Closing report on OTKA project 76481

In this project, we applied parametric and nonparametric methods to find the underlying structure of large, undirected simple or edge-weighted graphs.

In Section 1, we describe the parametric models we investigated (some of them we newly introduced) and the algorithms developed for parameter estimation. In Section 2, we consider nonparametric statistics (multiway cuts and normalized modularities, some of them we newly introduced), and estimate the constant of the volume-regularity of cluster pairs via the modularity spectra, hence extending the expander mixing lemma to several clusters. Testability issues are also considered. In Section 3, related work, conference trips, and involvement of students in the applications are discussed. In the References we only include the publications to which we directly refer in this report.

1 Parameter estimation in random graph models

We investigated two types of parametric random graph models, and gave algorithms for the maximum likelihood estimation of the parameters. Both models are capable to find hidden partitions of a simple graph's vertices for given number of clusters. About the number of clusters we can tell more in the next section.

1.1 EM algorithm for estimating the parameters of the block model

The so-called stochastic block model (introduced by Holland, later investigated by Bickel,Karrer,Rohe) is a generalization of the classical Erdős–Rényi model for several clusters. We formulated it in terms of mixtures so that to apply the EM algorithm for parameter estimation, which has not been applied to this situation yet. Our statistical sample was the adjacency matrix of the underlying simple graph, and considered it as an incomplete data specification, since the cluster memberships were missing. Therefore, it was straightforward to use the Expectation-Maximization (briefly, EM) algorithm proposed by Dempster, Laird, Rubin. This special application for mixtures is sometimes called collaborative filtering (Hofmann, Puzicha, Ungar, Foster). Roughly speaking, we had a mixture of binomial distributions, and above the binomial parameters we also estimated the parameters of the polynomially distributed latent membership vectors. According to the general theory of the EM algorithm, in exponential families (as in the present case), convergence to a local maximum is guaranteed. The algorithm was published in an extended abstract of the ASMDA'11 conference, see [6].

1.2 Parameter estimation in the α, β, and $k-\beta$ models

In the previous stochastic block model, within and between any pair of the vertex-clusters, edges came into existence with probability depending only on their endpoints' cluster memberships. However, there are more sophisticated models where the edge probabilities are not constant within the blocks. In

the one-cluster case, in [10], we investigated the following random graph model where the degree sequence is a sufficient statistic. (In fact, with different parameterization, this model was also introduced by Chatterjee, Diaconis, and Sly, but in [10] we also gave an algorithm for the maximum likelihood estimation of the parameters together with the proof of its convergence.) We have a simple random graph on n vertices, and for the probability p_{ij} that vertices i and j are connected we have that $\frac{p_{ij}}{1-p_{ij}} = \alpha_i\alpha_j$ $(i \neq j)$, and edges come into existence independently (but not with the same probability), where $\alpha_1, \ldots, \alpha_n$ are positive parameters. Instead of the odds, the log-odds (or logits) are used in the equivalent model of Chatterjee at al.: $\ln \frac{p_{ij}}{1-p_{ij}} = \beta_i + \beta_j$ $(i \neq j)$ with real parameters $\beta_1, \ldots, \beta_n$. We are looking for the maximum likelihood estimates of the parameters α's or β's based on the observed simple graph as a statistical sample. (It may seem that we have a one-element sample here, however, there are $\binom{n}{2}$ independent random variables in the background.)

We proved that in the above random graph model the degree sequence is a sufficient statistic. Further, if the degree sequence of the observed graph is an inner point of the polytope defined by the Erdős–Gallai conditions, then we proved that the maximum likelihood equation has a unique solution. We also recommended an algorithm and proved that the iteration of it converges to this unique solution.

This kind of an exponential model traces back Rasch and was used for psychological and educational measurements, later market research. The frequently cited Rasch-model involves categorical data, mainly binary variables, therefore the underlying random object can be thought of as a contingency table. We extended the Rasch- model to bipartite graphs with parameter sets $\beta_1, \ldots, \beta_m$ and $\gamma_1, \ldots, \gamma_n$ in the following way: $\ln \frac{p_{ij}}{1-p_{ij}} = \beta_i + \gamma_j$ $(i \in A,\, j \in B)$, where A and B are disjoint independent sets of the underlying bipartite graph's vertices, and the edges between A and B come into existence independently.

