Explainable Subgraphs with Surprising Densities: A Subgroup Discovery Approach*

Junning Deng † Bo Kang † Jefrey Lijffijt † Tijl De Bie †

Abstract

The connectivity structure of graphs is typically related to the attributes of the nodes. In social networks for example, the probability of a friendship between two people depends on their attributes, such as their age, address, and hobbies. The connectivity of a graph can thus possibly be understood in terms of patterns of the form 'the subgroup of individuals with properties X are often (or rarely) friends with individuals in another subgroup with properties Y'. Such rules present potentially actionable and generalizable insights into the graph. We present a method that finds pairs of node subgroups between which the edge density is interestingly high or low, using an information-theoretic definition of interestingness. This interestingness is quantified subjectively, to contrast with prior information an analyst may have about the graph. This view immediately enables iterative mining of such patterns. Our work generalizes prior work on dense subgraph mining (i.e. subgraphs induced by a *single* subgroup). Moreover, not only is the proposed method more general, we also demonstrate considerable practical advantages for the single subgroup special case.

Keywords

Graph mining, Subgroup Discovery, Subjective interestingness, Community detection

1 Introduction

Real-life graphs (*aka* networks) often contain attributes for the nodes. In social networks for example, nodes correspond to individuals and node attributes can include their age, address, hobbies, etc. A network's connectivity is usually related to those attributes: individuals' attributes affect the likelihood of them meeting, and, if they meet, of becoming friends. Hence, to a certain extent, it should be possible to understand the connectivity of a graph in terms of those attributes.

One approach to identify the relations between the connectivity and the attributes is to train a link prediction classifier, with as input the attribute values for a pair of nodes, and predicting the edge as present or absent. Such global models often fail to provide insight though, much like a global classifier on any data type may fail to provide insight in other classification problems. To address this, the local pattern mining community introduced the concept of *subgroup discovery*, which aims to identify subgroups of data points for which a target attribute has homogeneous and/or outstanding values. Subgroups are local patterns, in that they provide information only about a certain part of the data.

Research on local pattern mining in attributed graphs has so far focused on identifying dense node-induced subgraphs, dubbed *communities*, that are coherent also in terms of attributes. There are two complementary approaches. The first explores the space of communities that meet certain criteria in terms of density, in search for those that are homogeneous. The second explores the space of rules over the attributes, in search for those that define subgroups of nodes that form a dense community. This is effectively a subgroup discovery approach to dense subgraph mining.

Limitations of the state-of-the-art. Both of these approaches make use of attribute homophily: the tendency of links to exist between nodes sharing similar attributes. While the homophily assumption is often reasonable, it also limits the scope of application of prior work to finding dense communities with homogeneous attributes. A *first limitation* of the state-of-the-art is thus its inability to find e.g. sparse subgraphs.

A *second limitation* is that the interestingness of such patterns has invariably been quantified by objective measures—i.e. measures independent of the data analyst's prior knowledge. Yet, the most 'interesting' patterns found are often obvious and implied by such prior knowledge (e.g. communities involving high-degree nodes, or in a student friendship network, communities involving individuals practicing the same sport), making them subjectively uninteresting.

*This research was funded by the ERC under the EU's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement no. 615517, the Flemish government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme, the FWO (project no. G091017N, G0F9816N), and from the EU's Horizon 2020 research and innovation programme and the FWO under the Marie Sklodowska-Curie Grant Agreement no. 665501.

†IDLab, Ghent University; Firstname.Lastname@UGent.be

A *third limitation* of prior work is that the patterns describe only the connectivity *within* communities and not *between* subgroups of nodes. As an obvious example, this excludes patterns that describe friendships between a particular subgroup of female and a subgroup of male individuals in a social network. The experiments on real-life networks contain many less obvious examples.

