The Impact of Graph Symmetry on Clustering

Fabian Ball and Andreas Geyer-Schulz

Abstract This article investigates the effect of graph symmetry on modularity optimal graph clustering partitions. The key finding is that there actually exists an impact of graph symmetry, as more than 22 % of the analyzed graphs have an unstable partition. The results are based on an empirical analysis of 1254 symmetric graphs, which are a subset of the 1699 graphs that were analyzed by Ball and Geyer-Schulz (2018a). For each graph a modularity optimal partition is computed by a graph clustering algorithm. Additionally, the generating sets for the automorphism group of each graph are obtained. All computed partitions are tested for stability, which means that the symmetry that is captured by the automorphism group does not change this partition. Furthermore, definitions that allow to distinguish local and global symmetry of graphs are presented.

Fabian Ball · Andreas Geyer-Schulz
Karlsruhe Institute of Technology (KIT)
Kaiserstraße 12, D-76131 Karlsruhe, Germany
✉ fabian.ball@kit.edu
✉ andreas.geyer-schulz@kit.edu

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/09
ISSN 2363-9881

1 Introduction

The investigation of symmetry phenomena goes back to observations from the Stone Age (Reber, 2002) and examples of symmetry in ancient years (e.g. Sumerian and Greek engravings) are also used in the beginning of the relatively famous book of Weyl (1952). Moreover, in many scientific disciplines, especially in physics (Gross, 1996), symmetry effects play a major role in the theories and models. However, symmetry considerations in data analysis are not very common. For instance, Viana (2007) presents how a symmetry relation on the labels of structured data can be defined and how this leads to symmetric decompositions of the data. Murtagh (2009) shows how hierarchy in data, which can be represented by a dendrogram, can be interpreted in terms of equivalence by using the distances between the data points in the dendrogram. These distances are defined by an ultrametric. As a consequence, data points in the same branch of the dendrogram have the same distance to data points in another branch of the dendrogram. Thus, they are equivalent. Another example that is connected with data analysis is by Jabbour et al (2013). The authors introduce a pruning method for the well known `apriori` algorithm for frequent itemset mining (Agrawal and Srikant, 1994). The goal of `apriori` is to systematically identify sets of items that appear more often than a threshold from a set of transactions. Each transaction is itself a set of items. An obvious application is the analysis of (retail or online) shopping carts with the goal to generate recommendations on which products to buy in addition to those that are already in the shopping cart of a customer. The method of Jabbour et al (2013) finds equivalent items in the data, and if one of these items is infrequent, i.e. it can be excluded from the further search of `apriori`, all the equivalent items are infrequent, too.

Graphs can be used to model numerous situations, where it is necessary to portray the relations (the edges) between entities (the nodes). Besides this flexibility, graphs have the convenient property that the symmetry is already contained in the data and needs not to be defined separately. For example, the method of Jabbour et al (2013), roughly described above, transforms the frequent itemset problem into a graph representation and then finds the symmetries of the graph. This first example of symmetry that exists in graph data, however, shows a situation where symmetry is sought as part of a data mining algorithm instead of a situation where a given graph is analyzed exploratively and possibly even without prior knowledge about its structure. There are only few studies in

the literature that deal with symmetry of real-world graphs, i.e. graphs which emerge from some practical situation in reality: MacArthur et al (2008) show that symmetry exists in real-world networks, is caused by growth processes, often has a simple form, and they conjecture that symmetry may affect network properties such as robustness. For biological networks, Xiao et al (2008) study core relationships between biochemical control motifs by network quotients (the set of equivalence classes of vertices and the relationships between the equivalence classes). This removes the functional redundancy of biological networks introduced by growth processes with vertex and partial network duplication. In many biological networks, symmetry increases redundancy and thus serves as a reinforcement against damage. This seems also to hold for economic trading networks (Wang et al, 2009). Last, but not least, symmetry can be systematically exploited for speeding up graph symmetry discovery algorithms as Darga et al (2008) and Wang et al (2012) show. Consequences of symmetry are often known theoretically (e.g., shortest paths of equivalent nodes are all equivalent), but the practical impact on data analysis is neglected.

This paper is motivated by and part of a small research program on the role of symmetry in graph clustering algorithms and diagnostics. From a purely mathematical point of view, symmetry obviously affects the uniqueness and stability of optimal partitions (e.g. for the clustering of completely transitive graphs as the Petersen Graph (Ball and Geyer-Schulz, 2020)) and, for graph diagnostics, requires the use of invariant graph partition comparison measures (Ball and Geyer-Schulz, 2018b).

Practitioners of data analysis tend to discount such mathematical results as irrelevant, since they have been demonstrated on toy examples only. Many of them claim that, first, symmetry is unlikely to occur in real-life graphs and that, second, if it occurs, it does not affect the results of graph clustering algorithms.

The work of MacArthur et al (2008) is a first indication that symmetry exists in real-word graphs, but only for a very small number of graphs. We have investigated the claim that symmetry does not exist in real-world graphs in two related studies on datasets from `networkrepository.com`. The first study showed that only 272 of the 902 graphs investigated are asymmetric Ball and Geyer-Schulz (2018a). A more detailed analysis revealed that many of the asymmetric graphs have been artificially generated for benchmark purposes. The second study, which includes the graphs of the first study and has the largest published size ($n = 1699$), shows that over 70 % of all analyzed graphs

are symmetric (Ball and Geyer-Schulz, 2018a). Although the graphs from `networkrepository.com` do not constitute a random sample, our studies provide strong evidence for the existence of symmetry in real-world graphs.

This article is motivated by the second objection, namely that symmetry has no effect on the results of graph clustering algorithms. It poses the question, whether graph symmetry affects the graph clustering result, i.e. whether there are nodes in the resulting partition that can be exchanged between different clusters due to symmetry. A direct consequence of instability in terms of this article (we become more formal in Section 3) is the existence of multiple solutions of the same quality. This may not seem to be a big issue for practitioners, but when it comes to comparisons of graphs partitions, uniqueness of solutions is essential for traditional partition comparison measures. For non-unique partitions, similarity and dissimilarity measures are not unique and invariant versions of these measures are needed (Ball and Geyer-Schulz, 2018b). The practical conseqences of cluster instability depend on the application and are not studied in this contribution.