Making use of this extension, we generalized the β-model for the k-cluster case. Given $1 \leq k \leq n$, we are looking for k-partition $(V_1, \ldots, V_k)$ of the vertices such that vertices are independently assigned to clusters with certain probabilities (also to be estimated); and given the cluster memberships, vertices $i \in V_a$ and $j \in V_b$ are connected independently, with probability p_{ij} such that

$$\ln \frac{p_{ij}}{1-p_{ij}} = \beta_{ib} + \beta_{ja}$$

for any $1 \leq a, b \leq k$ pair. To estimate the parameters, the EM algorithm can again be used. More precisely, for the within-cluster edges we use the parameter estimation of the α- or β-model, hence obtain estimates of β_{ia}'s $(i \in V_a)$ in each cluster separately $(a = 1, \ldots, k)$; whereas, for the inter-cluster edges we use the extended Rasch-model to bipartite graphs. Note that here the parameter β_{ib} for $i \in V_a$ embodies the affinity of vertex i of cluster V_a towards vertices of cluster V_b; and likewise, β_{ja} for $j \in V_b$ embodies the affinity of vertex j of cluster V_b towards vertices of cluster V_a. By the model, this affinities are added together on the level of the log-odds. This model was introduced as k-β model in [11] and is is applicable to social networks, where attitudes of persons in the same social group (say, a) are the same toward members of another social group (say, b), though, this attitude also depends on the person in group a, and vice versa. The model may also be applied to biological networks, where the clusters consist, for example, of different functioning cells or units of the brain.

2 Nonparametric methods for clustering networks

In this setup, we were interested in finding homogeneous clusters of the vertices, also making considerations about the optimal number of clusters. To this end, nonparametric statistics (multiway cuts and modularities) were minimized or maximized over the k-partitions of the vertices. Under certain balancing conditions, we proved the testability of some of these statistics in the sense of L. Lovász, B. Szegedy, et al. The estimations were given in terms of the eigenvalues of unnormalized or normalized Laplacian and modularity matrices. Based on the spectral gaps between some structural and the other eigenvalues, we were able to find the optimal number of clusters, and by means of the eigenvectors – corresponding to the so-called structural eigenvalues – also gave algorithms to find the optimal partitions. The methods were extended to directed graphs and rectangular arrays too.

2.1 Modularity spectra and discrepancy

In [4] we defined end extended to edge-weighted graphs the balanced and normalized versions of the the Newman–Girvan modularity that focuses on the diagonal blocks and is capable to find clusters of intra-cluster connections larger than expected under independent attachment of the vertices. We proved that for given integer k (less than the number of the positive eigenvalues of the normalized modularity matrix) the k-partition of the vertices which maximizes the normalized modularity can be obtained by applying the k-means algorithm to the representatives of the vertices based on the eigenvectors corresponding to the $k-1$ largest eigenvalues of the normalized modularity matrix. We demonstrated through examples that the proper dimension depends on the number of eigenvalues of positive sign, whereas negative eigenvalues indicate an anticommunity structure (lower inter-cluster connections than expected under independent attachment of vertices). More generally, in [5], we managed to estimate the constant of volume-regularity of the cluster pairs by means of the gap in the spectrum between some structural (large absolute value) and the other eigenvalues and the k-variance of the representatives (objective function of the k-means algorithm). We also extended these concepts to the SVD of binary arrays or contingency tables, see [1].

We used the general framework of an edge-weighted graph. We defined the normalized version of the modularity matrix whose spectrum is in the [-1,1] interval, 0 is always an eigenvalue, and 1 is not an eigenvalue if the underlying graph is connected (the weight matrix of the edges is irreducible in the edge-weighted case). In fact, the introduction of this matrix $\mathbf{M}_D$ is rather technical, the expander mixing lemma can better be formulated with it, see [3]. Since the classical spectral gap of the underlying graph is $1 - \|\mathbf{M}_D\|$, a large spectral gap indicates small discrepancy as a quasi-random property discussed in Chung, Graham, Linial, et al. If there is a gap not at the ends of the spectrum, we want to partition the vertices into clusters so that a relation similar to the above property for the edge-densities between the cluster pairs would hold. For this purpose, we used a slightly modified version of the volume regularity's notion introduced by Alon, Coja-Oghlan, Han, Kang, Rödl, and Schacht, and defined so-called α-volume regular cluster pairs, where α is the smallest discrepancy of the pairs.

Since generalized random graphs (discussed in Subsection 1.1) can be viewed as edge-weighted graphs with a special block-structure burdened with random noise, we were able to give the following spectral characterization of them (in the submitted book of M. Bolla and the rectangular analogue in [1]). Fixing k, and tending with n to infinity in such a way that the cluster sizes grow at the same rate, there exists a positive number $\theta < 1$, independent of n, such that for every $0 < \tau < 1/2$ there are exactly $k-1$ eigenvalues of $\mathbf{M}_D$ greater than $\theta - n^{-\tau}$, while all the others are at most $n^{-\tau}$ in absolute value. Further, the k-variance of the vertex representatives constructed by the $k-1$ transformed structural eigenvectors is $\mathcal{O}(n^{-2\tau})$, and the cluster pairs are α-volume regular with any small α, almost surely. Note that generalized quasirandom graphs, defined by L. Lovász and V. T. Sós, are deterministic counterparts of generalized random graphs with the same spectral properties.

