# 2. Subgraphgroups

**Soham Roy**
Martikulation Number: 7028704
soro00002@stud.uni-saarland.de

## 1 1st idea

### 1.1 Do we really need subgroup discovery in this case for reasonable sized graphs

The first idea which simply compute the degree in X of each population member and stores rhem as continuous values .if we take a slightly different approach than sungroup discovery , for a sparse graph this can be done in theoretic polynomial time parameterized by the highest degree . for eg if we are trying to find a clique like structure in the graph which is a very common representation of a dense structure. Means we are trying to find a k-clique in a graph of maximum degree $\delta$. This can be done in polynomial time without even considering subgroup discovery algorithms: if we guess one vertex v in the clique, then the remaining vertices in the clique must be among the $\delta$ neighbors of v. Thus we can try all of the 2 $^{\delta}$ subsets of the neighbors of v, and return the largest clique like structure that we found. The total running time of this algorithm is O($2^{\delta} * \delta^{2} * n$), which is quite feasible for $\delta$ = 20 and small graphs. [1] This direct approach will work better than subgroup discovery algorithms also in reasonable time, so in essence it would be a [poor use of a subgroup discovery algorithm in this setting which has the potential to do more in more complicated settings.

### 1.2 Limitation and assumption

However,if we disregard that and follow the subgroup discovery , it is important to acknowledge that this approach has limitations and is not reliable. It fails to provide detailed information about the structure of the dense network and has limited capabilities in capturing its characteristics.

The primary objective of subgroup discovery is to identify subgroups with exceptional distributions. However, merely finding a set of nodes with high degrees does not yield meaningful subgroups when the entire graph is dense. This approach works only under the assumption that the graph is non-dense initially. If the entire population is dense, it will generate numerous subgroups, but none of them will provide significant insights. This assumption needs to hold true for the approach to be effective.

### 1.3 relevance of degree information

In the case of cliques, calculating the cluster coefficient can provide information about the connectivity among neighbors. However, this becomes interesting only after a certain degree threshold. Clusters with a low degree, such as a degree of 3, are rarely deemed significant, and manual intervention is often required to determine their relevance.

### 1.4 Loss of structural information

By converting the degree values into a continuous target attribute, the approach loses the structural information embedded in the friendship graph. Although the degree of a node represents its number of connections, it fails to capture the arrangement and pattern of those connections. Focusing solely

[1] Idea derived from page 10 , Parameterized algorithms by Cygan et.al- https://link.springer.com/book/10.1007/978-3-319-21275-3.

on the average degree disregards the specific topology and organization of the subgraphs within the graph.

## 2 2nd idea

### 2.1 Simplification of Target Attribute

Converting the densest subgraph information into a binary target attribute Y' simplifies the problem by reducing the complexity of the target variable. This approach combines the strengths of mining dense subgraphs and traditional subgroup discovery algorithms. It allows for the discovery of subgroups that exhibit a strong affinity to a well-defined dense structure, enabling more targeted and meaningful analyses of the graph data.

### 2.2 Limitations and Assumptions

The second idea also has limitations and assumptions that need to be considered. This approach assumes that the graph contains dense regions in specific areas and aims to mine the densest subgraph. However, it may overlook other dense regions or subgraphs that exist in different parts of the graph. While it can work well when there is only one prominent dense region, for graphs with multiple dense regions or caveman-like structures, it may miss out on important subgraphs.

### 2.3 Neglecting Negative Labels and Potential Information

Focusing solely on positive labels (membership in the densest subgraph) can lead to overlooking valuable information. Negative labels (non-membership in the densest subgraph) could provide insights into independent graphs or disconnected nodes that do not share edges. By disregarding negative labels, potential subgraphs or structures that are not part of the densest subgraph may be missed. For example, there could be independent graphs formed by nodes that do not have shared edges, and these provide meaningful information as well.

### 2.4 Loss of Structural Information

Additionally, converting the graph into a binary target attribute results in a loss of the structural information embedded in the friendship graph. The specific arrangement, connectivity patterns, and organization of subgraphs are not captured when considering only membership in the densest subgraph. There could be subgroups with interesting connectivity patterns or unique characteristics that are not captured by the densest subgraph alone.

### 2.5 Impact of Different Algorithms

It is important to consider that different algorithms may yield different results in terms of the identified densest subgraph, which can potentially impact the subsequent subgroup discovery process and the quality of the obtained results. The choice of algorithm plays a role in determining which dense subgraph is identified and utilized for subgroup discovery.

## 3 A word on ROSI and PICS

### 3.1 ROSI

ROSI (Kalofolias, Boley, and Vreeken 2019) emphasizes robust connectivity and simple descriptions in its objective. The goal of ROSI is to discover subgraphs that exhibit both robust connectivity and easily interpretable descriptions. To achieve this objective, ROSI goes beyond traditional density-based approaches and incorporates the notion of k-coreness while leveraging subgroup discovery algorithms. By efficiently mining large attributed graphs, ROSI aims to identify meaningful and robust subgraphs.

ROSI utilizes an optimization approach that combines the refinement operator $\rho$ and the optimistic estimator $\hat{f}$. The refinement operator generates a non redundant selector tree, ensuring that subgraph selectors are not duplicated. The optimistic estimator provides upper bounds on the objective function, enabling efficient pruning of sub-branches with suboptimal estimations. By using the Branch-and-Bound (BNB) algorithm, ROSI strives to discover the optimal solution, effectively finding robustly connected subgraphs with simple descriptions in an efficient manner.

