Local Subgroup Discovery on Attributed Network Graphs

Carl Vico Heinrich, Tommie Lombarts, Jules Mallens, Luc Tortike, David Wolf, and Wouter Duivesteijn (✉)

Technische Universiteit Eindhoven, Eindhoven, the Netherlands
{c.v.heinrich,t.lombarts,j.m.mallens,l.m.tortike,d.wolf}@student.tue.nl, w.duivesteijn@tue.nl

Abstract. We find locally exceptional subgroups of nodes in attributed graphs, combining both node attributes and structural information to assess subgroup exceptionality. Subgroups are locally exceptional if their behavior deviates from the behavior of a well-chosen local peer group, as opposed to the more common generic subgroup discovery approach where behavior is compared to the global behavior across the full dataset. This notion of Local Subgroup Discovery had been introduced for traditional flat-table data; to the best of our knowledge, we are the first to incorporate this notion explicitly in graph data. Our approach combines shortest-path distance with Gower's Distance, integrating both network structure and node attributes to rank nodes in relation to a prototype node. Combining this notion of exceptionality with existing LSD techniques, we discover local subgroups in three attributed graph datasets.

Keywords: Subgroup Discovery · Local Subgroup Discovery · Network Analysis · Attributed Graphs · Graph Distances · Pattern Mining.

1 Introduction

The increasing availability of large-scale network data [35, 41], combined with advances in analytical techniques [12], has enabled researchers to explore interesting patterns within various types of networks [11], such as biological [39], communication [2], or social networks [6]. Networks provide valuable insights not only into global graph structures but also into local phenomena and relations within the graph. Community structures, or clusters, are especially valuable in these analyses, as they reflect meaningful groupings [19].

One powerful method for uncovering these insights is Subgroup Discovery (SD) [29], which identifies subsets of nodes that deviate from expected norms. In traditional SD, methods typically compare characteristics of the subgroups to global characteristics [7, 32]. This approach works well in environments where the global norm or tendency is clear. In many real-world networks, nodes represent entities with a variety of characteristics and behaviors [11]. In such heterogeneous settings, comparing the subgroups to the entire network can obscure relevant patterns [13, 19].

We propose a framework for applying Local Subgroup Discovery (LSD) [30] to general network structures. In LSD, the interestingness of a subgroup is not compared to the global population but rather to a reference group, in which people share similar characteristics. Thus, rather than finding exceptional nodes or behavior within a graph concerning the global setting, we define subgroups based on their relation to local reference groups. This localized approach allows the discovery of subgroups whose traits differ from expected behavior compared to their peers, capturing both structural relationships and node attributes.

We introduce a novel application of LSD to the domain of network analysis, combining both node attributes and structural information to identify exceptional local subgroups. We contribute a framework adaptable to diverse quality and distance measures, solving ties according to the node attributes.

2 Related Work

2.1 Subgroup Discovery, Local Patterns, Subgraphs

Subgroup Discovery (SD) seeks subgroups with respect to a variable of interest [7]. A quality measure evaluates the interestingness of a subgroup compared to the overall dataset [24]. In Local Subgroup Discovery (LSD) [30], the quality of a subgroup is instead compared to a subset of the data: the reference group. This method makes use of a ranking, often generated using a distance measure, to identify relevant reference groups. Different data structures will require different methods for creating a ranking and quality measure. Social network data is often structured as an attributed graph. For attributed network data, traditional SD techniques exist in literature [8], but LSD techniques do not.

Exception Rules [47, 48] seek local patterns that deviate w.r.t. a local context, specifically: an association rule with a single item on the right-hand side, forming an exception to a more general common sense rule. Similarly, Anomalous Association Rules [10] seek association rules hidden by a dominant rule. None of these papers concern themselves with graph data. Conversely, the work on Exceptional Contextual Subgraph Mining [28] does discover interesting subgraphs defined in terms of graph attributes (the "context"), but subgraph behavior is compared to behavior on the entire graph as opposed to a local reference group.

2.2 Network Graphs

Network graphs are widely used in data mining to represent complex systems, such as social, biological, and communication networks, where entities are modeled as nodes and relationships as edges [40]. One key feature of network graphs is their ability to capture both local and global structural properties of the data, allowing researchers to study interactions at various scales [16]. The study of attributed networks is a specific extension of traditional graph theory, where nodes and/or edges have associated attributes (e.g., demographic information in social networks or molecular characteristics in biological networks) [52].