Our main theorem in [5] roughly states the following asymptotic result. Assume that $n \to \infty$ such that there are no dominant vertices, and the eigenvalues of $\mathbf{M}_D$, enumerated in decreasing absolute values, are

$$1 \geq |\mu_1| \geq \cdots \geq |\mu_{k-1}| > \varepsilon \geq |\mu_k| \geq \cdots \geq |\mu_n| = 0.$$

The partition $(V_1, \ldots, V_k)$ of the vertices is defined so that it minimizes the weighted k-variance s^2 of the optimum vertex representatives – obtained by means of the eigenvectors corresponding to $\mu_1, \ldots, \mu_{k-1}$. Then, with some other technical assumptions, the (V_i, V_j) pairs are $\mathcal{O}(\sqrt{2ks}+\varepsilon)$-volume regular $(i \neq j)$ and a similar relation holds for the intra-cluster discrepancies.

Note that, provided the underlying graph is connected, $|\mu_1| = 1$ can only be if $\mu_1 = -1$, and hence, in the 2-cluster case, our graph is a bipartite expander, whose spectrum was characterized earlier by N. Alon.

2.2 Spectral clustering of graphs and biclustering of contingency tables

Algorithms, using the results of Subsection 2.1 were introduced to find minimum normalized or regular cuts of edge-weighted graphs. Given a connected edge-weighted graph, we inspect the largest absolute value eigenvalues of its normalized modularity matrix. If n is 'very large', it suffices to find some leading eigenvalues. For this purpose fast numerical algorithms are available, e.g. the Lánczos method. Then select a k such that there is a gap between $|\mu_{k-1}|$ and $|\mu_k|$. We distinguish between the three following cases.

- If $\mu_1, \ldots, \mu_{k-1}$ are all positive, the output of the algorithm will be an approximation for the minimum normalized k-way cut of the graph with intra-cluster edge densities significantly higher than the inter-cluster ones (community structure).

- If $\mu_1, \ldots, \mu_{k-1}$ are all negative, the output of the algorithm will be an approximation for the maximum normalized k-way cut of the graph with intra-cluster edge densities significantly lower than the inter-cluster ones (anticommunity structure).

- If there are both positive and negative ones among $\mu_1, \ldots, \mu_{k-1}$, the output of the algorithm will be a clustering of the vertices with relatively

homogeneous edge-densities within the clusters and between any pairs of them (clustering with 'small' discrepancy). Some of the clusters or pairs of the clusters may have low, some of them may have high intra- or inter-cluster edge-density, as special cases.

Sometimes we want to classify data points, and first we put them into a Gaussian kernel to obtain the weight matrix of an edge-weighted graph. In this context, we clarified the role of reproducing kernel Hilbert spaces (in the submitted book of M. Bolla).

We proved similar results for so-called bicuts of contingency tables when we want to classify rows and columns simultaneously. We used the factors obtained by correspondence analysis to find biclustering of a contingency table such that the row–column cluster pairs are regular, i.e., they have small discrepancy. In the main theorem of [9] the constant of the so-called volume-regularity is related to the SVD of the normalized contingency table. Our result is applicable to two-way cuts when both the rows and columns are divided into the same number of clusters, thus extending partly the result of Butler estimating the discrepancy of a contingency table by the second largest singular value of the normalized table (one-cluster, rectangular case), and partly the result of [5] for estimating the constant of volume-regularity by the structural eigenvalues and the distances of the corresponding eigen-subspaces of the normalized modularity matrix of an edge-weighted graph (several clusters, symmetric case).

We applied the results for directed graphs, the weight matrix of which is quadratic but not symmetric, and we used in- and out-degrees for the normalization, see [9].

2.3 Testability of minimum balanced multiway cuts and normalized modularity spectrum with eigen-subspaces

L. Lovász, B. Szegedy, and coauthors defined the testability of simple graph parameters and proved equivalent notions of this testability. They also anticipated that their results remain valid if they consider weighted graph sequences with edge-weights in the [0,1] interval and no dominant vertex-weights. To this end, in [7], we slightly modified the definition of a testable graph parameter for weighted graphs and proved the testability of some minimum balanced multiway cut densities.

