOM GRAPHS

László Aszalós (Debrecen, Hungary)

János Kormos (Debrecen, Hungary)

Dávid Nagy (Debrecen, Hungary)

Dedicated to Professor András Benczúr on the occasion of his 70th birthday

Communicated by János Demetrovics

(Received June 1, 2014; accepted July 1, 2014)

Abstract. The Correlation Clustering is a classical, NP hard optimization problem with many social, economic, physical, biological and computer science applications. We had implemented several methods to find near optimal solutions for particular problems. Here we summarize the results of our experiments on random graphs in particular with regard to phase transitions.

Key words and phrases: Phase transition, correlation clustering, random graphs
The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

1. Introduction

The aim of clustering is to discover the structure of objects and group them based on similarity, without any previous information about their structure. We would like similar objects to get into the same clusters, and different object to get into different ones. This condition is very general, so we cannot wonder how many different clustering methods exist. Most of the classification methods are based on some distance functions. Based on this distance we can say that two objects are similar or dissimilar. Correlation clustering [5] is different. It uses a similarity relation, so two objects are similar if this relation holds, and dissimilar if does not. This kind of clustering is sleekly modeling physical processes [25], biological relations [26] or social coalitions [29].

Sometimes it is interesting to cluster a given structure [3]. However, other times we have no exact knowledge about the stucture, just about some of its parameters. Néda at al. [25] had shown that in the case of a complete signed graph there is a phase transition: the function $r(q)$, which denotes the relative size of the maximal cluster, has a transition at $q = 1/2$.

Clustering of random graphs is an especially engaging problem [12]. Unfortunately the clustering is so complicated, that we cannot give a mathematically strict analysis, just Monte Carlo simulations. Néda at al. [24] made numerical tests on some special Erdős-Rényi graphs as well as Barabasi-Albert type scale-free graphs. Our investigations were inspired by that article. We implemented several well-known algorithms along with some new ones [2], to replace the tools used at simulations in [24]. This, and a new storage method of graphs enables us to raise the number of nodes from $100 - 150$ to 500 and generate one curve of the graphs within an hour on an ordinary desktop computer, which needs thousands of clusterings. We tested these simulations by random graphs which can be written with a few parameters.

András Benczúr almost 40 years ago drew attention to the special problems and hardness in handling of the large systems as well as data (database) and program level. He gave good solutions concerning this ’big data’ world in real practice [8, 22], and he got very nice theoretical results [1, 9, 10], too. Another field of his interest is to understand the deep meaning of information, the essence of *information boom* of information society [11, 7, 10]. In this paper we give a small contribution to the investigation of large networks (graphs) from side of algorithms. Our experiments fully support Benczúr's approach.

1.1. Random networks

In this point we give a short summary of questions and answers from theory of random graphs concerning to our main topic. This is based mostly on the book of L. Lovász published in 2012 [20]. Random Networks has no other source which would provide a better explanatory overview than Chapter 1 in Lovász's book. The meaning might have been modified by rephrasing the author's sentences, therefore we quoted pieces verbatim from that chapter.

That is the fact that a large number of the most interesting structures and phenomena of the world can be described by networks. Some examples: the Internet, the network of hyperlinks; the acquaintance graph of all living people with about 7 billion nodes; the human brain, a network of neurons having about a hundred billion nodes. One can say that the whole universe is a single network, where the nodes are events (interactions between elementary particles), and the edges are the particles themselves. This is a network with perhaps 10^{80} nodes.

Very large networks are never completely known, in most cases they are not even well defined. Data about them can be collected only by indirect means like random local sampling or by monitoring the behavior of various global processes.

The most important and widely investigated questions: What is the average degree of nodes? Is the graph connected? Where is the largest cut in the graph? How to classify the nodes (vertices)? One of the crucial questions is how to observe graph processes? Poperties of very large graphs can be studied by randomly sampling small subgraphs. It turns out that this sample contains enough information to determine many properties and parameters of the graph, with some error of course.

There were introduced some different models. The simplest random graph model was developed by Erdős–Rényi [16] and Gilbert [19]. One can generate a random graph by taking the nodes, and connecting any two of them with a given probability making an independent decision about each pair of nodes. There are alternate models, which are essentially equivalent from the point of view of many properties. Two of these were introduced in the early papers by Erdős–Rényi [16, 17]. Another model, closer to some of the more recent developments, is evolving random graphs, where edges are added one by one, always choosing uniformly from the set of unconnected pairs. Random graphs have many interesting, often surprising properties, and a huge literature, see Bollobas [13].

Random graph models on a fixed set of nodes, discussed above, fail to reproduce important properties of real-life networks. In 1999 Albert and Barabasi [6] created a new random network model. Perhaps the main new feature compared with the Erdős–Rényi graph evolution model is that not only edges, but also nodes are added by natural rules of growing. The Albert–Barabási graphs reproduce the *heavy tail* behavior of the degree sequences of real-life graphs. In this paper our investigation is related to these classical models introduced above. Since then a great variety of growing networks were introduced, reproducing this and other empirical properties of real-life networks among them Móri [23], Fazekas and Porvázsnyik [18].

