

Recent Advances in Modularity Optimization and Their Application in Retailing

Andreas Geyer-Schulz and Michael Ovelgönne

Abstract In this contribution we report on three recent advances in modularity optimization, namely:

1. The randomized greedy (RG) family of modularity optimization algorithms are state-of-the-art graph clustering algorithms which are near optimal, fast, and scalable.
2. The extension of the RG family to multi-level clustering.
3. A new entropy based cluster index which allows the detection of the proper clustering levels and of stable core clusters at each level.

Last, but not least, several marketing applications of these algorithms for customer enablement and empowerment are discussed: e.g. the detection of low-level cluster structures from retail purchase data, the analysis of the co-usage structure of scientific documents for detecting multi-level category structures for scientific libraries, and the analysis of social groups from the friend relation of social network sites.

Andreas Geyer-Schulz

Informationsdienste und elektronische Märkte, Karlsruhe Institute of Technology (KIT), Karlsruhe

✉ andreas.geyer-schulz@kit.edu

Michael Ovelgönne

UMIACS, University of Maryland, College Park, MD

✉ mov@umiacs.umd.edu

CUSTOMER & SERVICE SYSTEMS
KIT SCIENTIFIC PUBLISHING
Vol. 1, No. 1, S. 37–48, 2014

DOI 10.5445/KSP/1000038784/05
ISSN 2198-8005



1 The RG Family of Algorithms

In this section we give a short presentation of the RG family of algorithms which optimize the modularity measure introduced by Newman and Girvan (2004):

$$Q(G, C) = \sum_{i=1}^p (e_{ii} - \alpha_i^2) \quad (1)$$

with $e_{ij} = \frac{\sum_{v_x \in C_i} \sum_{v_y \in C_j} m_{xy}}{\sum_{v_x \in V} \sum_{v_y \in V} m_{xy}}$ and $\alpha_i = \sum_j e_{ij}$, where $G = (V, E)$ is a loop-free graph and $C = \{C_1, \dots, C_p\}$ is a partition of V and M is the adjacency matrix of G defined by $m_{xy} = m_{yx} = 1$ if $\{v_x, v_y\} \in E$ and 0 otherwise.

Ovelgönne et al (2010a) and Ovelgönne and Geyer-Schulz (2010) modified Newman's greedy algorithm (Newman, 2004) by randomizing the selection of the first join candidate and by restricting the search for the second join candidate to the neighboring clusters of the first candidate. The randomized greedy (RG) algorithm accepts the best local improvement and, if no local improvement after k repetitions can be found, it accepts the join with the least decrease in modularity (like a walksat algorithm). The effect of the randomization of the greedy algorithm is dramatic: In the 10th DIMACS Implementation Challenge (2012), the RG algorithm won the Pareto challenge and is currently the most efficient graph-clustering algorithm. The main reason for the efficiency of the randomization lies in the large number of equivalence classes of joins with the same increase in modularity in real data sets (Ovelgönne and Geyer-Schulz, 2012a).

A greedy algorithm – even with the modifications described above – always finds a local optimum only. Ovelgönne and Geyer-Schulz (2012b) and Geyer-Schulz and Ovelgönne (2013) introduced a second idea to find heuristics which lead to better optima in modularity maximization: the core group graph clustering (CGGC) scheme. The CGGC scheme combines the locally optimal solutions found by several runs of the RG algorithm (or any other modularity optimization algorithm) in such a way that all the vertices which are in the same cluster in all locally optimal solutions form a core group. The result of this is a core group partition. The clusters of a core group partition contain the elements which have been in the same cluster in all locally optimal partitions the core group partition has been built from. Geyer-Schulz and Ovelgönne (2013) and

Ovelgönne and Geyer-Schulz (2013) characterize the core group partition as a saddle point on a Morse graph. Since a saddle point is always a point from which several local optima can be reached, saddle points are good restart points for an ensemble learning algorithm. The CGGC scheme may be used repeatedly, we denote this variant as CGGCi.