We restrict our investigation to modularity graph clustering (Newman and Girvan, 2004), which is, however, no restriction in general, as the whole analysis framework that we propose is irrespective of the used clustering method. More precisely, the analysis framework is not even restricted only to clustering, but works on a partition of the graph's node set and the symmetry group that acts on the nodes.

The following pages are structured as follows: Section 2 provides the necessary preliminaries, namely the definitions of a graph, of graph clustering, and of graph symmetry. In Section 3 we present five measures that allow the quantification of local and global graph symmetry, as well as the average size of a cluster and the average number of nodes affected by local symmetry. Local symmetry means that the different symmetries of a graph can be divided into several independent subsymmetries in comparison to the overall symmetry of the graph. We use these measures in Section 4, where we briefly describe how the analysis is conducted. After that, we present the analysis results, which are divided into simple and simplified graphs. Simplification tends to increase symmetry. Therefore, for the latter, even more partitions are unstable compared to the former. Finally, in Section 5, we wrap up our results and point into directions of future research.

2 Preliminaries

Before we can analyze graphs in terms of clustering partition stability we have to introduce and define several concepts. A graph $G = (V, E)$ is called simple if $V = \{1, 2, \ldots, n\}$ is a finite set of nodes and $E \subseteq \{\{u, v\} \mid u, v \in V,\ u \neq v\}$ is a set of symmetric binary relations, called edges. Edges $\{u, v\}$ are abbreviated uv. Furthermore, we assume graphs to be connected, which implies $m := |E| \geq n-1$ and for every possible partition $V = V_1 \cup V_2$ ($V_1 \cap V_2 = \emptyset$ and $V_1, V_2 \neq \emptyset$) there exist $u \in V_1, v \in V_2$ so that $uv \in E$. Graphs that are not simple as defined above are called non-simple. This means they either have directed, weighted, or multiple edges, possibly contain loops (i.e. edges from one node to itself), or are characterized by some other additional properties that are defined on the nodes or edges. Of course, also a combination of more than one such property results in a non-simple graph.

There does not exist a unique definition for graph clustering, but in general this means to partition the set of nodes V by some algorithm. A partition $P(V)$ (or just P if the context is clear) is a set of clusters $C_i \neq \emptyset$ so that $\bigcup_i C_i = V$ and $C_i \cap C_j = \emptyset$ holds for all $C_i, C_j \in P$, $i \neq j$. How a partition of a graph is formed, depends on the used method. Some try to divide the graph into a fixed number of clusters (e.g. Sanders and Schulz, 2013), others (e.g. Flake et al, 2004) minimize the cuts between clusters and maximize the cuts within clusters. A cut between two clusters is simply the number of edges that connect the two clusters. The nowadays still very popular *modularity* of Newman and Girvan (2004) is defined in a similar manner. It is a measure of graph partition quality, which is formed as the difference of the intra-cluster edge fraction (e_{ii}) and the expected quantity of edges that have the same type in terms of their incidence (a_i^2). These per cluster differences are summed up for the whole partition

$$Q := \sum_{C_i \in P} \left(e_{ii} - a_i^2 \right). \tag{1}$$

Given two clusters C_i and C_j, $e_{ij} := \frac{|\{uv \in E \mid u \in C_i, v \in C_j\}|}{2m}$ is the number of edges that connect a node in C_i to a node in C_j divided by twice the number of edges. As $e_{ij} = e_{ji}$, $2e_{ij}$ is the ratio of edges in G that connect C_i and C_j; e_{ii} is the fraction of edges within cluster C_i. Each cluster C_i is characterized by its incidence information, i.e. the ratio of edges of G which are incident to the cluster. This is $a_i := \sum_{C_j \in P} e_{ij}$. Therefore, e_{ii} is the probability for an

edge of G to be part of C_i and a_i^2 is the probability for an edge to connect two other (randomly chosen) clusters with the same incidence characteristics as C_i. That concludes the definition of Q: If $e_{ii} - a_i^2$ is large, most of the structurally equivalent nodes (characterized by edge incidence) are part of the same cluster and the cluster is a good one. Consequently, if this is true for all clusters, the partition is a good one.

More details on the mathematical definition can be found in Newman and Girvan (2004) or Geyer-Schulz and Ovelgönne (2012); Ovelgönne and Geyer-Schulz (2013). For the rest of this article we understand graph clustering as the optimization of modularity, which here means maximization. However, most of our upcoming definitions and analyses are completely independent of the used clustering approach and can also be applied to other graph clustering strategies, as in fact only a graph partition is needed, independent of its origin.

Graph symmetry can informally be described as the possibility of a cyclic reassignment of the node labels under the constraint that the edge relations stay the same. Formally, the relabeling is defined as a permutation, which is called *graph automorphism* in this context. A permutation $\pi : V \to V$ is a bijective map of nodes onto nodes, i.e. for every $u \in V$ there exists a unique image $u \mapsto u^\pi$ and $u^\pi \in V$, where u^π is the image of π applied to u. To explicitly write a permutation we use the condensed cycle notation $\pi = c_1 \cdots c_l$. Each c_i is a cycle of nodes $c_i = (u\ v \ldots w)$ so that $u \mapsto v$, $v \mapsto \ldots$, and finally $w \mapsto u$. Cycles of length one fix this node and they are normally omitted in the notation. A permutation π is an automorphism of G iff $G^\pi = G$. That means $(V^\pi, E^\pi) = (V, E)$ with $V^\pi := \{u^\pi \mid u \in V\}$ and $E^\pi := \{u^\pi v^\pi \mid uv \in E\}$. The condition $V = V^\pi$ always holds by definition and $E = E^\pi$ reflects the constraint of unchanged edge relations. These definitions imply the applicability of permutations not only on single nodes but also on sets (like clusters) and on other combinatorial objects (like partitions). In particular, $P^\pi := \{C^\pi \mid C \in P\}$, and C^π is covered by the definition of V^π above.