One of the most important questions is how to assign limits to sequences of graphs? The growing graph sequences tend to have a well-defined structure, for almost all of the possible random choices along the way. In the limit, the randomness disappears (similarly to the law of large numbers), and the asymptotic behaviour of the sequence can be described by a well-defined limit object. We have to mention again a Hungarian mathematician whose contribution is fundamental in this field. Namely Szemerédi and his regularity lemma from 1975 [27].

Many authors considered a growing sequence of graphs whose number of nodes tends to infinity, to define when such a sequence is convergent, and to assign a limit object to convergent graph sequences, which somehow incorporates all the properties we want to be remembered. For dense graphs, this notion of convergence was defined by Borgs, Chayes, Lovász, Soós and Vesztergombi [14]. More explicit descriptions of these limit objects can also be given, in the form of a two-variable measurable function, called a graphon (Lovász and Szegedy [21]).

There are several related questions here, depending on what we need as a result. The easiest setup is when we want to compute a numerical parameter of the graph; say, how large is the maximum cut, or what fraction of the triples induce a triangle, or we want to find a perfect matching in the graph, or a maximum cut, or a regularity partition in a huge dense graph. To handle these questions we must define similarity distance between two nodes of a graph. We could try considering two nodes similar, if their neighbourhoods differ by little.

1.2. Correlation clustering

At correlation clustering mathematically speaking we have a graph $G = (V, E)$, where V is the set of object we would like to cluster. And we have signed edges, more precisely a function $s : E \to \{+, -\}$ which assigns a sign for each edge. Here the sign $+$ denotes the similarity and $-$ denotes the dissimilarity. We refer to the signed graph as (G, s) in the following. Naturally we can use two colours instead of signs, but the signed graph is the traditional terminology according to Bansal at al. [5]. The sign of edges here can arise from any source, e.g. from similarity distance or by real distance.

The correlation clustering minimizes the disagreements: the number of pairs of dissimilar objects within clusters plus the number of pairs of similar objects in different clusters. If $p : V \to \mathbb{N}$ is a partition of the object, then we can assign a cost value to this partition as

$$f^s_G(p) := \Big| \{(i,j) \in E \,|\, s(i,j) = + \text{ iff } p(i) = p(j)\} \Big|.$$

The fact, that the goodness of a clustering is measured by a number, enables us to compare the different clusterings/partitions. This comparability of clusterings is not common, at other methods thumb rules to help the users to choose the right parameters: to get a good clustering. This is not the case here. The comparison enables us to choose the best one. The result of a correlation clustering of some signed graph (G, s) is a partition $p^\star$ where $f^s_G(p^\star)$ is minimal, i.e. $f^s_G(p^\star) \leq f^s_G(p)$ for each partition $p : V \to \mathbb{N}$.

When the number of the objects is less than 15, a full search gives for us the global optimum. In special cases (for special graphs) we can get the exact optimum for more nodes, too; but in general case not.

Unfortunately the number of partitions is an exponential function of the number of objects, so at practical cases we can only approximate the optimal partition. The authors and their students implemented several combinatorial optimization methods [2, 4]. These methods were tested for correlation clustering, and a fast and effective method had been chosen to run experiment for this article. This method named *Contraction* is a simple greedy algorithm. It starts with singletons as a partition and iteratively selects the pair of clusters worth to join. According to the greedy method it selects the pair where the cut in value $f_G^s(p)$ is maximal. It stops when there is no pair of clusters worth to join.

2. Theoretical and experimental results for complete graph

For a given set of object V in case of complete graphs the set of edges E is uniquely defined. But in case of signed complete graphs, many $s : E \to \{+, -\}$ *colouring* functions exist. To be able to compare these colourings, we define the rate q_s of positive edges as follows:

$$q_s = \frac{|\{e \in E|s(e) = +\}|}{|E|}.$$

It is obvious that only one colouring s' exists such that $q_{s'} = 0$. In this case all the edges are negative, and it is easy to check, that for the partition p' which contains only singletons $f_G^{s'}(p') = 0$. Similarly, there uniquely exists a colouring s'' such that $q_{s''} = 1$. In this case for the partition $p'' = \{V\}$ holds that $f_G^{s''}(p'') = 0$. In other cases we examined statistically colourings s, which have the same rate q_s: we have chosen several samples, and applied the correlation clustering for the signed graph (G, s).