The ensemble algorithm based on a combination of the RG algorithm and the CGGC scheme won the modularity quality challenge of the 10th DIMACS Implementation Challenge (2012) and currently is the algorithm with the highest modularity for large graphs (e.g. 1.3 million vertices, 14 million edges). Computing time ranges from approximately 30 seconds (RG) to 9 minutes (CGGCi/RG) on standard PCs for a graph with approximately 860 000 vertices and 16 130 000 edges. A recent benchmark of implementations of the randomized greedy algorithm in five different programming languages (C++, Java, C#, F#, and Python) has appeared in Stein and Geyer-Schulz (2013). For additional results, see Ovelgönne and Geyer-Schulz (2012b), Geyer-Schulz and Ovelgönne (2013), and Ovelgönne and Geyer-Schulz (2013).

Fortunato and Barthélemy (2007) showed the problem of the resolution limit of modularity clustering, namely that the number of clusters chosen by modularity based graph clustering algorithms is approximately the square root of the number of edges. The second implication of the resolution limit is that modularity clustering favors partitions with clusters of equal size. To eliminate the resolution limit from modularity clustering (Geyer-Schulz et al, 2013) introduce a link parametrized modularity function with the parameter λ replacing the number of edges in the graph:

$$Q(G, C, \lambda) = \sum_{i=1}^p \left(\frac{l_i}{\lambda} - \left(\frac{d_i}{2\lambda} \right)^2 \right) = \frac{1}{\lambda} \sum_{i=1}^p \left(l_i - \frac{d_i^2}{4\lambda} \right) \quad (2)$$

where l_i is the number of edges in cluster C_i and $d_i = 2l_i + l_i^{out}$ with l_i^{out} defined as the number of edges connecting vertices in C_i with vertices in the rest of the graph. For $\lambda \leq 4$ we get the singleton partition (the size of C_i is 1) and for $\lambda \gg 4L$ we get a single cluster which contains the whole graph. This modification is the basis for using the RG-family of algorithms for multi-level clustering. Note, that the second implication of the resolution limit is still in place: The parametrized modularity function still favors partitions with clusters of approximately equal size.

2 Recognizing Clusters

Last, but not least, there remains the open problem of assessing the quality of a partition of vertices produced by a graph clustering algorithm. Since assessing the quality of a cluster partition by human experts which is still the gold standard in many areas of data analysis is impossible because of the sheer size of the graphs becoming available from Internet data sources, the proper evaluation of the result of clustering algorithms has become a major research problem. A common formal approach to this problem is to repeatedly apply a clustering algorithm and to assess the stability of the solution by measuring the similarity of pairs of the produced partitions and to evaluate the distribution of the similarities. The measures suggested can be classified as classical measures (e.g. the Jaccard or the adjusted RAND index), set matching measures, and information based metrics. Two recent surveys on such measures are Meilă (2007) and Vinh et al (2010).

The key problem of this approach is that the similarity is assessed only between pairs of partitions and that the symmetries in solution sets are not properly respected: Consider e.g. a ring structure of n vertices. A modularity optimization algorithm will divide the ring in \sqrt{n} clusters with \sqrt{n} vertices – or as near as possible if n is not a square number. However, the partition we get will depend on the randomly selected starting point. And, of course, there exist n starting points and thus we may get n different partitions. Consequently, the ring structure has no cluster structure. However, measuring the similarity of pairs of partitions may not detect this.

Geyer-Schulz et al (2013) propose a new information measure to assess the information in a set of partitions. The basic idea of the measure is the following fact: Graph symmetries give rise to permutation groups: e.g. for a 9-element ring the permutation (912345678) is the generator of an automorphism group of the graph of this ring.

Let $Aut_s(P)$ be the set of all partitions generated from partition P by applying all permutations in the automorphism subgroup of graph G to partition P .