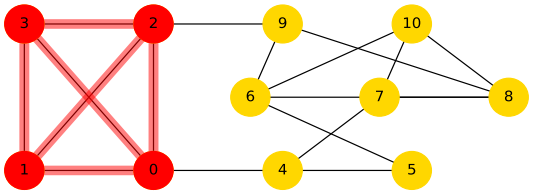
Contributions. We depart from the existing literature in formalizing a *subjective* interestingness measure, building on the ideas from the FORSIED framework [4], and this for *sparse* as well as for *dense* subgraph patterns. In this way, we overcome the first and second limitations of prior work discussed above. Moreover, this interestingness measure is naturally applicable for patterns describing the graph density between a pair of subgroups, to which we will refer as *bi-subgroup patterns*. Hence, our method overcomes the third limitation of prior work. Our specific contributions are: (1) Novel definitions of single-subgroup patterns and bi-subgroup patterns [Sec. 2]. (2) A formalization of *Subjective Interestingness* (SI), based on the analyst's evolving prior beliefs [Sec. 3]. (3) A beam-search algorithm to mine the subjectively most interesting bi-subgroup patterns [Sec. 4]. (4) An empirical evaluation on real-world data, confirming our method's ability to identify subjectively interesting patterns [Sec. 5].

2 Subgroup pattern syntaxes for graphs

This section formalizes single-subgroup and bi-subgroup patterns for graphs, beginning with some notation.

An attributed graph is denoted as a triplet $G = (V, E, A)$ where V is a set of $n = |V|$ vertices, and $E \subseteq \binom{V}{2}$ is a set of $m = |E|$ undirected edges[1], and A is a set of attributes $a \in A$ defined as functions $a : V \to \text{Dom}_a$, where Dom_a is the set of values the attribute can take over V. For each attribute $a \in A$ with nominal Dom_a and for each $y \in \text{Dom}_a$, we introduce a Boolean function $s_{a,y} : V \to \{\text{true}, \text{false}\}$, defined as true for $v \in V$ iff $a(v) = y$. Analogously, for each $a \in A$ with real-valued Dom_a and for each $l < u$ and $l, u \in \text{Dom}_a$, we define $s_{a,[l,u]} : V \to \{\text{true}, \text{false}\}$, with $s_{a,[l,u]}(v) \triangleq$ true iff $a(v) \in [l, u]$. We call these functions *selectors*, and denote the set of all selectors as S. A *description* or *rule* W is a conjunction of a subset of selectors: $W = s_1 \wedge s_2 \ldots \wedge s_{|W|}$. The *extension* $\varepsilon(W)$ of a rule W is defined as the subset of vertices that satisfy it: $\varepsilon(W) \triangleq \{v \in V | W(v) = \text{true}\}$. We informally also refer to the extension as the *subgroup*. Now a

[1] We consider undirected graphs without self-edges for the sake of presentation and consistency with most literature. However, all our results can be easily extended to directed graphs and graphs with self-edges.



(a) Graph

Vertex	0	1	2	3	4	5	6	7	8	9	10
a	3.5	2.6	3.8	3.2	1.8	1.2	5.4	0.9	6.7	2.3	3.1
b	1	1	1	1	1	0	0	1	0	0	0
c	0	0	1	0	0	0	1	1	1	1	1
d	1	0	1	1	1	0	0	0	0	1	0

(b) Vertex attributes

Figure 1: Example attributed graph with 11 vertices (0-10) and 4 associated attributes (*a*-*d*). The subgraph induced by the description ($W = s_{a,[2,4]} \wedge s_{b,1}$) is highlighted in red.

description-induced subgraph can be formally defined as:

DEFINITION 1. (Description-induced-subgraph) *Given an attributed graph $G = (V, E, A)$ and a description W, we say that a subgraph $G[W] = (V_W, E_W, A)$ where $V_W \subseteq V, E_W \subseteq E$, is induced by W if:*

(i) $V_W = \varepsilon(W)$, i.e., the set of vertices from V that is the extension of the description W, and

(ii) $E_W = (V_W \times V_W) \cap E$, i.e., the set of edges from E that have both endpoints in V_W.