As Erdős and Rényi examined the connectivity of the whole graph for random graphs [16], it is evident that it is needed to examine the size of the maximal cluster of the solution of the correlation clustering on complete graphs. As we noted in the introduction, Néda at al. had shown that the correlation clustering has two district phases separated by 1/2 [25]. In the asymptotic case when $q \in [0, 1/2)$, the relative size of the

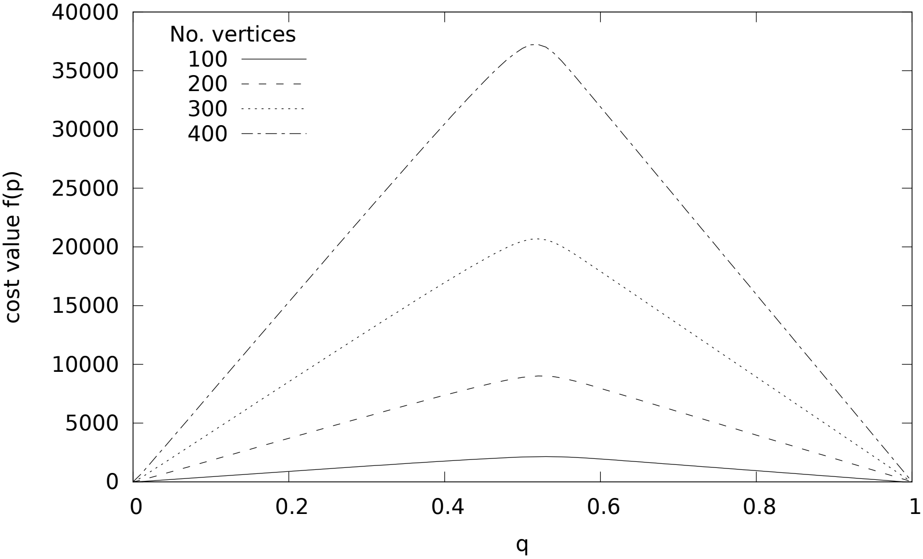

Figure 1. As we have more edges, the value of the cost function is bigger.

maximal cluster is 0, and when $q \in (1/2, 1]$ then the maximal cluster contains all the vertices. Fig. 2 and 3 justify that our experimental results are consistent with this theoretical result, i.e. as n grows we get closer to the asymptotic limit.

To show some half-hidden properties we present also the average of the values of $f_G^s(p)$ as the function of q_s, see Fig. 1. As we can have many different colourings with the same rate, we only plot the mean. In this case, the deviation is not remarkable. The figure contains several curves for complete graphs with different sizes. Note, that the symmetry of these functions, and the turning point is around 1/2 by the figure. If we examine the data carefully, in these cases this value is around 0.53 and not 0.5. It is an open question *whether this difference is the effect/error of the optimization algorithm we used, or it has other reason?*

Fig. 3 shows the most interesting part of the curve of the averages of the maximal cluster sizes which can be found on Fig. 2. Here we denote relative sizes of the maximal clusters according to the number of vertices of the complete graphs for several cases, to show the tendencies.

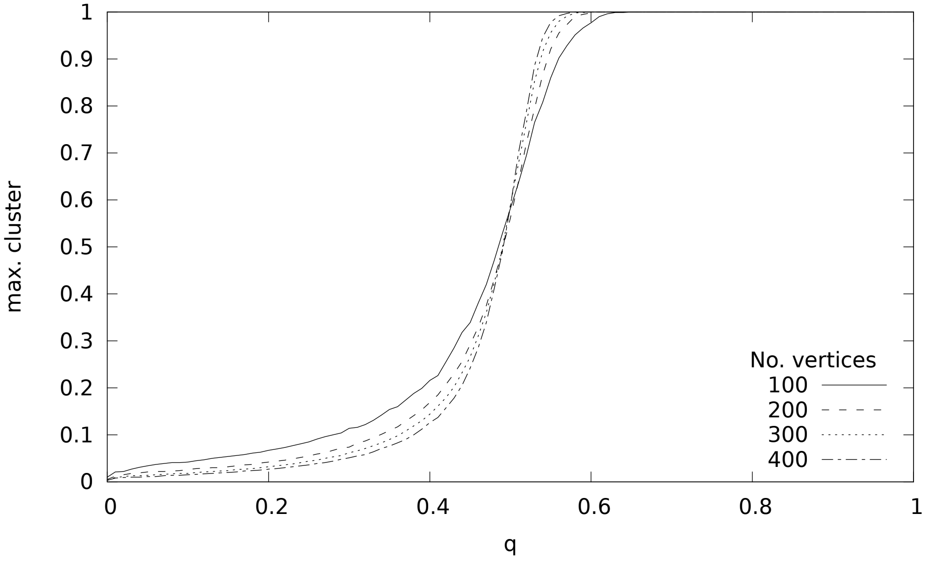

Figure 2. Relative sizes of the maximal clusters.

3. Clustering of Erdős-Rényi type graphs

Apart from the uniqueness of the complete graph, there are other options to choose graphs to colour. As the colourings are random, it is reasonable to choose the graphs randomly, too. Two well-known types of random graph exist; in this and the next section we analyse the clustering of their colourized versions.

At Erdős-Rényi type graphs with parameters (n, p) we drop out each edge of the complete graph with n vertices with probability $1-p$ and are left with probability p. As the value p is high, the results are near to the results of the previous section. The tendency becomes different, if p is small. On Fig. 4 we illustrate the result of clustering several graphs have near the same edges. *It can be seen on the figure as p is decreasing the curves become more and more asymmetric.* Table 1 shows that this is not an illusion, by the experiments the curves really have this property. To examine this carefully we need to choose an even smaller p, an even bigger n, and connected graphs only.