### 3.2 PICS

The objective of the PICS (Akoglu et al. 2012) algorithm is to find meaningful patterns in attributed graphs, including clusters, bridges, and outliers. It employs a Minimum Description Length (MDL) approach to minimize the encoding cost. The objective function of PICS aims to find the partitioning that results in the lowest total encoding cost, which consists of the model description cost and the data description cost. PICS employs a greedy heuristic approach to iteratively refine the partitioning and find a good working model that compresses the data effectively.

## 4 Analysis

### 4.1 Comparison between PICS and ROSI: Greedy vs. Optimal Approach

While PICS utilizes a greedy approach to minimize the total encoding cost, it is not guaranteed to find the optimal result. In contrast, ROSI defines its objective density function in such a way that it can resort to finding the optimal value without relying on greedy approaches. By incorporating the BNB algorithm and an efficient optimization strategy, ROSI aims to achieve optimal solutions for discovering robustly connected subgraphs with simple descriptions. In ROSI, the mined subgroups are independent of each other, meaning that the discovery of one subgroup does not influence the discovery of others. Each subgroup is explored separately, allowing for a diverse range of subgroups to be identified.

Similarly, in PICS, the subgroups discovered are disjoint from each other. This means that the identified subgroups do not overlap or share common elements. PICS achieves disjointness through its own approach, which focuses on returning a set of disjoint cluster features.

both ROSI and PICS employ different strategies to avoid redundancy in their subgroup discovery processes. ROSI focuses on closed patterns and independent exploration, while PICS prioritizes disjointness among the identified cluster features.

### 4.2 Control and flexibility - Paremeters - A boon or curse

ROSI utilizes two parameters, dmax and yeta, which provide control and flexibility in its operation. In datasets with numerous features, ROSI demonstrates superior real-life applicability due to the ability to control the depth of the branch and bound (BnB) algorithm. This feature proves advantageous in scenarios where finding connectedness is crucial, such as determining the strength of collaboration between two communities. The higher the number of interconnected edges between communities, the stronger their collaboration. By manipulating the yeta parameter, ROSI can be tailored to find solutions that capture this collaboration strength.

In contrast, PICS highlights automation as one of its key contributions. However, there is no mechanism to convey specific information, such as the importance of connectedness, to the algorithm. This limitation becomes apparent when presented with adversarial graphs lacking significant density but exhibiting high connectedness. PICS may struggle in such cases, while ROSI, with a higher yeta value, would indicate the absence of such subgroups. Conversely, a lower yeta value in ROSI would identify the connected structures.

It is important to note that ROSI operates within the depth restriction of dmax, which may result in the risk of losing out on fine-grained patterns. As the patterns in ROSI can only contain conjunctive terms of at most dmax size, there is a potential limitation in capturing intricate patterns. On the other hand, PICS faces no constraints on depth and can potentially explore all features, allowing for comprehensive and diverse subgroup discovery.

In summary, ROSI proves beneficial when specific depth restrictions exist or when focusing on fine-grained patterns with limited conjunctions. Conversely, PICS shines when there are no depth restrictions, and a wide range of features and combinations need to be explored to capture comprehensive and diverse subgroups.

### 4.3 Choosing between them , maybe both

In one complete run of the programs, rosi can only give only top K subgroup while k is a parameter , the optimal number of K might need expertise over the field/graph. The outcome of rosi is a conjunction of pattern which maximizes its density objective function but in one run pics give all the subgroups it has discovered in the form of feature clusters and node groups corresponding to which feature clusters. If multiple subgroups are desired and the optimal K is uncertain, running PICS first to understand the feature clusters can inform subsequent runs of ROSI. However, if mining a limited number of top subgroups is the goal, ROSI is the preferred choice.

### 4.4 Adaptability to Dynamic Graphs

In dynamic graph scenarios, such as social media networks where nodes and edges continuously change based on friendships and new connections, ROSI demonstrates better adaptability compared to PICS. ROSI's design allows it to efficiently handle dynamic graphs by avoiding redundant calculations, making it well-suited for these situations. ROSIs adaptive nature enables it to respond to changes in the graph without the need for recalculating the entire range of calculations performed by PICS. When a subtree within the graph requires modification, ROSI can selectively update the density score calculation and refinement procedure, or choose to explore a previously pruned path based on the changes. It does not have to start from the root and traverse down the entire tree, enabling more efficient and targeted adjustments in response to graph dynamics. In contrast, PICS lacks an easy way to adapt to dynamic graphs while avoiding extensive recalculations. Its automated approach and comprehensive exploration make it less suitable for dynamic scenarios, as it would require repetitive computations on the entire graph each time a change occurs. Overall, ROSI's design and adaptability allow it to better handle dynamic graphs by efficiently incorporating changes without redundant calculations, providing an advantage over PICS in such dynamic scenarios.

### 4.5 Interpretability and Actionability of Discovered Patterns

ROSIs output tends to provide language descriptions that are concise and succinct, allowing for easy interpretation and understanding of the discovered patterns. The patterns generated by ROSI can often tell a coherent and actionable story about the data. In contrast, PICS may not produce language descriptions as succinctly as ROSI. While PICS excels in providing comprehensive subgroups and feature clusters, the resulting patterns may require further analysis and interpretation to extract meaningful insights and actionable information.

## References

Akoglu, Leman et al. (2012). “Pics: Parameter-free identification of cohesive subgroups in large attributed graphs”. In: *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, pp. 439–450.

Kalofolias, Janis, Mario Boley, and Jilles Vreeken (2019). “Discovering robustly connected subgraphs with simple descriptions”. In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 1150–1155.