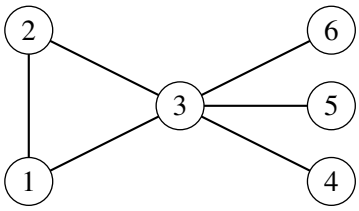
Two permutations π and τ can be catenated so that $u^{\pi\tau} = (u^\pi)^\tau$ is the successive application from left to right and $\pi\tau = \rho$ is again a permutation. The set of all automorphisms of G is denoted $Aut(G)$, which is a permutation group with the following properties:

- Identity: $\mathbf{1} \in Aut(G) : \mathbf{1}\pi = \pi\mathbf{1} = \pi \quad \forall \pi \in Aut(G)$
- Inverses: $\pi \in Aut(G) \iff \pi^{-1} \in Aut(G)$ with $\pi\pi^{-1} = \pi^{-1}\pi = \mathbf{1}$
- Closure: $\forall \pi, \tau \in Aut(G) : \pi\tau \in Aut(G)$
- Associativity: $\forall \pi, \tau, \rho \in Aut(G) : (\pi\tau)\rho = \pi(\tau\rho)$

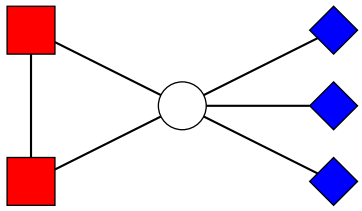
These properties allow a compact representation of $Aut(G)$ in terms of a subset $S \subset Aut(G)$ that *generates* the whole group, denoted $\langle S \rangle = Aut(G)$. The generation informally means to successively create new permutations by catenating all possible combinations of elements of S until $Aut(G)$ is recreated.

Often it is possible to decompose an automorphism group into several normal subgroups $H_1, \ldots, H_k$, denoted $H_i \lhd Aut(G)$. A subgroup is simply a subset of the permutation group $Aut(G)$, which is itself a group, and normality of a subgroup is the property that $\forall \pi \in Aut(G) : \pi H_i \pi^{-1} = H_i$. The term $\pi H_i \pi^{-1}$ is defined as the set $\left\{\pi\tau\pi^{-1} \mid \tau \in H_i\right\}$ and thus $\pi H_i \pi^{-1} = H_i$ informally means that H_i is unaffected by every $\pi \in Aut(G)$.

Transferred to graphs, this can be interpreted as the existence of an area in the graph, expressed by a subset of nodes, that is symmetric but independent of other such areas. This independence is nicely described by MacArthur et al (2008) who define these groups to be *support disjoint*. The support of a permutation π is the set of nodes not fixed by π: $\text{supp}(\pi) \coloneqq \{u \in V \mid u^\pi \neq u\}$ and, consequently, the support of a group is $\text{supp}(H_i) \coloneqq \bigcup_{\pi \in H_i} \text{supp}(\pi)$. Composing all the support disjoint H_i, H_j (i.e. $\text{supp}(H_i) \cap \text{supp}\left(H_j\right) = \emptyset$ $(i \neq j)$) yields the automorphism group $\prod_i H_i = Aut(G)$. If such a decomposition exists (i.e. $k > 1$, $Aut(G) \trianglelefteq Aut(G)$ is always true) it also can be carried out on the generating set S, given its elements satisfy two conditions (*irreducibility* and *uniqueness* (MacArthur et al, 2008, p. 3527)). We do not give the details, but assume that these conditions hold in the following. As a consequence, if decomposition is possible, $Aut(G) = \prod_i H_i = \prod_i \langle S_i \rangle$ where $H_i = \langle S_i \rangle$, $i = 1, \ldots, k$. An example is given in Figure 1.



(a) A small graph with the automorphism group $Aut(G) = \{\mathbf{1}, (1\ 2), (4\ 5), (4\ 6), (5\ 6), (4\ 5\ 6), (4\ 6\ 5), (1\ 2)(4\ 5), (1\ 2)(4\ 6), (1\ 2)(5\ 6), (1\ 2)(4\ 5\ 6), (1\ 2)(4\ 6\ 5)\}$.



(b) The squared red and the diamond blue nodes can both be mapped onto each other color-wise but the two symmetries are independent: $H_1 = \{\mathbf{1}, (1\ 2)\}$, $H_2 = \{\mathbf{1}, (4\ 5), (4\ 6), (5\ 6), (4\ 5\ 6), (4\ 6\ 5)\}$ with $Aut(G) = H_1 \times H_2$. The "missing" permutations of $Aut(G)$ emerge from the possible combinations $\pi\tau,\ \pi \in H_1, \tau \in H_2$.

Figure 1: Example of a small symmetric graph which has a decomposable automorphism group (left). The two independent symmetric areas are differently colored (right) and a decomposition of the generating set is $S = \{(1\ 2)\} \cup \{(4\ 5), (4\ 5\ 6)\}$.

3 Measures

A graph partition P is stable under the symmetry group $Aut(G)$, if there does not exist an automorphism $\pi \in Aut(G)$ so that $P \neq P^{\pi}$. However, if $P \neq P^{\pi'}$, π' is said to affect P and, therefore, symmetry has an impact on the partition. $|Aut(G)| = 1$ means only the identity automorphism exists, which leaves all nodes unchanged (mapped onto themselves), and the graph is *asymmetric*. As a consequence, only symmetric graphs at all may have an impact on partitions and asymmetric graphs can be excluded from our investigation.

Three equivalent measures of partition stability are presented by Ball and Geyer-Schulz (2018c). One definition to capture this property is to partition the generating set S, depending on whether an element $\pi \in S$ affects a partition P or not. The set S is split into two subsets $\tilde{\Pi}_{intra} \coloneqq \{\pi \in S \mid P^{\pi} = P\}$ and $\tilde{\Pi}_{inter} \coloneqq S \setminus \tilde{\Pi}_{intra}$, either of them can possibly be empty. If $\tilde{\Pi}_{inter} = \emptyset$, the partition is said to be stable, as no permutation exists for which P is mapped to $P^{\pi} \neq P$. A simple measure that can be derived is

$$ig_S \coloneqq \frac{|\tilde{\Pi}_{inter}|}{|S|}, \tag{2}$$

which is the fraction of the number of generating elements that cause instability compared to the size of the generator as a whole. It clearly takes values between 0 and 1. Smaller values indicate less instability, larger values indicate more instability.