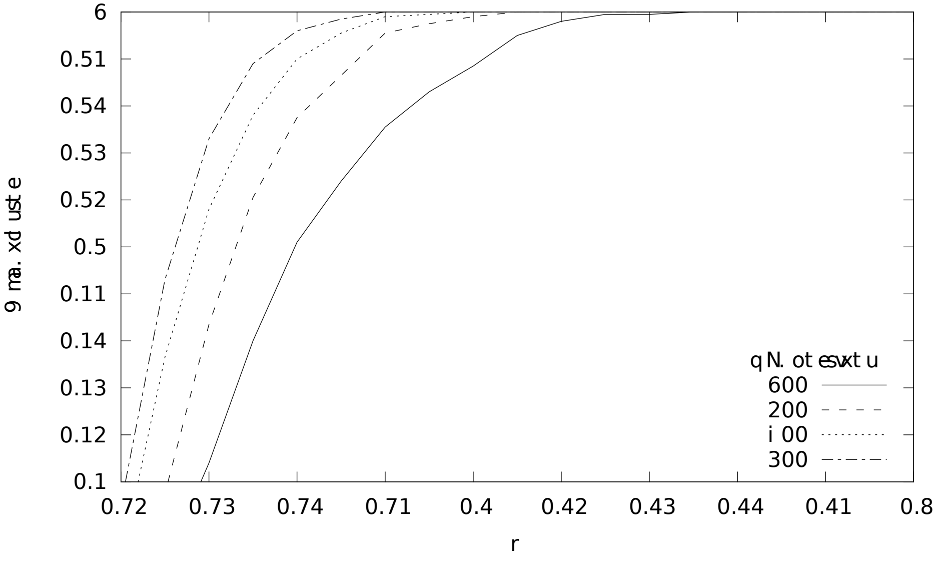

Figure 3. Relative sizes of the maximal clusters.

Table 1. Optimum of the curves in Fig. 4

p	1	0.25	0.11	0.06	0.04
q	0.55	0.56	0.57	0.58	0.60

Fig. 5 shows that decreasing p has opposite effects from increasing n: as p becomes smaller and smaller we get away from the previous asymptotic limit. It is an open question *whether for any fixed $p > 0$, as n heads to infinity the asymptotic limit becomes the same as in the previous section, or not?*

4. Clustering of Barabási-Albert type graphs

In the last decade of the previous century the interest on big graphs has increased. As Barabási and Albert found out the method of gener-

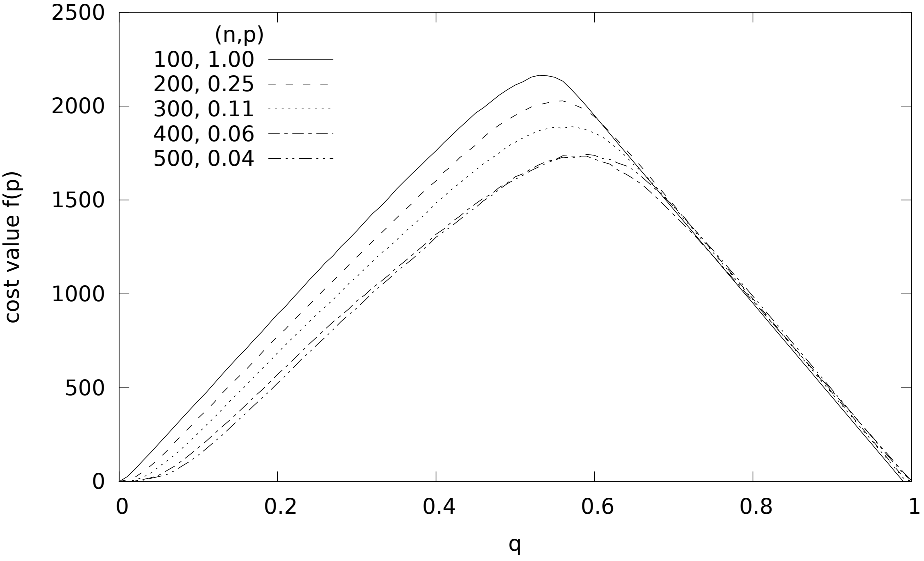

Figure 4. The turning point moves right if the p becomes small.

ation these graphs [6] the interest even jumped.