The information content of each vertex v is given by

$$H(v) = \min_{l \in L} H(v, l) = - \sum_{i=1}^p P(i, v, l) \log_2 P(i, v, l) \quad (3)$$

where p is the number of clusters and $P(i, v)$ the probability that vertex v is in cluster C_i labelled l_i for all $Aut_s(P)$. L is the set of all possible ways to label the clusters in the partitions in P . The information content of vertices can be used to identify vertices which belong to a cluster (stable core groups) and vertices whose cluster membership changes.

The total entropy for the set of partitions $Aut_s(P)$ is then simply $H(Aut_s(P)) = \sum_{v \in V} H(v)$. This measure still depends on the choice of the assignment of vertices to clusters. To make the measure unique, we select the assignment of vertices to clusters which minimizes $H(Aut_s(P))$. And for partitions of regular graph structures like a ring, the entropy will be maximal, if the set of partitions we have used is $Aut_s(P)$. The last condition also indicates the main weakness of the proposed measures, namely the missing capability to efficiently compute $Aut_s(P)$.

3 Customer Empowerment, Customer Enablement and Modularity Clustering

Bowen and Lawler (1992) require that empowerment of service employees must be complemented by enabling service employees by management support, knowledge support, and technical support, so that empowerment works. The same holds for customer empowerment: Customer empowerment will only work, if customers are also enabled. And enablement of the customer depends on customer-oriented service processes. Critical for the success of such services is the development of a proper theoretical framework for each concrete service which spans the gap between marketing and computer science.

In the following we present three scenarios in which we highlight the potential of the RG family of modularity clustering algorithms for implementing innovative customer-oriented service processes.

The first scenario is a retail scenario: Recently (Die Zeit, 19.12.2012 (Schadwinkel, 2012c,a,b)), several retailers (and direct marketers) offer the time-buying customer (Berry, 1979) new services like “the menu in the bag” (Perfetto, Karstadt), “the walk-through receipt book” (Kochhaus, Hamburg and Berlin) and “the delivered menu” (KommtEssen, Kochzauer, KochAbo, Schlemmertüte, HelloFresh). What the customer buys, is a bundle of complementary products (the ingredients of the menu (more

or less preprocessed) in the proper weight ratio) together with a process description for cooking the menu.

Even a rather crude value analysis of these menu services (see e.g. Anderson et al (2006)) reveals that customer benefits are not restricted to time savings, but that a rather subtle picture of consumer benefits exists: The bundle of complementary products in one bag eliminates the search for menu ingredients (and the uncertainty of choice in combining ingredients). Preprocessed ingredients reduce the number of preprocessing steps (washing, cooking, blending, cutting to shape, ...) and the time needed for them. Packing ingredients in the proper weight ratio eliminates weighing (and errors in weighing) and reduces losses from excess ingredients caused by unfitting package sizes and the effort of waste reduction by proper menu sequencing. Eliminating the need for menu sequencing reduces the consumer's constraints in menu selection. Finally, a proper process description eliminates the search for the cooking receipt and eliminates uncertainty and potential pitfalls in the cooking process. Process descriptions, of course, can be multi-media based and distributed via mobile and/or social media (e.g. cooking video clips on YouTube).

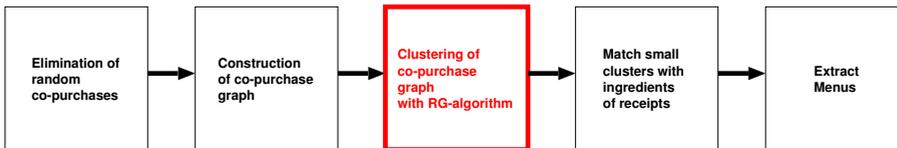


Fig. 1 The data analysis process for menu services

The micro-economic background theory for these services is the analysis of complementarities (induced by receipts as technological production functions) as presented e.g. in (Milgrom and Roberts, 1992, pp. 106-113). Clearly, analysing complementarities in a retailer's assortment is a promising application of RG algorithms (and of Ehrenberg's repeat-buying theory (see Ehrenberg (1988)), a classic quantitative marketing theory): Figure 1 shows a generic 5-step data analysis process for supporting this type of service from a retailer's point of view:

1. The data source are market baskets from POS scanner data. In this preprocessing step random co-purchases are eliminated with the help of Ehrenberg's logarithmic series models (see Böhm et al (2003)).