The partition $P = \{\{1,2,3\},\{4,5\},\{6\}\}$ of the small graph in Figure 1 is affected by the symmetry: The generating set $S = \{(1\ 2),(4\ 5),(4\ 5\ 6)\}$ is divided into $\tilde{\Pi}_{intra} = \{(1\ 2),(4\ 5)\}$ and $\tilde{\Pi}_{inter} = \{(4\ 5\ 6)\}$, as $P^{(4\ 5\ 6)} = \{\{1,2,3\}, \{5,6\},\{4\}\} \neq P$. Hence, $ig_S = \frac{1}{3}$.

Let us now come to measures that quantify the symmetry of a graph itself. We call the first one *relative symmetry*, as it is defined as the fraction of nodes that are affected by symmetry:

$$rs_G \coloneqq \frac{|\mathrm{supp}\,(Aut(G))\,|}{n}. \tag{3}$$

It takes values between 0 (asymmetric graph) and 1 (completely symmetric graph). Based on this definition, we define

$$\overline{gs}_G \coloneqq \frac{rs_G}{k} \tag{4}$$

to be the *mean global symmetry*, where k is the number of support disjoint subgroups $H_1 \times \ldots \times H_k = Aut(G)$. Therefore, $\overline{gs}_G$ is the average fraction of affected

nodes of each independent symmetric area in the graph. The idea is to distinguish between the graph symmetry as such and local symmetry, which means that $\overline{gs}_G$ becomes small if there are many small areas of independent symmetry.

When we apply those definitions to the graph in Figure 1, $\text{supp}\,(Aut(G)) = \{1,2,4,5,6\}$ and $k = 2$, as $Aut(G) = H_1 \times H_2$ (see Figure 1b). The relative symmetry is then $rs_G = \frac{5}{6}$, which is a high value because all but one node (the center node) are affected by symmetry. However, as the automorphism group can be decomposed into two independent subsymmetries, the mean global symmetry is smaller ($\overline{gs}_G = \frac{5}{12}$).

The last two measures quantify the average/maximum number of nodes that are affected by an independent symmetry subgroup or are part of a cluster. Average support is defined as

$$\text{avg}\,(\text{supp}\,(Aut(G))) \coloneqq \frac{|\text{supp}\,(Aut(G))\,|}{k}, \tag{5}$$

where k is again the number of support disjoint subgroups of $Aut(G)$. To measure the average number of nodes per cluster we define

$$\text{avg}\,(C) \coloneqq \frac{n}{|P|}. \tag{6}$$

This definition is in line with the findings of Fortunato and Barthélemy (2007), which support the implicit assumption of balanced cluster sizes if modularity clustering is used.

Again, applying these measures on the small graph G (Figure 1) and the partition P from above, which consists of three clusters, yields $\text{avg}\,(\text{supp}\,(Aut(G))) = \frac{5}{2}$ and $\text{avg}\,(C) = \frac{6}{3} = 2$.

4 Empirical Analysis

Following our goal to find out whether graph symmetry possibly affects the graph clustering results, i.e. more precisely, if the resulting partition is stable, we carry out an empirical analysis of a large collection of graphs. The data base and the analysis approach of it is similar as described by Ball and Geyer-Schulz (2018a) and we only outline the essentials here:

- The meta-repository `http://www.networkrepository.com` is used as source of many graphs of diverse sizes and from different domains like (online) social networks, chemical / biological networks, recommendation

networks, ... We selected all datasets with a compressed size of at most 70 megabytes (a total of 3015 datasets; Ball, 2019). We must not assume that our selection of graphs generalizes to every possible area of application. Nonetheless, we cover a wide range of graphs which are relatively well known in the scientific community (e.g. the Karate network) and believe that our convenience sample is at least not too unrepresentative.

- Only symmetric graphs ($|Aut(G)| > 1$) are taken into account as it makes no sense to investigate symmetry effects of asymmetric graphs.
- We distinguish simple and non-simple graphs and we take only connected graphs into account. Simple graphs are defined as above (undirected, unweighted, no loops or multiple edges). A non-simple graph is characterized to have at least one of the above properties, and it can be transformed into a simple graph by removing those additional properties: Directed edges are removed by replacing them by undirected edges, weights are removed by forcing the weight to 1 (which is equivalent to the binary relation defined above), loops are removed without replacement, and multiple edges between two nodes are replaced by a single edge. This leads to the phenomenon that a simplified graph is possibly more symmetric than its non-simple counterpart, because there are more degrees of freedom in the symmetry definition. For simple graphs, two nodes are already equivalent if their adjacency relations are equivalent. Two nodes of a non-simple graphs are only equivalent if additionally to the adjacency (i.e. the structure) the corresponding edges have the same weights, both nodes have a loop, etc. Therefore, relaxing these properties may turn an asymmetric non-simple graph into a symmetric simple graph. For an example, see Ball and Geyer-Schulz (2018a, pp. 4–5).
- Partitions are computed using the "Core Groups Graph Clustering Randomized Greedy" (CGGCRG) algorithm (Ovelgönne and Geyer-Schulz, 2013). As it is a heuristic approach, one cannot be sure to find a partition that has globally the maximal modularity. However, we refer to the resulting partition and its modularity in the following as being optimal in the sense of "best solution found". The algorithm's result is always assured to be at least a (very) good locally optimal solution.

- A generating set for the automorphism group is obtained by the latest version of `saucy` (Darga et al, 2004, 2008; Katebi et al, 2012, `http://vlsi-cad.eecs.umich.edu/BK/SAUCY/`, accessed September 3, 2018). Other practically suitable algorithms are available: e.g. `nauty` (McKay, 1981; McKay and Piperno, 2014), `bliss` (Junttila and Kaski, 2007), `traces` (McKay and Piperno, 2014), `conauto` (López-Presa et al, 2014). `saucy` is optimized to perform well on large and sparse graphs.