The Erdős-Rényi type graphs can be treated as dense graphs, as the degree of any node is np on the average, therefore its limit, when n tends to infinity, is ∞ for any positive p. The generation of the Barabási-Albert kind of scale free networks is the following: it starts from a complete graph with m_0 node, and each newer node connects to m older nodes using the preference attachment. We call these graphs as m_0/m type Barabási-Albert graphs. Here by adding a new node to the network, the sum of degrees increases by $2m$ in each step. Hence the limit of the degree of any node on the average is $2m$, i.e. a constant. This suggests, that the asymptotic behaviour of the clusters will be different in these two cases. Fig. 6 shows different sizes 3/2 type Barabási-Albert graphs. For the sake of scare, here we used error-bars. It is obvious that if the super-node (the node with the most edges) have many positive edges, have many negative edges, or have about equal many positive and negative edges, then the optimal clustering will be very different. As the structure of Erdős-Rényi type graphs is more symmetric, the difference of the maximal and minimal values of the cost function are not as big

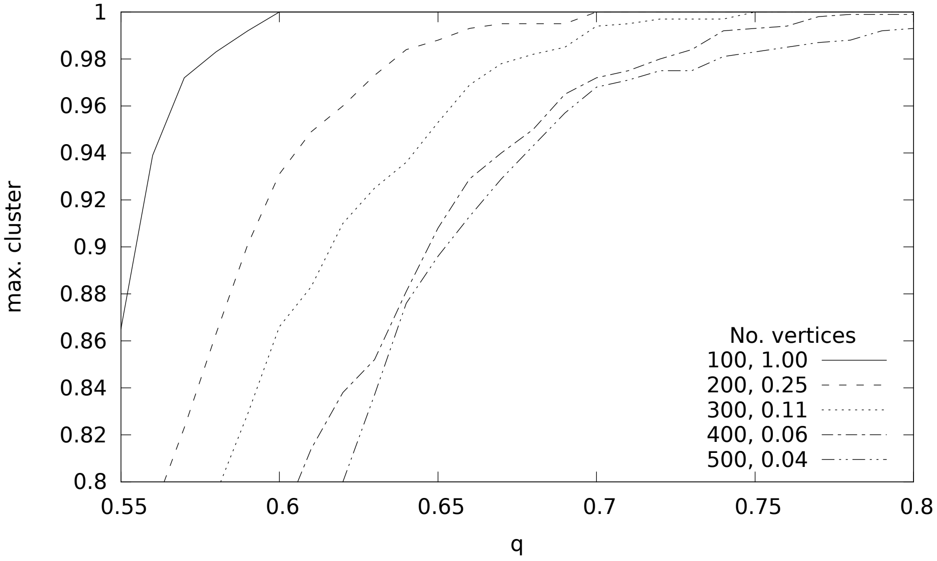

Figure 5. As p decreases the curve moves away from the previous limit.

as here. The number of edges in these graphs is a small fraction of the complete graphs, hence the value of the cost function is just its fraction. The asymmetry of the curves is obtrusive, by examining the data these functions have extrema around 0.7. Fig. 7 illustrates the relative size of the maximal clusters. In this case we show the whole graph given that almost everywhere it differs from the theoretical result on the complete graphs. As the size of the graph grows, the curve moving away from the line $y = 1$. We have the conjecture that *the limit of the relative size will be* 0 *for* $q \in [0, 1)$ *and* 1 *for* $q = 1$, but without calculating these curves for even larger graphs we cannot neither support nor reject it.

As these scale-free graphs have two parameters, we can examine how the clusters vary, as we change these parameters. The experiments show that as we increase the value of m, the same size graphs have more edges and hence have more conflicts, so the values of the cost function increase. Similarly the size of the maximal cluster increases, and in the interval $[0.8, 1]$ the curve is getting closer to the line $y = 1$.

Increasing m_0 a little at first has only a small effect, since it does not dramatically change the number of edges. However, the size of the core

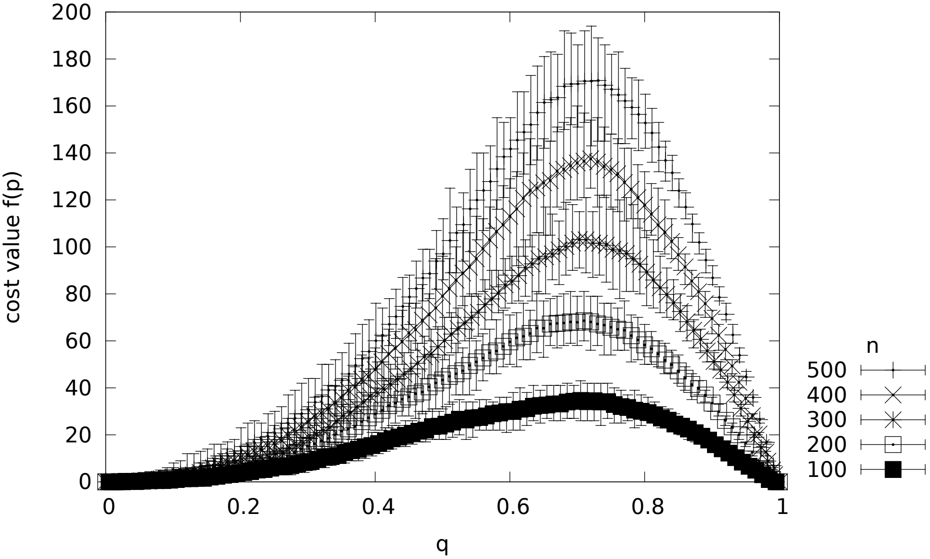

Figure 6. Asymmetry at curves of 3/2 type BA graphs

determines the degree of the competition for better preference values, and finally dives the scale-free parameter of the whole network. Fig. 8 shows, that *if m_0 is bigger, then the extrema tends to left.* Fig. 9 shows the size of the maximal clusters. In this case, like at Fig. 7 we can find a weak tendency, that *the size of the maximal cluster gets bigger.*