2. Next, an unweighted co-purchase graph is constructed from the non-random co-purchases generated in the previous step.
3. The unweighted co-purchase graph is clustered with the RG algorithm or one of its variants. For the innovative retail services discussed above a scale transformation of the cluster criterium (see formula 2) which produces relatively small clusters must be found. In addition, stable clusters must be identified with the help of the methods presented in section 2.
4. These clusters are matched with the ingredient list of a large recipe data base to identify popular menus which are candidates to be offered by such services.
5. The candidate menus are evaluated with regard to their potential margin contributions, popularity, the availability of the basic ingredients and the capacity of the prepackaging facility. Finally, the menus satisfying the retailer's profitability and availability criteria are selected and extracted.

The second scenario is the analysis of the co-usage structure of scientific documents. Data sources are Google Scholar or the BibTip GmbH (<http://www.bibtip.com>), a small German scientific recommender service provider. BibTip currently has recommender information on 150 million scientific documents with a growth rate of approximately 1.3 million transactions per day. The preprocessing step corresponds to the first subprocess in the first scenario.

However, in this scenario, the theoretical justification of the data analysis performed rests on the economic principles of incentive compatibility and self-selection (see (Milgrom and Roberts, 1992, pp. 140-146 and pp. 154-158) and Spence (1974)). Time pressures on researchers as well as students lead to incentive compatibility in information search: Almost all of the users of a scientific library conduct their literature search task-specific and as efficiently as possible. As a consequence, their choice behavior truthfully reflects their search interests and the search logs record their true behavior (incentive compatibility). The task-specific nature of the search behavior leads to self-selection: Users choosing the same information object have similar search interests and perform similar research tasks. This self-selection effect allows the transfer of Ehrenberg's repeat-buying theory to the analysis of anonymous session data and the identification of recommendations: All anonymous market baskets which contain

the same information object are considered as the purchase history of a latent, locally homogenous cluster of users whose revealed preferences (by their choices) are recorded, aggregated, and analyzed. The privacy of the user is respected. The results of the preprocessing step are the basis of BibTip's recommender service (Geyer-Schulz et al, 2003).

Construction of a recommender graph and clustering of the recommender graph can be done by the RG family of algorithms. Multi-level clusters can be exploited for an improvement of the categorization of scientific documents, since currently only about 12 percent of scientific documents are properly tagged in the German scientific library system. Prototypes of new user interfaces which enable the user to navigate on the recommendation network and thus through a semantic knowledge network have already been built (Neumann et al, 2008). Additional analysis options may include the automatic analysis of knowledge diffusion and knowledge transfer.

A direct application of social network analysis techniques for the purpose of analysing information needs, information exposure, information legitimation, information routes, and information opportunities as promoted e.g. by (Haythornthwaite, 1996) requires the observation of information and communication flows between scientists and students. Monitoring information and communication flows, however, is a violation of the privacy of users and of data protection laws in most countries. Librarians strongly oppose this type of system instrumentation. However, a qualitative analysis of the information flows in a university environment reveals that a strengthening of the information routes from the most experienced and advanced researchers in a disciplines to the novice students is desirable. A prototype of such a system (myVU) based on pseudo-identities and self-assessment of experience had been implemented and deployed in a university environment by (Geyer-Schulz et al, 2001). However, user acceptance and support for this system was rather low. The introduction of role-based extensions of scientific recommender systems has failed because of privacy concerns. The challenge in introducing even moderately personalized information systems is to do it in a privacy-aware manner which is accepted by users.