In recent years, significant attention has been given to mining interesting patterns from attributed networks, particularly in the context of subgroups with shared structural and attribute-based characteristics [8]. Such methods often focus on attributes of nodes and edges alone, such as an attributed 'to' and 'from' approach to edges, using these attributed edges as datapoints [26]. As data becomes more complex, integrating network structure into the analysis becomes increasingly important for finding novel patterns [1].

2.3 EgoNets

Ego Networks (EgoNets) represent a specific perspective in network analysis, focusing on a central node (the "ego") and its immediate neighbors (the "alters"). The structure of EgoNets is especially useful for studying local properties and immediate environments of nodes, allowing exploration of network dynamics at the micro-level [17]. EgoNets have been widely used in Social Network Analysis (SNA) [4, 36, 37], where relationships between individuals or entities are key areas of analysis. In social networks, EgoNets enable the examination of personal networks, including measures such as size, density, and centrality [20].

EgoNets provide insights into local connectivity and the specific role of the ego within its surrounding network, often enriched with attribute data to assess relationships at both the structural and attribute levels [37]. The relevance of EgoNets beyond SNA, for instance in LSD which considers both node attributes and graph structure, remains underexplored [8]. While EgoNets focus on local structures centered on the ego and its immediate connections [5], they do not encompass the broader pattern discovery that LSD entails, aimed at identifying recurring substructures across the network [49].

3 Methods

A *network* is represented as a graph $G = (V, E)$, where V is the set of nodes (vertices) and $E \subseteq V \times V$ is the set of edges between nodes. Each node $v \in V$ may have associated attributes $\boldsymbol{a}(v) = \{a_1(v), \ldots, a_h(v)\}$ and a binary target variable $t(v) \in \{0, 1\}$. A *subgroup* is defined as any subset of nodes $S \subset V$. Our methods require a distance measure $d : V \times V \to \mathbb{R}^+$ which quantifies the distance between any two nodes $u, v \in V$, to be instantiated in Section 3.1.

A *prototype* is any node $x \in V$ that serves as the center for defining a subgroup S. Given prototype x we define two parameters: $\sigma, \rho \in \mathbb{N}$, where $\sigma < \rho$. The *local subgroup* S_σ consists of the σ nearest neighbors to x, according to the distance measure d. The *reference group* R_ρ contains the ρ nearest neighbors to x. Therefore, by construction, $S_\sigma \subset R_\rho$. The goal is to find subgroups R_ρ and S_σ such that the distribution of the target variable t within these subgroups reveals distinct local patterns. Specifically, we seek to identify R_ρ and S_σ where:

1. the target distribution within R_ρ differs from the entire graph G;
2. the target distribution within S_σ differs from the reference group R_ρ.

To gauge the interestingness of local subgroups S_σ, a *quality measure* $q \in \mathbb{R}$ assigns a numerical value to any subset of nodes $S \subset V$. Given a *prototype* x and a distance measure d, we obtain a ranking of all nodes $v_i \in V$ based on their distance to x. This ranks the target values $t(v_i)$ for the nodes:

$$\boldsymbol{r}(x) = \{t(v_1),\, t(v_2),\, \ldots,\, t(v_\sigma),\, \ldots,\, t(v_\rho),\, \ldots\}$$

where the nodes v_i are ordered such that:

$$d(x, v_1) \leq d(x, v_2) \leq \ldots \leq d(x, v_\sigma) \leq \ldots \leq d(x, v_\rho) \leq \ldots$$

where $S_\sigma = \{x, v_1, v_2, \ldots, v_\sigma\}$ and $R_\rho = \{x, v_1, v_2, \ldots, v_\sigma, \ldots, v_\rho\}$.

The objective, in the end, is to identify pairs (R, S), such that q is maximized, indicating that subgroup S is exceptional compared to its reference group R.

3.1 Distance Measure - Shortest Path Distance

We propose to use the shortest-path distance, denoted by d_{uv}, which quantifies the distance between any two nodes $u, v \in V$. The shortest-path distance between nodes u, v is defined as the minimum number of edges that need to be traversed from node u to v:

$$d_{uv} = \min_{P_{uv}} \left(\sum_{(u,v) \in P_{uv}} w(u, v) \right)$$

where P_{uv} represents the set of all possible paths between u and v, and $w(u, v)$ represents the weight of the edge between nodes u and v.

Several algorithms exist for finding the shortest path in a social network graph, such as the Bellman-Ford algorithm [9, 18], Breadth-First Search [38], Dijkstra's algorithm [15], and the A* algorithm [22]. We choose to compute the shortest path distance through Dijkstra's algorithm.