5. Technical details

For our students it was evident at implementing correlation clustering, to store the graphs in a modified adjacency matrix, where some ones were replaced by -1, if the corresponding vertices are connected with negative edges. The partition stored as a vector of numbers. The authors reimplemented these with bit-matrices, hence the calculation of the cost value reduced to bitwise operations. This speeded up the calculation alone about a hundred times for some optimization methods.

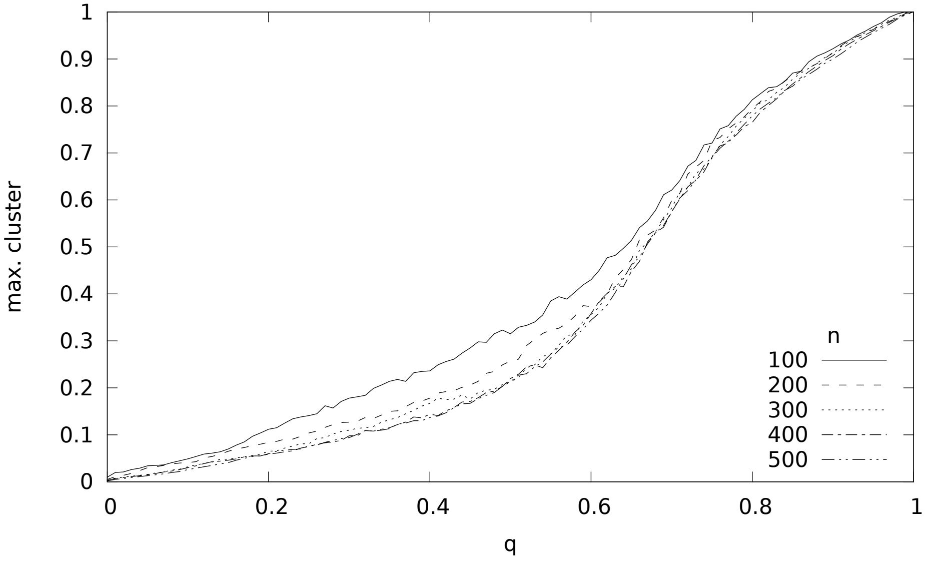

Figure 7. Relative sizes of the max. clusters of 3/2 type BA graphs.

Moreover the authors invented several optimization methods which use the specialities of the correlation clustering [2]. With these methods one can get results very close (about one percent distance) to result of the best implemented algorithm, the taboo method.

In the case of sparse graphs the adjacency matrix stores many zeros needlessly. Hence the search for non-zero elements could take more time than the calculation. Therefore we had implemented a storage method variant using the Yale Sparse Matrix Format [15] and an extra row, to be able to sort the signed edges of the generated graph. Although our graphs are symmetric, we store the edges in both directions, so after sorting the edges belonging to some verteces are together, we can process them in succession. Fig. 10 shows the running time of the clusterings using different storage methods. With this new sparse representation the clustering is two times slower than with bitwise operation, nevertheless we believe that the calculations could be speeded up.

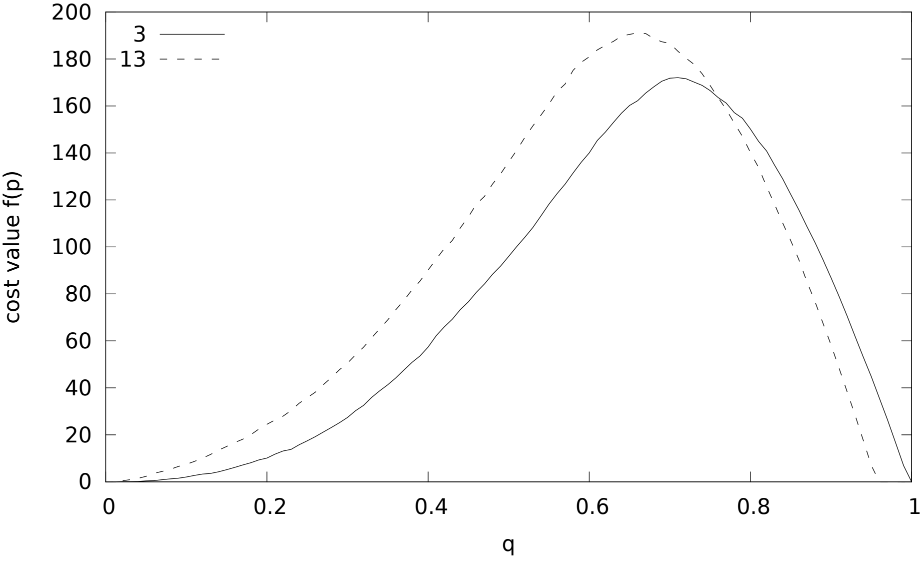

Figure 8. Curves of the cost values of BA graphs with different cores.