EXAMPLE 1. *Fig. 1(a) displays an example attributed graph with 11 vertices, 18 edges. Each node is annotated with 1 real-valued attribute (a) and 3 binary attributes (b, c, d). Consider a description $W = s_{a,[2,4]} \wedge s_{b,1}$. The extension of this description is the set of nodes with attribute a value from 2 to 4 and attribute b as 1, i.e., $\varepsilon(W) = \{0, 1, 2, 3\}$. The subgraph induced by W is formed from $\varepsilon(W)$ and all the edges connecting pairs of vertices in that set (highlighted with red in Fig. 1(a)).*

2.1 Single-subgroup pattern A first pattern syntax we consider informs the analyst about the density of a description-induced subgraph $G[W]$. We assume the analyst is satisfied by knowing whether the density is unusually small, or unusually large, and given this does not expect to know the precise density. It thus suffices for the pattern syntax to indicate whether the density is either smaller than, or larger than, a specified value. We thus formally define the *single-subgroup* pattern syntax as a triplet (W, I, k_W), where W is a description and $I \in \{0, 1\}$ indicates whether the number of edges E_W in

subgraph $G[W]$ induced by W is greater (or less) than k_W. Thus, $I = 1$ indicates the induced subgraph is sparse, whereas $I = 0$ characterizes a dense subgraph. The maximum number of edges in $G[W]$ is denoted by n_W, equal to $\frac{1}{2}|\varepsilon(W)|(|\varepsilon(W)| - 1)$.

2.2 Bi-subgroup pattern We also define a pattern syntax informing the analyst about the edge density between two different subgroups. More formally, we define a *bi-subgroup pattern* as a quadruplet (W_1, W_2, I, k_W), where W_1 and W_2 are two descriptions, and $I \in \{0, 1\}$ indicates whether the number of connections between $\varepsilon(W_1)$ and $\varepsilon(W_2)$ is upper bounded (1) or lower bounded (0) by the threshold k_W. The maximum number of connections between the extensions $\varepsilon(W_1)$ and $\varepsilon(W_2)$ is denoted by $n_W \triangleq |\varepsilon(W_1)||\varepsilon(W_2)| - \frac{1}{2}|\varepsilon(W_1 \wedge W_2)|(|\varepsilon(W_1 \wedge W_2)| + 1)$. Note that single-subgroup patterns are a special case of bi-subgroup patterns when $W_1 \equiv W_2$.

REMARK 1. *Although k_W for a pattern (W_1, W_2, I, k_W) can be any value with which the number of connections between $\varepsilon(W_1)$ and $\varepsilon(W_2)$ (or within $\varepsilon(W_1)$ when $W_1 \equiv W_2$) are bounded, our work focus on identifying patterns whose k_W is the actual number of connections between these two subgroups (or within this single subgroup when $W_1 \equiv W_2$), as such patterns are maximally informative.*

3 Formalizing the subjective interestingness

Previous work on mining patterns in attributed graphs focuses on identifying dense communities, with *density* quantified in an objective way (see Sec. 6). However, given prior information on the graph, the resulting patterns may be trivial, containing limited information that is novel to the analyst. Tackling this necessitates the use of subjective measures of interestingness.

3.1 General approach We follow the approach as outlined by De Bie [5] to quantify the SI of a pattern. In this framework, the analyst's belief state is modeled by a *background distribution* over the data space. This background distribution represents any prior beliefs the analyst may have by assigning a probability (density) to each possible value for the data according to how plausible the analyst thinks this value is. As such, the background distribution also makes it possible to assess the surprise in the analyst when informed about the presence of a pattern. It was argued that a good choice for the background distribution is the maximum entropy distribution subject to some particular constraints that represent the analyst's prior beliefs about the data. As the analyst is informed about a pattern, the knowledge about the data will increase, and the background distribution will change. For details see Sec. 3.2.