The first analysis to identify symmetric graphs (Ball and Geyer-Schulz, 2018a) involved 902 simple and 797 simplified graphs. In this analysis, we build on the previous study: For a total of 629 simple and 625 simplified **symmetric** graphs identified in the first study we compute a modularity optimal partition and a generating set for the automorphism group. From this, the indicators n, m, $|P|$, Q, k, $|\tilde{\Pi}_{intra}|$, $|\tilde{\Pi}_{inter}|$, $|\text{supp}\,(Aut(G))\,|$, and $\max|\text{supp}\,(Aut(G))\,|$ are derived for each graph, which allows the calculation of all measures presented in the previous section. n and m are the number of nodes and edges of G, Q is the modularity for the computed partition P. In (Ball and Geyer-Schulz, 2018a) we report 630 simple and 634 simplified symmetric graphs. The slight difference is due to the high computational complexity so that we needed to exclude some graphs from this analysis.

In Table 1 we present descriptive statistics of some of the defined measures for simple and simplified graphs. A comparison reveals that the former are smaller (cf. m) on average and have a lower modularity (cf. Q). This is, however, not unusual as modularity normally increases with the graphs' size (Fortunato and Barthélemy, 2007). This fact also explains the differences in the distribution of the average cluster sizes. Furthermore, simple graphs are less symmetric (cf. rs_G) and have much more often a stable partition (cf. ig_S). The values of $\overline{gs}_G$ compared to rs_G for either type of graphs show that the automorphism groups are often decomposable (i.e. $k > 1$), but also completely symmetric graphs with an indecomposable group exist (maximum value $rs_G = \overline{gs}_G = 1$).

Table 2 shows how many simple and simplified graphs were analyzed originally (Ball and Geyer-Schulz, 2018a) and how many of them are symmetric. Furthermore, we present the number of graphs with unstable partitions and the percentages that describe the proportion of graphs compared to the number of all graphs and to the number of symmetric graphs.

Table 1: Partition stability statistics for `networkrepository.com` data sets for (a) simple and (b) simplified graphs. The simplified graphs on average are substantially larger, have a higher modularity, are more symmetric, and have more often an unstable partition.

	m	Q	$\frac{n}{\lvert P\rvert}$	$\frac{\lvert\mathrm{supp}(Aut(G))\rvert}{k}$	rs_G	$\overline{gs}_G$	ig_S
count	629	629	629	629	629	629	629
mean	3.713×10^5	0.558	478.500	2750.600	0.323	0.149	0.055
std	1.700×10^6	0.211	2138.700	45,495	0.285	0.249	0.220
min	1	0	2	2	1.280×10^{-5}	3.557×10^{-7}	0
25 %	47	0.474	6.667	2	0.103	0.049	0
50 %	71	0.600	8.400	2.115	0.238	0.073	0
75 %	103	0.663	13	3	0.446	0.114	0
max	1.785×10^7	0.999	35,517	1.049×10^6	1	1	1

(a) A total of 72 graphs have an unstable modularity optimal partition.

	m	Q	$\frac{n}{\lvert P\rvert}$	$\frac{\lvert\mathrm{supp}(Aut(G))\rvert}{k}$	rs_G	$\overline{gs}_G$	ig_S
count	625	625	625	625	625	625	625
mean	6.965×10^5	0.766	2887.500	14,394	0.709	0.267	0.208
std	1.522×10^6	0.184	21,983	69,580	0.389	0.424	0.393
min	46	0	4.001	2	3.333×10^{-5}	3.539×10^{-6}	0
25 %	12,540	0.708	144.600	2.946	0.380	0.000	0
50 %	1.200×10^5	0.820	510.310	12.440	0.976	0.002	1.839×10^{-5}
75 %	5.821×10^5	0.894	1612.800	986	1.000	0.500	0.057
max	1.723×10^7	0.996	4.738×10^5	1×10^6	1	1	1

(b) In contrast to the simple graphs, only 310 of 625 graphs have a stable modularity optimal partition.

Table 2: Overview of the total number of analyzed simple and simplified graphs by Ball and Geyer-Schulz (2018c), the number of symmetric graphs analyzed in this article, and the number of graphs with unstable partitions that were found. Additionally, the percentages of the symmetric graphs and of graphs with unstable partitions are given. The abbreviation "abs." means "absolute"; "rel." means "relative".

	Total	**Symmetric graphs**		**Graphs with unstable partitions**		
		abs.	rel. to total	abs.	rel. to total	rel. to symmetric
Simple graphs	902	629	69.734 %	72	7.982 %	11.447 %
Simplified graphs	797	625	78.419 %	315	39.523 %	50.400 %
All graphs	1699	1254	73.808 %	387	22.778 %	30.861 %

How well the automorphism groups can be decomposed is underlined by the comparison of the distributions of rs_G and $\overline{gs}_G$ for simple graphs in Figure 2. Graphs with $rs_G > 0.95$ seem to have an indecomposable group. The situation is similar for simplified graphs, however, the difference is even larger: Over 50 % of the graphs have $rs_G > 0.95$ but 75 % (more than 400) have $\overline{gs}_G \leq 0.05$, and for about 150 graphs $\overline{gs}_G > 0.95$ holds. We show these results in Figure 3. Please note that the definition of $\overline{gs}_G$ implies $\overline{gs}_G \leq 0.5$ if $k > 1$. These findings allow to state that symmetric real-world graphs often have either a very local symmetry or a very global one.