6. Conclusion and further works

In this article—based on previous work of Néda at al. [24, 25] – we have examined several random signed graphs, and their optimal partition. We have inspected the relation between the shapes of the cost-value and maximal cluster function and the parameters of the different type of random graphs. We have formulated several conjectures about behaviour of these curves. Although we had overstep the size of previous experiments, but one need to analyse the result of clustering of even bigger graphs to check the tendencies and interpret the conjectures in detail. The new data type to store graphs is a good step into the right direction, but other tricks need, to speed up the calculation. We have many open questions and there are several other type of random graphs to analyse [18, 28].

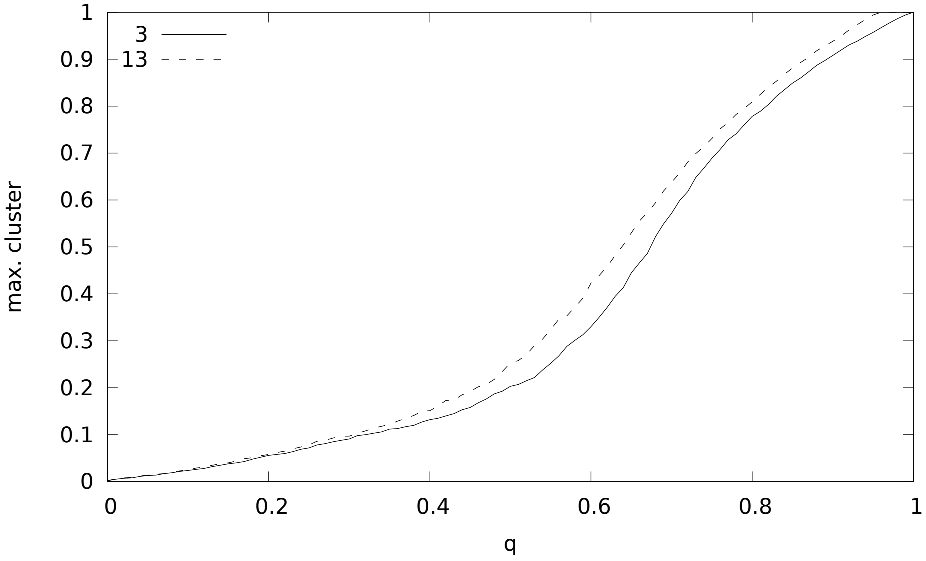

Figure 9. Relative sizes of the max. clusters of BA graphs with different cores.

References

[1] **Arató, M., A. Benczúr and A. Krámli,** On the solution of optimal performance of page storage hierarchies with an independent reference string, *Banach Center Publ.,* **6** (1) (1980), 9-15.

[2] **Aszalós, L. and M. Bakó,** *Fejlett keresőalgoritmusok,* Tankönyvtár, 2012. `http://morse.inf.unideb.hu/∼aszalos/diak/fka/`.

[3] **Aszalós, L., L. Hajdu and A. Pethő,** *On a correlational clustering of integers,* arXiv, preprint arXiv:1404.0904, 2014.

[4] **Bakó, M. and L. Aszalós,** Combinatorial optimization methods for correlation clustering, *Coping with complexity,* eds. D. Du-

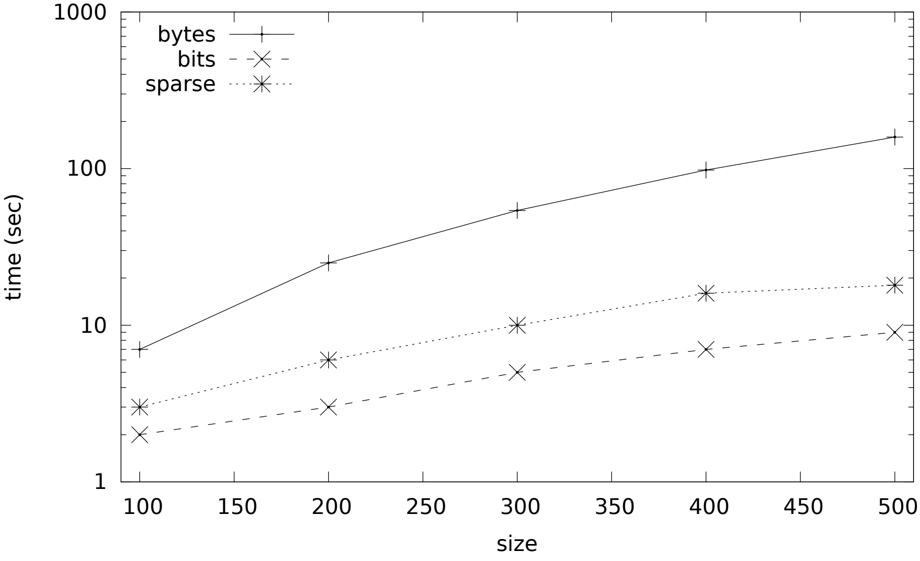

Figure 10. Running time of clustering 3/2 type BA graphs.

mitrescu, R.I. Lung and L. Cremene, Casa Cartii de Stiinta, Cluj-Napoca, 2011, 2-12.