3.2 Distance Tie-breaker - Gower's Distance

When identifying reference groups, the shortest path distance frequently assigns ties to multiple nodes. We use node attributes to break these ties, applying Gower's Distance [21]. Since each node $v \in V$ not only has edges but also an associated set of attributes $\boldsymbol{a}(v)$, we can evaluate the similarity between the prototype x and any node v. Gower's Distance is particularly useful here, as it accommodates various types of attributes, including binary, categorical, ordinal, and continuous data. For any two nodes $u, v \in V$, Gower's Distance $d(u, v)$ is computed by considering the differences across all node attributes:

$$d(u, v) = \frac{1}{h} \sum_{k=1}^{h} \delta_k(u, v)$$

where h is the attribute dimensionality, and $\delta_k(u, v)$ is the partial distance between nodes u and v with respect to attribute k.

For *numerical* attributes, the partial distance $\delta_k(u, v)$ is computed as the scaled absolute difference:

$$\delta_k(u, v) = \frac{|a_k(u) - a_k(v)|}{\max_{x \in V} \big(a_k(x)\big) - \min_{x \in V} \big(a_k(x)\big)}$$

with the value $a_k(x)$ being replaced with the rank of the attribute for node x in case of *ordinal* attributes, as proposed by [27, pp. 35–36].

For *categorical* attributes, the distance is binary:

$$\delta_k(u, v) = \mathbb{1}_{\{a_k(u) \neq a_k(v)\}}$$

Gower's Distance allows us to integrate both node connectivity and node attributes into the subgroup detection process. Gower's Distance is only applied when nodes are equidistant according to the primary distance measure.

Whilst Gower's Distance is not guaranteed to break all ties, such as with 2 nodes with identical attributes at an identical distance from the prototype node, it does become significantly better at breaking ties for data with more features, specifically features that have a large domain size. This is because an increase in dimensions and feature domain size adds more unique sets of attributes, and thus more unique possibilities arise. In the case that Gower's Distance still results in a tie, the nodes will be added to the group in random order (based on their node names in the original graph).

3.3 Quality Measure

To quantify the exceptionality of the local subgroup S relative to its reference group R, we employ Weighted Kullback-Leibler divergence (WKL) [34] as our quality measure $q(S, R)$. WKL is an adaptation of Kullback-Leibler (KL) divergence [31], a fundamental concept in information theory [45]. It measures how one probability distribution diverges from a second, reference distribution, with the key enhancement of weighting by subgroup size. KL divergence is widely used in many fields, including Subgroup Discovery [3, 33, 51], as a tool for comparing distributions. In this context, KL divergence evaluates the difference between the distributions of the target variable $t(x)$ within the subgroup S and the distribution of the same target variable in the reference group R.

This measure captures how distinct the subgroup S is in comparison to its reference group R, a critical aspect of LSD. Specifically, it quantifies how much the presence of the target variable differs between the two subgroups:

$$q(S, R) = \frac{|S|}{|R|} \cdot \sum_{y \in \{0,1\}} P_S(y) \log \left(\frac{P_S(y)}{P_R(y)} \right)$$

where $P_A(y) = \mathbb{P}\big(t(x) = y \mid x \in A\big)$.

Algorithm 1 Generate Distance Ranking

procedure RANKING(x, S)
  $D \leftarrow Dijkstra(x, S)$
  $R \leftarrow SortOn(R, D)$
  for d **in** *unique*(R) **do**
    $r, l \leftarrow idxmin(R, d), idxmax(R, d)$
    $gowerlist \leftarrow Gower(x, n_i)\ \forall\ n \in R[r : l]$
    $SortOn(R[r : l], gowerlist)$
  return R

Algorithm 2 Generate ρ and σ

procedure DISCOVERY($R \leftarrow ranking$)
  $\rho, \sigma, b \leftarrow 0, 0, 0$
  for i **in** *len*(R) **do**
    $q \leftarrow Quality(R[0 : i], S)$
    if $q \geq b$ **then**
      $b \leftarrow q$
      $\rho \leftarrow i$
  $b \leftarrow 0$
  for i **in** $[0 : \rho]$ **do**
    $q \leftarrow Quality(R[0 : i], R[0 : \rho])$
    if $q \geq b$ **then**
      $b \leftarrow q$
      $\sigma \leftarrow i$
  return ρ, σ

3.4 Local Subgroup Discovery

To discover our distance-based subgroups we build from a *prototype*, which can be any point x within our nodes V. The subgroup S_σ is then the σ closest neighbors to our prototype x and the reference group R_ρ the ρ closest neighbors. Firstly, we select a prototype x randomly from our total population V and store the attributes $[a_1(x), \ldots, a_h(x)]$. The shortest path and Gower's distance between nodes are used to create a ranking, such that the ranking can be used to easily compute quality features for different size references and local subgroups. To generate this ranking, data points are all sorted based on their distance to the prototype x using Dijkstra's algorithm [15], breaking ties with Gower's Distance.