Given a background distribution, the SI of a pattern can be quantified as the ratio of the *Information Content* (IC) and the *Description Length* (DL) of a pattern. The IC is defined as the amount of information gained when informed about the pattern's presence, computed as the negative log probability of the pattern w.r.t. the background distribution P. The DL quantifies the code length needed to communicate the pattern to the analyst. These are discussed in more detail in Sec. 3.3, but first we further explain the background distribution.

3.2 The background distribution

The initial background distribution Here we recapitulate how prior beliefs of the following types can be modelled in a background distribution: (i) on individual vertex degrees; (ii) on the overall graph density; (iii) on densities between bins.

Type (i) and (ii): Prior beliefs on individual vertex degrees and on the overall graph density. Given prior beliefs about the degree of each vertex, the maximum entropy distribution is a product of independent Bernoulli distributions, one for each of the random variable $h_{u,v}$ defined as 1 if $(u, v) \in E$ and 0 otherwise [5]. Denoting the probability that $h_{u,v} = 1$ by $p_{u,v}$, this distribution is of the form:

$$P(E) = \prod_{u,v} {p_{u,v}}^{h_{u,v}} \cdot (1 - p_{u,v})^{1-h_{u,v}},$$

$$\text{where} \quad p_{u,v} = \frac{\exp(\lambda_u^r + \lambda_v^c)}{1 + \exp(\lambda_u^r + \lambda_v^c)}.$$

This can be conveniently expressed as:

$$P(E) = \prod_{u,v} \frac{\exp((\lambda_u^r + \lambda_v^c) \cdot h_{u,v})}{1 + \exp(\lambda_u^r + \lambda_v^c)}.$$

The parameters λ_u^r and λ_v^c can be computed efficiently. For a prior belief on the overall density, every edge probability $p_{u,v}$ simply equals the assumed density.

Type (iii): Additional prior beliefs on densities between bins. We can partition nodes in an attributed graph into bins according to their value for a particular attribute. For example, nodes representing people in a university social network can be partitioned by class year. Then expressing prior beliefs regarding the edge density between two bins is possible. This would allow the data analyst to express, for example, an expectation about the probability that people in class year y_1 is connected to those in class year y_2. If the analyst believes that people in different class years are less likely to connect with each other, the discovered pattern would end up being more informative and useful as it contrasts more with this kind of belief. As

shown by Adriaens et al. [1], the resulting background distribution is also a product of Bernoulli distributions, one for each of the random variable $h_{u,v} \in \{0,1\}$):

$$P(E) = \prod_{u,v} \frac{\exp((\lambda_u^r + \lambda_v^c + \gamma_{k_{u,v}}) \cdot h_{u,v})}{1 + \exp(\lambda_u^r + \lambda_v^c + \gamma_{k_{u,v}})},$$

where $k_{u,v}$ indexes the block formed by the intersecting part of two bins which vertex u and v belongs to correspondingly, λ_u^r ,λ_v^c and $\gamma_{k_{u,v}}$ are efficiently computable parameters. Note that the background distribution can model a prior belief simultaneously for the edge densities between bins resulting from multiple partitions.

Updating the background distribution Upon being represented with a pattern, the background distribution should be updated to reflect the data analyst's newly acquired knowledge. The beliefs attached to any value for the data that does not contain the pattern should become zero. In the present context, once we present a pattern (W_1, W_2, I, k) to the analyst, the updated background distribution P' should be such that $\phi_W(E) \geq k_W$ (if $I = 0$) or $\phi_W(E) \leq k_W$ (if $I = 1$) holds with probability one, where $\phi_W(E)$ denotes a function counting the number of edges between $\varepsilon(W_1)$ and $\varepsilon(W_2)$. By De Bie [4], it was argued to choose P' as the *I-projection* of the previous background distribution onto the set of distributions consistent with the presented pattern. Then Van Leeuwen et al. [20] showed that the resulting P' is again a product of Bernoulli distribution:

$$P'(E) = \prod_{u,v} {p'_{u,v}}^{h_{u,v}} \cdot (1 - p'_{u,v})^{1-h_{u,v}}$$

$$\text{where} \quad p'_{u,v} = \begin{cases} p_{u,v} & \text{if} \quad \neg\big(u \in \varepsilon(W_1), v \in \varepsilon(W_2)\big), \\ \frac{p_{u,v} \cdot \exp(\lambda_W)}{1 - p_{u,v} + p_{u,v} \cdot \exp(\lambda_W)} & \text{otherwise.} \end{cases}$$

How to compute λ_W is also given in [20].

3.3 The subjective interestingness measure

The Information Content (IC). Given a pattern (W_1, W_2, I, k_W), and a background distribution defined by P, the probability of the presence of the pattern is the probability of getting k_W or more (for $I = 0$), or fewer than k_W (for $I = 1$) successes in n_W trials with possibly different success probabilities $p_{u,v}$. While it is impractical to compute these probabilities exactly, using the same approach as Van Leeuwen et al. [20] they can be tightly upper bounded using the general Chernoff/Hoeffding bound [10], as follows:

$$\mathbf{Pr}[(W_1, W_2, I = 0, k_W)] \leq \exp\Bigg(-n_W \mathbf{KL}\bigg(\frac{k_W}{n_W} \parallel p_W\bigg)\Bigg),$$

$$\mathbf{Pr}[(W_1, W_2, I = 1, k_W)] \leq \exp\Bigg(-n_W \mathbf{KL}\bigg(1 - \frac{k_W}{n_W} \parallel 1 - p_W\bigg)\Bigg),$$

where $p_W = \frac{1}{n_W} \sum_{u \in \varepsilon(W_1), v \in \varepsilon(W_2)} p_{u,v}$.

$\mathbf{KL}\Big(\frac{k_W}{n_W} \parallel p_W\Big)$ is the Kullback-Leibler divergence between two Bernoulli distribution with success probabilities $\frac{k_W}{n_W}$ and p_W respectively. Note that:

$$\begin{aligned} \mathbf{KL}\big(\frac{k_W}{n_W} \parallel p_W\big) &= \mathbf{KL}\big(1 - \frac{k_W}{n_W} \parallel 1 - p_W\big), \\ &= \frac{k_W}{n_W} \log\big(\frac{k_W/n_W}{p_W}\big) + \\ &\quad \big(1 - \frac{k_W}{n_W}\big) \log\big(\frac{1 - k_W/n_W}{1 - p_W}\big). \end{aligned}$$

We can thus write:

$$\mathbf{Pr}[(W_1, W_2, I, k_W)] \leq \exp\Bigg(-n_W \mathbf{KL}\bigg(\frac{k_W}{n_W} \parallel p_W\bigg)\Bigg).$$

The IC is the negative log probability of the pattern being present under the background distribution:

$$\begin{aligned} \mathrm{IC}[(W_1, W_2, I, k_W)] &= -\log(\mathbf{Pr}[(W_1, W_2, I, k_W)]), \\ &\geq n_W \mathbf{KL}\bigg(\frac{k_W}{n_W} \parallel p_W\bigg). \end{aligned} \tag{3.1}$$