The third scenario is the identification of social groups from the friendship link relation of social network sites like Facebook, Google+, or Xing. In this setting, no preprocessing is necessary, the RG algorithms can be directly applied to the complete friendship graph. However, depending on

the application, partitions of clusters with appropriate size (corresponding to social groups at different scales) have to be selected. Doing this in a proper way remains difficult without additional data enrichment because of the trend to have considerably more friends in the Internet than in real life. At least for the network provider, data enrichment by attributes from personal profiles, additional communication data, etc. is an option. A first application scenario of social clustering for emergency alerts which is based on the social psychological analysis of bystander behavior has been described by Geyer-Schulz et al (2010), Ovelgönne et al (2010b), and Geyer-Schulz et al (2012). Again, the theoretical backing for this scenario is rather intricate: Darley and Latané provided a careful analysis of bystander behavior in emergencies (see Latané and Darley (1970)) which has been verified by social psychological experiments over the next 25 years (for a recent survey see Brehm et al (2005)). The key finding of Darley and Latané (and their followers) is that a victim in an emergency is more likely to receive help if at least a weak social tie with one bystander exists. This finding provides the link to Granovetter's theory of the role of weak links in society (Granovetter, 1973) and to social clustering.

The exploration of social clustering for customer empowerment and enablement remains an important topic for further research. As the three examples given above show, the design of services for customer empowerment and enablement require a sound theoretical foundation.

References

- Anderson JC, Narus JA, Rossum WV (2006) Customer value propositions in business markets. *Harvard Business Review* 84(3):91 – 99
- Berry LL (1979) The time-buying consumer. *Journal of Retailing* 55(4):58–69
- Böhm W, Geyer-Schulz A, Hahsler M, Jahn M (2003) Repeat-buying theory and its application for recommender services. In: Schwaiger M, Opitz O (eds) *Exploratory Data Analysis in Empirical Research*, Springer-Verlag, Heidelberg, *Studies in Classification, Data Analysis, and Knowledge Organization*, vol 22, pp 229 – 239

- Bowen DE, Lawler EW (1992) The empowerment of service workers: What, why, how, and when. *Sloan Management Review* 33(3):31–39
- Brehm SS, Kassin S, Fein S (2005) *Social Psychology*, 6th edn. Houghton Mifflin Company, Boston
- Ehrenberg AS (1988) *Repeat-Buying: Facts, Theory and Applications: Facts, Theory and Applications*, 2nd edn. Charles Griffin & Company Ltd, London
- Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proc National Academy of Sciences of the United States of America* 104(1):36 – 41
- Geyer-Schulz A, Ovelgönne M (2013) The randomized greedy modularity clustering algorithm and the core groups graph clustering scheme. In: Gaul W, Geyer-Schulz A, Okada A, Baba Y (eds) *German-Japanese Interchange of Data Analysis Results*, Springer, Heidelberg, *Studies in Classification, Data Analysis, and Knowledge Organization*, pp 15–34
- Geyer-Schulz A, Hahsler M, Jahn M (2001) Educational and scientific recommender systems: Designing the information channels of the virtual university. *International Journal of Engineering Education* 17(2):153 – 163
- Geyer-Schulz A, Neumann A, Thede A (2003) An architecture for behavior-based library recommender systems. *Information Technology and Libraries* 22(4):165 – 174
- Geyer-Schulz A, Ovelgönne M, Sonnenbichler A (2010) Getting help in a crowd – a social emergency alert service. In: Institute for Systems and Technologies of Information, Control and Communication (ed) *Proceedings of the International Conference on e-Business*, Athens, Greece, pp 207–220
- Geyer-Schulz A, Ovelgönne M, Sonnenbichler AC (2012) A social location-based emergency service to eliminate the bystander effect. In: MS Obaidat JF GA Tsihrantzis (ed) *e-Business and Telecommunications, Communications in Computer and Information Science*, vol 222, Springer Berlin / Heidelberg, pp 112 – 130
- Geyer-Schulz A, Ovelgönne M, Stein M (2013) Modified randomized modularity clustering: Adapting the resolution limit. In: Lausen B, van den Poel D, Alfred U (eds) *Algorithms from and for Nature and Life: Classification and Data Analysis*, Springer, Heidelberg, *Studies in Classification, Data Analysis, and Knowledge Organization*, pp 355–364