Algorithm 1 describes how such a ranking is created using $Dijkstra(x, S)$ to generate a list of size $|S|$ with shortest path distances starting from x to all nodes in S. Here, $SortOn(A, B)$ sorts a list A based on the values in list B, and $idxmin(L, x)$ and $idxmax(L, x)$ find the lowest and highest index in list L where value x appears. We then traverse the ranked target variables and for every possible reference group size $i \in [0 : n]$ we find the reference group size ρ such that $q(R_\rho, G)$ is maximized. Then, we repeat the process of finding a local subgroup within said reference group for all subgroup sizes of $i \in [0 : \rho]$ to find the subgroup size σ such that $Quality(S_\sigma, R_\rho)$ is maximized (cf. Algorithm 2).

This is done by first iteratively generating a quality measure for every possible size of the reference group and storing the best size, and then generating a quality measure for every possible size of the local subgroup and generating their quality measure with regards to the reference group, storing the most effective one.

3.5 Top-k Postprocessing

Selecting two prototypes u, v such that the d_{uv} is very low causes nearly identical rankings due to the properties of the shortest path distance. Hence, subgroups of neighboring nodes overlap substantially. We postprocess the set of discovered subgroups, based on the Dice-Sørensen coefficient [14, 46] measuring the overlap between subgroups S_1 and S_2:

$$O(S_1, S_2) = \frac{2}{|S_1| + |S_2|} \sum_{n \in S_1} \mathbb{1}_{\{n \in S_2\}}$$

We set a threshold θ, as an upper bound of the maximum allowed overlap between subgroups. We sort all candidate subgroups in order of decreasing quality measure, considering them from the top down. While considering candidate subgroup S_i, we check whether $\exists_{j \in \{1,\ldots,i-1\}}$ s.t. $O(S_j, S_i) > \theta$; if so, subgroup S_i is discarded. We continue this process until k subgroups are retained, or until we run out of candidates.

4 Experimental Setup

We demonstrate our version of LSD on attributed graph data[1] by deploying it on three datasets: `OGBG-MolPCBA` [42], `Twitch PT` [44], and `WebKB Cornell` [43]. Each dataset offers other structural challenges and attributes, making them suitable for testing our method on various domains.

The datasets must meet several conditions to be considered. First, they need to be graph-based, with clearly defined nodes and edges. Second, each dataset must contain meaningful node attributes, ensuring the identification and interpretation of subgroups within the graph. Third, the datasets must be small enough to allow for computationally intensive calculations, as larger datasets can impose significant computational challenges. Thus, we aim to balance rich feature sets and complex structures with a manageable size. For the primary dataset, interpretability is essential. In the context of LSD, non-anonymized features are preferable to facilitate clear subgroup interpretation. However, finding datasets that do not contain sensitive information, and thus are not anonymized, can be challenging. We select the `OGBG-MolPCBA` dataset [42] as our primary dataset due to its non-anonymized features, ensuring interpretability. We augment these experiments with two further datasets where this interpretability is obscured by anonymization, `Twitch PT` [44] and `WebKB Cornell` [43], to demonstrate that our method generalizes beyond a single dataset.

[1] source code at `https://github.com/TUeEMM/LSD-ATNG` and [23]

Table 1: Features and Possible Values in `OGBG-MolPCBA` [50]

Feature	Possible Values
Atomic Number	1 to 118, misc
Chirality	CHI_UNSPECIFIED, CHI_TETRAHEDRAL_CW, CHI_TETRAHEDRAL_CCW, CHI_OTHER, misc
Degree	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, misc
Formal Charge	-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, misc
No. of Hydrogen	0, 1, 2, 3, 4, 5, 6, 7, 8, misc
Radical Electrons	0, 1, 2, 3, 4, misc
Hybridization	SP, SP2, SP3, SP3D, SP3D2, misc
Aromaticity	False, True
In a Ring	False, True

4.1 OGBG-MolPCBA (Primary Dataset)

The `OGBG-MolPCBA` dataset from the Open Graph Benchmark [42], introduced as part of a benchmark study on molecular machine learning [50], is selected as the primary dataset for this study. Each graph in this dataset represents a molecule, with nodes corresponding to atoms and edges corresponding to chemical bonds. The node features are 9-dimensional [42], describing various atomic properties such as atomic number, chirality, and hybridization [25].