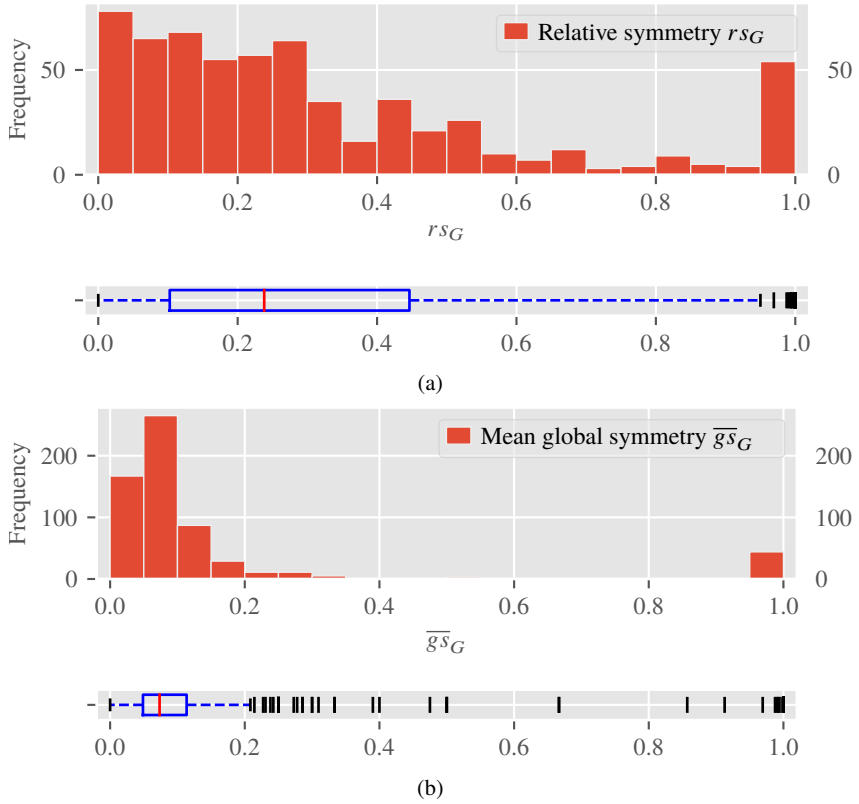


Figure 2: Comparison of the distributions for rs_G a and $\overline{gs}_G$ b for simple graphs. Most graphs – with only a few exceptions – have a relatively local symmetry.

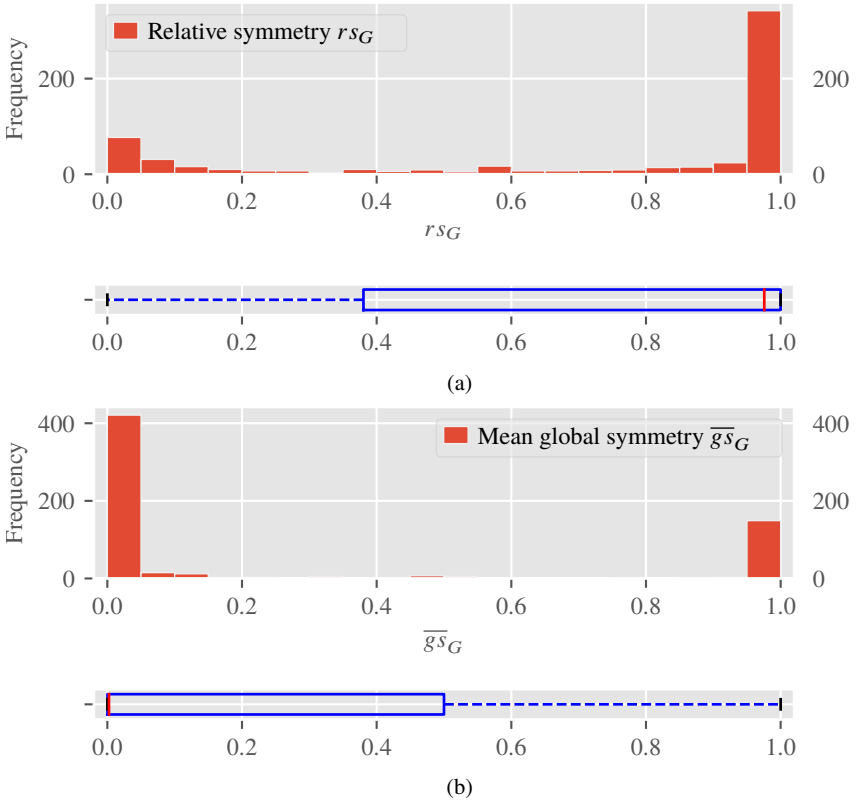


Figure 3: Comparison of the distributions for rs_G a and $\overline{gs}_G$ b for simplified graphs.

Another interesting observation is the distribution of the fraction of generating elements causing partition instability ig_S in Figure 4: Even though over 50 % of the simplified graphs have an unstable partition, either only very few or, in contrast, (nearly) all generating elements cause this instability. The finding for simple graphs is similar; therefore, we do not show a visualization of it.

In Table 3 we report the names of the simple graphs that we have found to have an unstable modularity optimal partition. The names often indicate their origin, for instance “ENZYMES” graphs are graph representations of molecules, “soc” indicates social networks, “web” relates to web graphs, and “rt” are retweet networks. These graphs can be used as a starting point for further work on the issue we outlined in this article. It would certainly be interesting to have a more detailed look into the origin and structure of these particular graphs to gain a deeper understanding of why the symmetry exists.

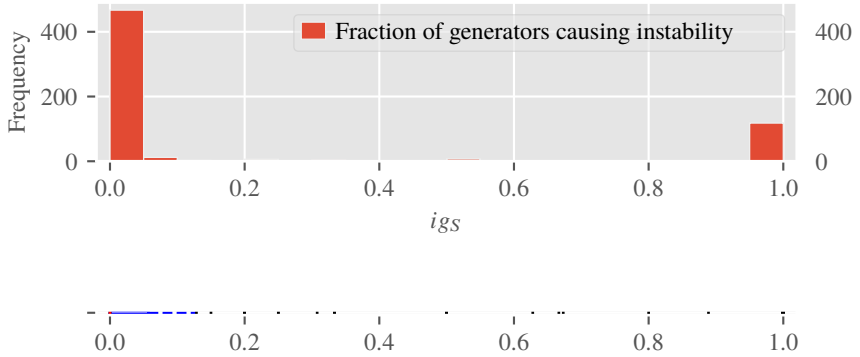


Figure 4: Distribution of ig_S for the simplified graphs. $ig_S \leq 0.057$ for 75 % of all graphs (cf. Table 1b), although about half of them have an unstable partition. In contrast, for a large number of graphs (nearly) the complete generator is responsible for instability.

Table 3: The list of all names of simple graphs that have an unstable modularity optimal partition in our analysis. The datasets can be found on `http://www.networkrepository.com`.