Furthermore, we proved the testability of the structural eigenvalues and the corresponding eigen-subspace of the normalized modularity matrix, see [8]. In view of this, spectral clustering methods can be performed on a smaller part of the underlying graph and give good approximation for the cluster structure. For the proofs we used theory of compact operators (normalized modularity matrix between finite dimensional Hilbert spaces and its continuous generalization which is a normalized graphon). In [2] we extended the theory to the convergence of contingency tables.

3 Related work, conference trips, applications

We gave talks on the following conferences (participation and/or travel was covered by this OTKA grant). The contents of these talks appeared in proceedings or in journals later. M. Bolla : COMPSTAT 2010 (Paris), ASMDA

2011 (Rome), EUROCOMB 2011 (Budapest), Applications of Graph Spectra in Computer Science, CRM, 2012 (Barcelona). V. Csiszár: Prague Stochastics 2010, 41st Probability Summer School, 2011 (St. Flour, France).

The book M. Bolla: Spectral clustering and biclustering (Learning large graphs and contingency tables) was submitted to the Wiley in January 2013 (based on an accepted proposal and contract). It does contain the number of this project and is to appear around the end of this July. However, it is not included in our main publication list.

Papers [8] and [9] were submitted in 2012 (based on an invitation for the participants of the above EUROCOMB 11 and the Barcelona's conference). For invitation, the paper [2] became Project 003371 of the Industry Gateway database which serves as a platform between researchrs and possible industrial partners; whereas [3] was included in the Intellectual Archive online multidisciplinary journal as Paper 431.2012.06.08.

PhD students of G. Tusnády: V. Csiszár defended her PhD thesis in 2009 and P. Hussami in 2011 in topics related to that of the present project. MS Students of M. Bolla at the BME and CEU (Erik Bodzsár, László Nagy, Ildikó Priksz, Zsolt Szabó) were involved in the applications of spectral clustering: they tested the parametric and nonparametric algorithms on randomly generated and real-world data in the framework of their diploma thesis. Since September 2012, M. Bolla has a PhD student (Ahmed Abo-Zaid) with whom they work on extension of spectral clustering methods to investigate strategic interactions in networks. In the fall semester 2012 M. Bolla led an Elective Undergraduate Research course at the Budapest Semester of Mathematics with two students (Max Del Giudice and Joan Wang) who applied spectral clustering for the directed graph of immigration-emigration data and obtained interesting results which they are going to submit to an undergraduate research journal in the US.

There are some rather technical papers in our main publication list, the topic of which do not directly relate to the topic of the present project, however, gave useful tools for our proofs. For example, papers of G. Tusnády on possible refinements of large deviation tail probabilities, of V. Csiszár on the EM algorithm, and of M. Bolla on some matrix decompositions used in a dynamic factor analysis algorithm.

References

[1] Bolla, M., Friedl, K., Krámli, A., Singular value decomposition of large random matrices (for two-way classification of microarrays), *Journal of Multivariate Analysis* 101 (2010) 434-446.

[2] Bolla, M., Statistical inference on large contingency tables: convergence, testability, stability. In: Proc. of the COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris (Y. Lechevallier and G. Saporta eds). Physica-Verlag, Springer (2010), 817-824.

[3] Bolla, M., Beyond the expanders, *International Journal of Combinatorics*, Paper 787596 (2011).

[4] Bolla, M., Penalized versions of the Newman–Girvan modularity and their relation to multiway cuts and k-means clustering, *Physical Review E* **84**, 016108 (2011).

[5] Bolla, M., Spectra and structure of weighted graphs, *Electronic Notes in Discret. Math.* **38** (2011), 149-154.

[6] Bolla, M., Parametric and nonparametric approaches to recover regular graph partitions. In: Proc. of the 14th ASMDA Conference, ed. R. Manca and C. H. Skiadas, June 7-10, 2011, Universita di Sapiensa, Roma, pp. 164-171.

[7] Bolla, M., Kói, T., Krámli, A., Testability of minimum balanced multiway cut densities, *Discret. Appl. Math.* **160** (2012), 1019–1027.

[8] Bolla, M., Modularity spectra, eigen-subspaces, and structure of weighted graphs, arXiv:1301.5254 [math.ST], submitted to the European Journal of Combinatorics.

[9] Bolla, M., SVD, discrepancy, and regular structure of contingency tables, arXiv:1301.5259 [math.ST], submitted to the Discrete Applied Mathematics.

[10] Csiszár, V., Hussami, P., Komlós, J., Móri, T. F., Rejtő, L., Tusnády, G., When the degree sequence is a sufficient statistic, *Acta Math. Hungar.* **134** (2012), 45-53.

[11] Csiszár, V., Hussami, P., Komlós, J., Móri, T. F., Rejtő, L., Tusnády, G., Testing goodness of fit of random graph models, *Algorithms* **5** (2012), 629-635.