The detailed atomic features of the `OGBG-MolPCBA` dataset enable the discovery of subgroups of molecules that exhibit unique patterns in their chemical structure, potentially leading to insights that could guide future molecular studies. This is critical for the field of drug discovery and molecular biology, making the dataset relevant for scientific applications. The clear interpretability of the node features was a significant factor in selecting this dataset as our primary focus, as it allows for a thorough understanding of the subgroups identified by the LSD algorithm. Another key reason for choosing this dataset is its size. With 41 127 molecular graphs, each containing an average of 25.5 nodes and 27.5 edges [42], the dataset is large enough to test the scalability of the LSD algorithm while remaining computationally manageable.

Table 1 describes the various atomic properties that are present in the dataset. The *atomic number* represents the number of protons in the atom (with a *misc* category for undefined atoms). This feature acts as the binary target variable in our model, which evaluates to true if the atomic number is 6 or greater. *Chirality* refers to the geometric arrangement of atoms. The *degree* describes the number of bonds an atom forms and the *formal charge* indicates the hypothetical charge of an atom. The *number of hydrogen* atoms attached to an atom varies from 0 to 8, and the number of *radical electrons* (unpaired electrons) ranges from 0 to 4. *Hybridization* describes how atomic orbitals combine to form bonds with other atoms. *Aromaticity* tells whether an atom is part of an aromatic ring, and *in a ring* indicates if the atom is part of any ring system.

4.2 Twitch PT

The `Twitch PT` dataset, introduced as part of a study on multi-scale attributed node embedding [44], is a social network graph representing Twitch streamers based in Portugal. The 1 912 nodes represent individual streamers, and the 64 510 edges represent their followership relationships. The target represents explicit language usage. The 128-dimensional node attributes represent anonymized embeddings based on the games played by the streamers.

This dataset is chosen for its dense social network structure, allowing us to test the LSD algorithm in a context where connections between individuals form tight followership clusters. While the features are anonymized, the network itself is rich in structural information, enabling the discovery of subgroups of streamers who exhibit similar followership behaviors or content preferences. Thus, the `Twitch PT` experiment complements experiments on the molecular data in `OGBG-MolPCBA` by showing how LSD can identify communities of streamers based on behavioral patterns rather than explicit node attributes.

4.3 WebKB Cornell

The `WebKB Cornell` dataset, introduced in a study on geometric graph convolutional networks [43], is another graph-based dataset with anonymized features. It represents the network of web pages within Cornell University. and consists of nodes representing individual web pages and edges representing hyperlinks between them. Each node has 1 703 features, which are bag-of-words representations of the web page's content.

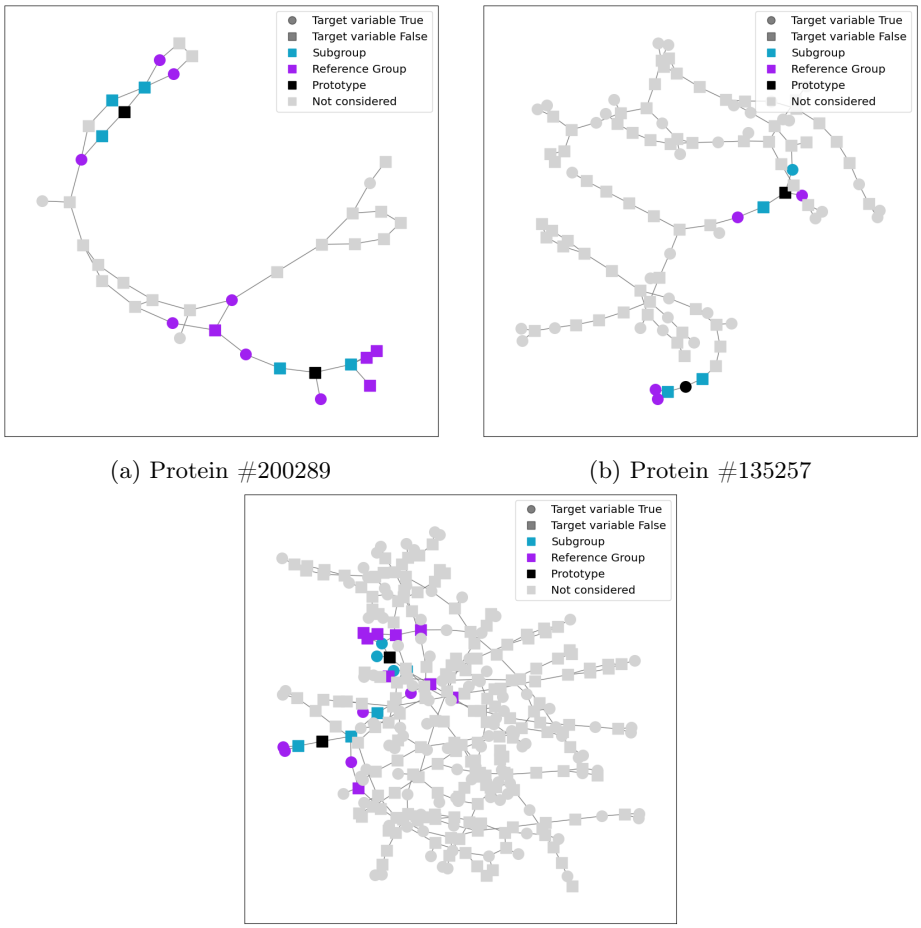
The high-dimensional bag-of-words features provide a detailed description of each web page's content, making this dataset ideal for testing the LSD algorithm's ability to handle textual and content-based node attributes. The `WebKB Cornell` dataset serves as a test case for how well LSD performs on a relatively small network with large feature vectors, demonstrating the algorithm's versatility in handling different types of graphs.