Names of simple graphs with unstable modularity optimal partitions				
as-22july06	com-dblp	EX1	johnson8-4-4	soc-youtube
auto	com-youtube	EX2	keller4	soc-youtube-snap
bfly	debr	EX4	keller6	tech-as-skitter
bio-dmela	diag	fe-sphere	MANN-a27	tech-internet-as
c-fat200-1	ENZYMES-g161	G48	MANN-a45	ukerbe1
c-fat200-2	ENZYMES-g272	G49	MANN-a81	ukerbe1-dual
c-fat500-1	ENZYMES-g293	G50	MANN-a9	web-arabic-2005
c-fat500-10	ENZYMES-g352	GD06-theory	power	web-edu
c-fat500-2	ENZYMES-g468	GD97-a	rt-islam	web-indochina-2004
c-fat500-5	ENZYMES-g509	GD98-c	rt-retweet-crawl	web-sk-2005
ca-citeseer	ENZYMES-g523	grid1	se	web-uk-2005
ca-dblp-2010	ENZYMES-g531	grid1-dual	soc-buzznet	web-wikipedia2009
cage	ENZYMES-g540	johnson16-2-4	soc-flickr	
cca	ENZYMES-g55	johnson32-2-4	soc-gowalla	
ccc	ENZYMES-g578	johnson8-2-4	soc-twitter-follows	

To summarize our main findings, first and foremost, we point out that the existence of more than 10 % (simple) and more than 50 % (simplified) graphs with modularity “optimal” unstable partitions is not a negligible percentage. Contrary to the statement of Garlaschelli et al (2010, p. 1705), we can not

confirm that "[t]he modular structure of real networks can be [...] seen as a symmetry-breaking property". Instead, we provide empirical evidence for the impact of graph symmetry on graph clustering. We defined this impact by simply checking whether the clustering solution is unique under the transformations of a permutation group, which is the graph automorphism group. As a further quantification of how the group acts on the graph and its modularity optimal partition, we used our measures defined in Section 3. It is interesting to see that there often is not just one large symmetric area in the graph, but many smaller and independent parts, which goes in line with the findings of MacArthur et al (2008). On the other hand, the fraction of generating elements which cause the instability of the partition shows that often the complete generator is involved. For a more in-depth analysis of the connections between local and global symmetry and an actual impact on the clustering partition we refer to the first author's PhD thesis (Ball, 2019).

A consequence of all this is that researchers working in applications of graph clustering should routinely check their results with regard to stability. Otherwise, it could happen that multiple equivalent clustering solutions exist of which one is not aware of. This problem is aggravated by the failure of the standard partition comparison measures for symmetric graphs. For this problem, an additional tool for diagnosing effects of graph automorphisms is the measure decomposition for partition comparison measures by Ball and Geyer-Schulz (2018b).

5 Conclusion

The presented results are the continuation of our work published before, where we (i) showed that graph symmetry exists in many real-world networks (Ball and Geyer-Schulz, 2018a), and (ii) defined stability of graph clustering partitions concerning the automorphism group of the graph (Ball and Geyer-Schulz, 2018c). Here, we could show that graph symmetry actually has an impact on the clustering result. In the end, 72 of 629 (11.447 %) simple graphs and 315 of 625 (50.400 %) simplified graphs have an unstable partition.

Future work could repeatedly compute partitions for the graphs with unstable partitions and check if they are always unstable or just randomly due to the random behavior of the used algorithm. Moreover, the complete analysis could be repeated with other modularity optimizing algorithms (e.g. Blondel et al, 2008) or methods that follow another optimization criterion (e.g. Raghavan et al, 2007). The last open problem that we want to point out is the question, how instabilities

should be handled. Possible solutions are to simply ignore them if the effect is small, break the symmetry in advance, use or develop other clustering criteria (e.g. maximize modularity for stable partitions) or methods (e.g. fuzzy clustering), or to reconsider the interpretation of the clustering result.

References

Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th Very Large Databases Conference (VLDB '94), Morgan Kaufmann Publishers Inc., Santiago de Chile (Chile), pp. 487–499. ISBN: 15-5860-153-8, URL: http://www.vldb.org/conf/1994/P487.PDF.

Ball F (2019) Impact of Symmetries in Graph Clustering. PhD thesis, Karlsruhe Institute of Technology (KIT), Institute for Information Systems and Marketing (IISM), Karlsruhe. DOI: 10.5445/IR/1000090492.

Ball F, Geyer-Schulz A (2018a) How Symmetric Are Real-World Graphs? A Large-Scale Study. Symmetry 10(1):29–1-17, Multidisciplinary Digital Publishing Institute (MDPI), Basel (Switzerland). DOI: 10.3390/sym10010029.

Ball F, Geyer-Schulz A (2018b) Invariant Graph Partition Comparison Measures. Symmetry 10(10):504–1-24, Multidisciplinary Digital Publishing Institute (MDPI), Basel (Switzerland). DOI: 10.3390/sym10100504.

Ball F, Geyer-Schulz A (2018c) Symmetry-Based Graph Clustering Partition Stability. Archives of Data Science, Series A 4(1):1–21, Karlsruhe Institute of Technology (KIT), Karlsruhe (Germany). DOI: 10.5445/KSP/1000085951/01.

Ball F, Geyer-Schulz A (2020) Comparing Partitions of the Petersen Graph. To appear.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10):1008–1-12, IOP Publishing Ltd, Bristol (United Kingdom). DOI: 10.1088/1742-5468/2008/10/P10008.

Darga PT, Liffiton MH, Sakallah KA, Markov IL (2004) Exploiting Structure in Symmetry Detection for CNF. In: Proceedings of the 41st Annual Design Automation Conference (DAC '04), Association for Computing Machinery (ACM), San Diego (USA), pp. 530–534. ISBN: 15-8113-828-8, DOI: 10.1145/996566.996712.

Darga PT, Sakallah KA, Markov IL (2008) Faster Symmetry Discovery Using Sparsity of Symmetries. In: Proceedings of the 45th ACM / IEEE Design Automation Conference (DAC '08), Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE), Anaheim (USA), pp. 149–154. ISBN: 978-1-605581-15-6, DOI: 10.1145/1391469.1391509.

Flake GW, Tarjan RE, Tsioutsiouliklis K (2004) Graph Clustering and Minimum Cut Trees. Internet Mathematics 1(4):385–408. DOI: 10.1080/15427951.2004.10129093.

Fortunato S, Barthélemy M (2007) Resolution Limit in Community Detection. Proceedings of the National Academy of Sciences of the United States of America 104(1):36–41, Siegmund DO (ed), National Academy of Sciences, Washington D.C. (USA). DOI: 10.1073/pnas.0605965104.