[5] **Bansal, N., A. Blum and S. Chawla,** Correlation clustering, *Machine Learning,* **56** (1-3) (2004), 89-113.

[6] **Barabási, A.L. and R. Albert,** Emergence of scaling in random networks, *Science,* **286** (5439) (1999), 509-512.

[7] **Benczúr, A.,** The evolution of human communication and the information revolution - A mathematical perspective, *Mathematical and Computer Modelling,* **38** (7) (2003), 691-708.

[8] **Benczúr, A.,** Nagy rendszerek implementálásának software problémái, *MTA SZTAKI Közlemények,* **15** (1975), 120-130.

[9] **Benczúr, A.,** Adatbáziskezelő rendszerek hatékonyságának jellemzése Kolmogorov algoritmikus információmennyisége alapján, *Alk. Mat. Lapok,* **13** (1987), 285-289.

[10] **Benczúr, A.,** The digital universe - An information theoretical analysis, *Proc. 14th Int. Conf. on Computer Systems and Technologies,* ACM, 2013, 1-10.

[11] **Benczúr, A. and J. Kormos,** Az informatikus szakmáról, *Informatika a felsőoktatásban,* University of Debrecen, 2002, 437-447.

[12] **Bolla, M. and G. Tusnády,** Spectra and optimal patitions of weighted graphs, *Discrete Mathematics,* **128** (1) (1994), 1-20.

[13] **Bollobás, B. and O.M. Riordan,** *Handbook of Graphs and Networks: From the Genom to the Internet,* Mathematical Results on Scale-Free Random Graphs, Wiley - VCH GmbH & Co, 2003.

[14] **Borgs, C., J.T. Chayes, L. Lovász, V.T. Sós and K. Vesztergombi,** Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing, *Advaces in Mathematics,* **219** (6) (2008), 1801-1851.

[15] **Eisenstat, S.C., M.C. Gursky, M.H. Schultz and A.H. Sherman,** *Yale sparse matrix package I. The symmetric codes,* tech. report, DTIC Document, 1977.

[16] **Erdős, P. and A. Rényi,** On random graphs, *Publ. Math. Debrecen,* **6** (1959), 290-297.

[17] **Erdős, P. and A. Rényi,** On the evolution of random graphs, *Magyar Tud. Akad. Mat. Kut. Int. Közl.,* **5** (1960), 17-61.

[18] **Fazekas, I. and B. Porvázsnyik,** Scale-free property for degrees and weights in a preferential attachment random graph model, *J. of Probability and Statistics,* Article ID 707960, 2013.

[19] **Gilbert, E.N.,** Random graphs, *The Annals of Math. Statistics,* (1959), 1141-1144.

[20] **Lovász, L.,** *Large networks and graph limits,* American Math. Soc. **60**, 2012.

[21] **Lovász, L. and B. Szegedy,** Limits of dense graphs sequences, J. Combinatorial Theory, ser. B, **96** (6) (2006), 933-957.

[22] **Molnár, B., Gy. Szabó and A. Benczúr,** Investigation of system of criteria within selection processes for ERP systems: A Middle-European perspective, *Infocommunications J.,* **6** (1) (2014), 26-35.

[23] **Móri, T.,** On random trees, *Studia Sci. Math. Hung.,* **39** (1) (2002), 143-155.

[24] **Néda, Z., R. Sumi, M. Ercsey-Ravasz, M. Varga, B. Molnár and Gy. Cseh,** Correlation clustering on networks, *J. of Physics A: Mathematical and Theoretical,* **42** (34): 345003, 2009.

[25] **Néda, Z., R. Florian, M. Ravasz, A. Libál and G. Györgyi,** Phase transition in an optimal clusterization model, *Physica A: Statistical Mechanics and its Applications,* **362** (2) (2006), 357-368.

[26] **Sprons, O., D.R. Chialvo, M. Kaiser and C.C. Hilgetag,** Organisation, development and function of complex brain networks, *Trends in Cognitive Sciences,* **8** (9) (2004), 418-425.

[27] **Szemerédi, E.,** On sets of integers containing no k elements in arithmetic progression, *Acta Arith.,* **27** (585) (1975), 199-245.

[28] **Varga, I. and G. Kocsis,** Generating networks topologies with clustering similar to online social networks, *Cellular Automata for Research and Industry,* Krakow, Poland (to appear)

[29] **Yang, Bo, W.K. Cheung and J. Liu,** Community mining from signed social networks. Knowledge and engineering, *IEEE Transactions,* **19** (10) (2007), 1333-1348.

L. Aszalós, J. Kormos and D. Nagy
Faculty of Informatics
University of Debrecen
laszalos@unideb.hu
kormos.janos@inf.unideb.hu
nagydavid900120@gmail.com