5 Experimental Results

5.1 OGBG-MolPCBA

Figure 1 displays the top-two non-redundant local subgroups (blue nodes), along with their respective prototypes (black node) and reference groups (purple+blue nodes), found in three molecules (sampled to vary in size). A node $v \in V$ is depicted as a circle if $t(v) = 1$, and as a square if $t(v) = 0$. We mostly find subgroups in less dense parts of the network.

To further investigate the resulting subgroups, we run our discovery algorithm on the largest protein in the dataset. Table 2 lists the top-10 subgroups discovered, with values for ρ, σ, q, and the accuracy of the binary target variable within R_ρ and S_σ. The subgroups display relatively high q values, also reflected by target variable accuracy differences. Values are near identical for



(a) Protein #200289 (b) Protein #135257

(c) Protein #182987

Fig. 1: Molecule graphs: each figure displays one protein, with the top-two non-redundant local subgroups and their respective reference groups.

the subgroups ranked $2-5$, which stands to reason: molecular structures can recur within a protein. If such a structure is an interesting subgroup, its 'copies' should also be found. Since the copies concern graph isomorphisms across possibly distinct sets of nodes, these copies will slip through our redundancy filtering (cf. Section 3.5) which merely identifies identical node sets.

5.2 Twitch PT

The top-five subgroups discovered on the `Twitch PT` dataset (cf. Table 3) display more variation in size. We find mostly local subgroups with a very low target variable distribution compared to that of their reference group. This shows that the

Table 2: Protein 349519 subgroups

Prot.	ρ	σ	q	**Acc.** R_ρ	**Acc.** S_σ
46	9	3	0.321	0.56	0.00
28	14	5	0.231	0.21	0.60
253	13	5	0.229	0.23	0.60
35	13	5	0.229	0.23	0.60
147	13	5	0.229	0.23	0.60
11	11	7	0.207	0.55	0.29
167	5	3	0.207	0.60	0.33
268	5	3	0.207	0.60	0.33
45	9	4	0.204	0.56	0.25
272	16	5	0.196	0.25	0.60

Table 3: `Twitch PT` subgroups

Prot.	ρ	σ	q	**Acc.** R_ρ	**Acc.** S_σ
385	9	4	0.278	0.67	0.25
1378	26	13	0.218	0.62	0.31
97	12	4	0.192	0.83	0.50
403	103	14	0.173	0.61	0.14
1032	220	58	0.164	0.49	0.17

Table 4: `WebKB Cornell` subgroups

Prot.	ρ	σ	q	**Acc.** R_ρ	**Acc.** S_σ
78	21	6	0.255	0.19	0.67
135	36	10	0.252	0.22	0.70
30	39	6	0.246	0.21	0.83
60	22	4	0.242	0.18	0.75
16	33	3	0.228	0.24	1.00

local subgroups are mostly groups of Twitch users that use no explicit language, within a network/reference group of streamers that do use explicit language.

5.3 WebKB Cornell

On the `WebKB Cornell` dataset, we find more subgroups with a high q score (cf. Table 4). Contrary to the `Twitch PT` dataset, here most subgroups have a significantly higher target variable accuracy than their respective reference groups, indicating a subgroup of web pages classified as project/staff within a reference group of papers that are mostly student/course/faculty related.

5.4 Ablation Study: Full Graph as Reference Group

The motivation of LSD is that subgroups can show interesting exceptional behavior when compared to a local reference group. Here, we perform an ablation study, observing whether the preceding experimental results change when we take $R_\rho = V$: if not, LSD provides no benefits over standard SD.