Garlaschelli D, Ruzzenenti F, Basosi R (2010) Complex Networks and Symmetry I: A Review. Symmetry 2(3):1683–1709, Multidisciplinary Digital Publishing Institute (MDPI), Basel (Switzerland). DOI: 10.3390/sym2031683.

Geyer-Schulz A, Ovelgönne M (2012) The Randomized Greedy (RG) Modularity Clustering Algorithm and the Core Groups Clustering (CGGC) Scheme. In: Proceedings of the German / Japanese Workshops Karlsruhe 2010 / Kyoto 2012, Gaul W, Geyer-Schulz A, Kunze J (eds), Springer, Berlin, Heidelberg, Studies in Classification, Data Analysis, and Knowledge Organization.

Gross DJ (1996) The Role of Symmetry in Fundamental Physics. Proceedings of the National Academy of Sciences of the United States of America 93(25):14256–14259, National Academy of Sciences, Washington D.C. (USA). DOI: 10.1073/pnas.93.25.14256.

Jabbour S, Khiari M, Sais L, Salhi Y, Tabia K (2013) Symmetry-Based Pruning in Itemset Mining. In: Proceedings of the 25th International Conference on Tools with Artificial Intelligence, Institute of Electrical and Electronics Engineers (IEEE), IEEE Computer Society, pp. 483–490. DOI: 10.1109/ICTAI.2013.78.

Junttila T, Kaski P (2007) Engineering an Efficient Canonical Labeling Tool for Large and Sparse Graphs. In: Proceedings of the 9th Workshop on Algorithm Engineering and Experiments (ALENEX), Society for Industrial and Applied Mathematics (SIAM), Philadelphia (USA), pp. 135–149. DOI: 10.1137/1.9781611972870.13.

Katebi H, Sakallah KA, Markov IL (2012) Conflict Anticipation in the Search for Graph Automorphisms. In: Logic for Programming, Artificial Intelligence, and Reasoning, Bjørner N, Voronkov A (eds), Lecture Notes in Computer Science, vol. 7180, Bjørner N, Voronkov A (eds). Springer, Berlin, Heidelberg, pp. 243–257. ISBN: 978-3-642287-17-6, DOI: 10.1007/978-3-642-28717-6_20.

López-Presa JL, Chiroque LF, Fernández Anta A (2014) Novel Techniques to Speed up the Computation of the Automorphism Group of a Graph. Journal of Applied Mathematics 2014:934637. ISSN: 1110-757X, DOI: 10.1155/2014/934637.

MacArthur BD, Sánchez-García RJ, Anderson JW (2008) Symmetry in Complex Networks. Discrete Applied Mathematics 156(18):3525–3531, Elsevier B.V. ISSN: 0166-218X, DOI: 10.1016/j.dam.2008.04.008.

McKay BD (1981) Practical Graph Isomorphism. Congressus Numerantium 30:45–87. ISSN: 0384-9864, URL: `http://users.cecs.anu.edu.au/~bdm/papers/pgi.pdf`.

McKay BD, Piperno A (2014) Practical Graph Isomorphism, II. Elsevier B.V. ISSN: 0747-7171, DOI: 10.1016/j.jsc.2013.09.003.

Murtagh F (2009) Symmetry in Data Mining and Analysis: A Unifying View Based on Hierarchy. Proceedings of the Steklov Institute of Mathematics 265(1):177–198, SP MAIK Nauka / Interperiodica, Moscow (Russia). DOI: 10.1134/S0081543809020175.

Newman MEJ, Girvan M (2004) Finding and Evaluating Community Structure in Networks. Physical Review E 69(2):026113–1-15, American Physical Society (APS), College Park (USA). DOI: 10.1103/PhysRevE.69.026113.

Ovelgönne M, Geyer-Schulz A (2013) An Ensemble Learning Strategy for Graph Clustering. In: Graph Partitioning and Graph Clustering, Bader DA, Meyerhenke H, Sanders P, Wagner D (eds), Contemporary Mathematics, vol. 588, Bader DA, Meyerhenke H, Sanders P, Wagner D (eds). American Mathematical Society (AMS), Providence, Rhode Island (USA), pp. 187–205. DOI: 10.1090/conm/588/11701.

Raghavan UN, Albert R, Kumara S (2007) Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. Physical Review E 76(3):036106–1-11, American Physical Society (APS), College Park (USA). DOI: 10.1103/PhysRevE.76.036106.

Reber R (2002) Reasons for the Preference for Symmetry. Behavioral and Brain Sciences 25(3):415–416, Cambridge University Press, Cambridge (United Kingdom). DOI: 10.1017/S0140525X02350076.

Sanders P, Schulz C (2013) Think Locally, Act Globally: Highly Balanced Graph Partitioning. In: Experimental Algorithms, Bonifaci V, Demetrescu C, Marchetti-Spaccamela A (eds), Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, vol. 7933, pp. 164–175. ISBN: 978-3-642385-27-8, DOI: 10.1007/978-3-642-38527-8_16.

Viana MAG (2007) Symmetry Studies for Data Analysis. Methodology and Computing in Applied Probability 9(2):325–341, Kluwer Academic Publishers – Plenum Publishers. DOI: 10.1007/s11009-007-9022-x.

Wang H, Yan G, Xiao Y (2009) Symmetry in World Trade Network. Journal of Systems Science and Complexity 22(2):280–290, Springer US. DOI: 10.1007/s11424-009-9163-9.

Wang J, Huang Y, Wu FX, Pan Y (2012) Symmetry Compression Method for Discovering Network Motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(6):1776–1789, Institute of Electrical and Electronics Engineers (IEEE). ISSN: 1545-5963, DOI: 10.1109/TCBB.2012.119.

Weyl H (1952) Symmetry. Princeton University Press, Princeton (USA). ISBN: 978-0-691023-74-8.

Xiao Y, MacArthur BD, Wang H, Xiong M, Wang W (2008) Network Quotients: Structural Skeletons of Complex Systems. Physical Review E 78(4):046102–1-7, American Physical Society (APS), College Park (USA). DOI: 10.1103/PhysRevE.78.046102.