- Granovetter MS (1973) The strength of weak ties. *The American Journal of Sociology* 78(6):1360 – 1380
- Haythornthwaite C (1996) Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research* 18(4):323 – 342
- Latané B, Darley J (1970) *The Unresponsive Bystander: Why doesn't he help?* Appleton-Century-Crofts, New York
- Meilă M (2007) Comparing clusterings – an information based distance. *Journal of Multivariate Analysis* 98(5):873–895
- Milgrom P, Roberts J (1992) *Economics, Organization and Management*, 1st edn. Prentice Hall
- Neumann AW, Philipp M, Riedel F (2008) Recodiver: Browsing behavior-based recommendations on dynamic graphs. *AI Communications* 21(2-3):177 – 183
- Newman MEJ (2004) Fast algorithm for detecting community structure in networks. *Physical Review E* 69(6):066,133
- Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Physical Review E* 69(2):026,113
- Ovelgönne M, Geyer-Schulz A (2010) Cluster cores and modularity maximization. In: *ICDMW '10. IEEE International Conference on Data Mining Workshops, Piscataway*, pp 1204 – 1213
- Ovelgönne M, Geyer-Schulz A (2012a) A comparison of agglomerative hierarchical algorithms for modularity clustering. In: Gaul W, Geyer-Schulz A, Schmidt-Thieme L, Kunze J (eds) *Proceedings of the 34th Conference of the German Classification Society, Springer, Heidelberg, Studies in Classification, Data Analysis, and Knowledge Organization*, pp 225 – 232
- Ovelgönne M, Geyer-Schulz A (2012b) An ensemble-learning strategy for graph-clustering. In: Bader DA, Meyerhenke H, Sanders P, Wagner D (eds) *10th DIMACS Implementation Challenge – Graph Partitioning and Graph Clustering, Rutgers University, DIMACS – Center for Discrete Mathematics and Theoretical Computer Science*, [http://www.cc.gatech.edu/dimacs10/papers/\[18\]-dimacs10_ovelgoennegeyersschulz.pdf](http://www.cc.gatech.edu/dimacs10/papers/[18]-dimacs10_ovelgoennegeyersschulz.pdf)
- Ovelgönne M, Geyer-Schulz A (2013) An ensemble learning strategy for graph clustering. In: Bader DA, Meyerhenke H, Sanders P, Wagner D (eds) *Graph Partitioning and Graph Clustering, American Mathe-*

- mathematical Society, Providence, Contemporary Mathematics, vol 588, pp 187–205
- Ovelgönne M, Geyer-Schulz A, Stein M (2010a) Randomized greedy modularity optimization for group detection in huge social networks. 4th ACM SNA-KDD Workshop on Social Network Mining and Analysis 2010:1–9
- Ovelgönne M, Sonnenbichler AC, Geyer-Schulz A (2010b) Social emergency alert service – a location-based privacy-aware personal safety service. In: Proceedings of the 2010 Fourth International Conference on Next Generation Mobile Applications, Services and Technologies, IEEE Computer Society, Los Alamitos, CA, USA, pp 84 – 89
- Schadwinkel A (2012a) Begehbare Rezept. *Die Zeit* 2012(52):36
- Schadwinkel A (2012b) Liefern statt kaufen. *Die Zeit* 2012(52):36
- Schadwinkel A (2012c) Menü in der Tüte. *Die Zeit* 2012(52):36
- Spence MA (1974) *Market Signaling: Information Transfer in Hiring and Related Screening Processes*. Harvard University Press, Cambridge, Massachusetts
- Stein M, Geyer-Schulz A (2013) A comparison of five programming languages in a graph clustering scenario. *Journal of Universal Computer Science* 19(3):428 – 456
- Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J Mach Learn Res* 11:2837–2854