Setting $R_\rho = V$ consistently produces subgroups markedly different from and less effective than those discovered by LSD. On `OGBG-MolPCBA`, the top-10 ablation subgroups are fully disjoint (in covered nodes) from the top-10 LSD subgroups; node overlaps are 0.62% on `Twitch PT` and 8.99% on `WebKB Cornell`. Measuring discriminative power through WRAcc [7, Equation (1)], the LSD subgroups outperform the ablation subgroups (`OGBG-MolPCBA`: 0.215 vs. 0.0533; `Twitch PT`: 0.2349 vs. 0.0798; `WebKB Cornell`: 0.2211 vs. 0.1103). LSD subgroups cover substantially higher proportions of nodes v with $t(v) = 1$ than ablation subgroups on `OGBG-MolPCBA` (31.56% vs. 0.00%) and `WebKB Cornell` (81% vs. 48.6%), but lower on `Twitch PT` (28.3% vs. 63.84%). We conclude that

LSD provides substantial benefits on attributed network graphs over SD with $R_\rho = V$: discovered subgroups are different, more discriminative, and ultimately more valuable.

6 Conclusions

We propose a method to apply Local Subgroup Discovery (LSD) on attributed network graphs. Our method employs a network-based distance between nodes to find interesting subgroups, which have a target variable distribution that differs from their local reference group. We expand on previously established LSD methods by using network-based distances and attribute-based similarity for the creation of a subgroup ranking. The algorithm finds reference groups and subgroups with high quality measures by iteratively exploring the dataset. A top-k post-processing procedure is used to prune highly similar subgroups. We analyze the results of running the LSD algorithm on three datasets, showing that the algorithm identifies pairs of groups (R_ρ, S_σ) such that S_σ is exceptional to R_ρ with regards to our proposed quality feature q.

The introduction of a new distance-based ranking method enables the application of existing Subgroup Discovery techniques to network-structured data, paving the way for future research into using Subgroup Discovery or Exceptional Model Mining methods on graph structures. Future work could involve adapting this approach to handle more complex target data, including numerical or multivariate targets, enabling richer subgroup discoveries.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aggarwal, C. (ed.): Social Network Data Analytics. Springer, New York, NY (2011)
2. Ahmad, W., Hasan, O., Pervez, U., Qadir, J.: Reliability modeling and analysis of communication networks. Journal of Network and Computer Applications **78**, 191–215 (11 2016)
3. Arab, A., Arora, D., Lu, J., Ester, M.: Subgroup discovery in unstructured data. CoRR **abs/2207.07781** (2022)
4. Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F.: Analysis of ego network structure in online social networks. In: Proc. SocialCom-PASSAT. pp. 31–40 (2012)
5. Atzmueller, M.: Mining social media: key players, sentiments, and communities. WIREs Data Mining and Knowledge Discovery **2**(5), 411–419 (2012)
6. Atzmueller, M.: Data mining on social interaction networks. Journal of Data Mining and Digital Humanities **1** (6 2014)
7. Atzmueller, M.: Subgroup discovery. WIREs Data Mining and Knowledge Discovery **5**(1), 35–49 (1 2015)
8. Atzmueller, M.: Compositional subgroup discovery on attributed social interaction networks. In: Proc. DS. p. 259–275 (2018)

9. Bellman, R.: On a routing problem. Quarterly of Applied Mathematics **16**(1), 87–90 (1958)
10. Berzal, F., Cubero, J.C., Marín, N., Gámez, M.: Anomalous association rules. In: Proc. IEEE ICDM workshop on Alternative Techniques for Data Mining and Knowledge Discovery (2004)
11. Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., Zanin, M.: The structure and dynamics of multilayer networks. Physics Reports **544**(1), 1–122 (2014)
12. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports **424**(4), 175–308 (2006)
13. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining: The ASA Data Science Journal **4**(5), 512–546 (2011)
14. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**, 297–302 (1945)
15. Dijkstra, E.: A note on two problems in connexion with graphs. Numerische Mathematik **1**(1), 269–271 (Dec 1959)
16. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press (2010)
17. Everett, M., Borgatti, S.: Ego network betweenness. Social Networks **27**, 31–38 (01 2005)
18. Ford, L.: Network Flow Theory. RAND Corporation, Santa Monica, CA (1956)
19. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports **659**, 1–44 (Nov 2016)
20. Freeman, L.C.: Centered graphs and the structure of ego networks. Mathematical Social Sciences **3**(3), 291–304 (October 1982)
21. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics **27**(4), 857–871 (1971)
22. Hart, P., Nilsson, N., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics **4**(2), 100–107 (1968)
23. Heinrich, C.V., Lombarts, T., Mallens, J., Tortike, L., Wolf, D., Duivesteijn, W.: Local subgroup discovery on attributed network graphs (Feb 2025). https://doi.org/10.5281/zenodo.14919732
24. Herrera, F., Carmona, C., González, P., Del Jesus, M.: An overview on subgroup discovery: foundations and applications. Knowledge and Information Systems **29**(3), 495–525 (Nov 2010)
25. Hu, W., Milesi, A.: Ogb/ogb/utils/features.py at master · snap-stanford/ogb (2022), `https://github.com/snap-stanford/ogb/blob/master/ogb/utils/features.py`
26. Jorge, C.C., Atzmueller, M., Heravi, B.M., Gibson, J.L., de Sá, C.R., Rossetti, R.J.: Mining exceptional social behaviour. In: Proc. EPIA. pp. 460–472 (2019)
27. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
28. Kaytoue, M., Plantevit, M., Zimmermann, A., Bendimerad, A.A., Robardet, C.: Exceptional contextual subgraph mining. Mach. Learn. **106**(8), 1171–1211 (2017)
29. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI/MIT Press (1996)
30. Konijn, R., Duivesteijn, W., Kowalczyk, W., Knobbe, A.: Discovering local subgroups, with an application to fraud detection. In: Proc. PAKDD. pp. 1–12 (2013)

31. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1), 79 – 86 (1951)
32. Lavrac, N.: Subgroup discovery techniques and applications. In: Proc. PKDD. pp. 2–14 (05 2005)
33. van Leeuwen, M.: Maximal exceptions with minimal descriptions. Data Mining and Knowledge Discovery **21**(2), 259–276 (2010)
34. van Leeuwen, M., Knobbe, A.: Non-redundant subgroup discovery in large and complex data. In: Proc. ECML PKDD. pp. 459–474 (2011)
35. Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M.: Statistical properties of community structure in large social and information networks. In: Proc. WWW. pp. 695–704 (2008)
36. Leskovec, J., Mcauley, J.: Learning to discover social circles in ego networks. Neural Information Processing Systems **25**, 539–547 (12 2012)
37. Marsden, P.: Network data and measurement. Annual Review of Sociology **16**, 435–463 (1990)
38. Moore, E.: The shortest path through a maze. In: Proceedings of an international symposium on the theory of switching, Part II. pp. 285–292. Harvard University Press, Cambridge, MA (1959)
39. Muzio, G., O'Bray, L., Borgwardt, K.: Biological network analysis with deep learning. Briefings in Bioinformatics **22**(2), 1515–1530 (9 2020)
40. Newman, M.E.J.: The structure and function of networks. Computer Physics Communications **147**(1), 40–45 (2002)
41. Newman, M.E.J.: The structure and function of complex networks. SIAM Review **45**(2), 167–256 (Jan 2003)
42. Open Graph Benchmark: Graph property prediction (2024), `https://ogb.stanford.edu/docs/graphprop/`
43. Pei, H., Wei, B., Chang, K.C., Lei, Y., Yang, B.: Geom-GCN: Geometric graph convolutional networks. In: Proc. ICLR (2020)
44. Rozemberczki, B., Allen, C., Sarkar, R.: Multi-scale attributed node embedding (2021), `https://arxiv.org/abs/1909.13021`
45. Shlens, J.: Notes on Kullback-Leibler divergence and likelihood. CoRR **abs/1404.2000** (2014)
46. Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skrifter/Kongelige Danske Videnskabernes Selskab **5**, 1 (1948)
47. Suzuki, E.: Autonomous discovery of reliable exception rules. In: Proc. KDD. pp. 259–262 (1997)
48. Suzuki, E., Kodratoff, Y.: Discovery of surprising exception rules based on intensity of implication. In: Proc. PKDD. pp. 10–18 (1998)
49. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences, Cambridge University Press (1994)
50. Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A., Leswing, K., Pande, V.: Moleculenet: a benchmark for molecular machine learning. Chemical Science **9**, 513–530 (2018)
51. Xu, S., Walter, N.P., Kalofolias, J., Vreeken, J.: Learning exceptional subgroups by end-to-end maximizing KL-divergence. CoRR **abs/2402.12930** (2024)
52. Zhou, Y., Cheng, H., Yu, J.: Graph clustering based on structural/attribute similarities. Proc. VLDB **2**, 718–729 (08 2009)