The Description Length (DL). A pattern with larger IC is more informative. Yet, sometimes it is harder for the analyst to assimilate as its description is more complex. A good SI measure should trade off IC with DL. The DL should capture the length of the description needed to communicate a pattern. Intuitively, the cost for the data analyst to assimilate a description W depends on the number of selectors in W, i.e., $|W|$. Let us assume communicating each selector in a description W has a constant cost of α and the cost for I and k_W is fixed. The total description length of a pattern (W_1, W_2, I, k_W) can be written as[2]:

$$\mathrm{DL}[(W_1, W_2, I, k_W)] = \alpha(|W_1| + |W_2|) + \beta. \tag{3.2}$$

The Subjective Interestingness (SI). Putting the IC and DL together finally yields the SI:

$$\begin{aligned} \mathrm{SI}[(W_1, W_2, I, k_W)] &= \frac{\mathrm{IC}[(W_1, W_2, I, k_W)]}{\mathrm{DL}[(W_1, W_2, I, k_W)]}, \\ &= \frac{n_W \mathbf{KL}\Big(\frac{k_W}{n_W} \parallel p_W\Big)}{\alpha(|W_1| + |W_2|) + \beta}. \end{aligned} \tag{3.3}$$

[2] In all our experiments, we use $\alpha = 0.3, \beta = 0.5$

4 Algorithm

This section describes the algorithm for obtaining a set of interesting patterns. Since the proposed SI interestingness measure is more complex than most objective measures, heuristic search strategies are inevitable for tractability, as described next.

4.1 Beam search For mining single-subgroup patterns, we applied a classical heuristic search strategy over the space of descriptions—the beam search. The general idea is to only store a certain number (called the *beam width*) of best partial description candidates of a certain length (number of selectors) according to the SI measure, and to expand those next with a new selector. This is then iterated. This approach is standard practice in subgroup discovery (used e.g. in Cortana [12] and pysubgroup [11]).

4.2 Nested beam search To search for the bi-subgroup patterns, however, a traditional beam search over both W_1 and W_2 simultaneously turned out to be more difficult to apply effectively: beams large enough for good quality results turned out to be too demanding. Instead, a nested beam search strategy, where one beam search is nested into the other, gives good results. Here, the outer beam search explores promising selectors for the description W_1, and the inner beam search expands those for W_2. Let us denote the width of the outer and inner beam by x_1 and x_2 respectively. The total number of patterns identified by our algorithm is $x_1 \cdot x_2$. To maintain a sufficient diversity among the discovered patterns, we constrain the outer beam to contain at least x_1 different W_1 descriptions. Further details are given in the supplement [6].

4.3 Implementation The implementation builds on *Pysubgroup* [11], a Python package for subgroup discovery implementation. We integrated our nested beam search algorithm and SI measure into this original interface. A Python implementation of the algorithms and the experiments is available.[3] All experiments were conducted on a PC running Ubuntu with i7-7700K 4.20GHz CPU and 32 GB of RAM.

5 Experiments

We evaluate our methods on three real-world networks. In the following, we first describe the datasets (Sec. 5.1). Then we discuss the properties of the discovered patterns (single-subgroup patterns in Sec. 5.2 and bi-subgroup patterns in Sec. 5.3), with a purpose to evaluate various aspects of our proposed SI measure. In addition, scalability evaluation for both cases is given.

[3] `https://bitbucket.org/ghentdatascience/essd_public`

Table 1: Dataset statistics summary

Dataset	Type	$\|V\|$	$\|E\|$	#Attributes	$\|S\|$
Caltech36	undirected	762	16651	7	602
Reed98	undirected	962	18812	7	748
Lastfm	undirected	1892	12717	11946	23892
DblpAffs	directed	6472	3066	116	232

5.1 Data For our experiments we used four datasets. Data size statistics are given in Table 1.

***Caltech36* and *Reed98*.** Two Facebook social networks from the Facebook100 [18] data set, gathered in September 2005: one for Caltech Facebook users, and one for Reed University. Node attributes describe the person's status (faculty or student), gender, major, minor, dorm/house, graduation year, and high school.

***Lastfm*.** [3] A social network generated from friendships between `Lastfm.com` users. A list of most-listened musical artists and tag assignments for each user is given in [user, tag, artist] tuples. We took the tags that a user ever assigned to any artist and assigned those to the user as binary attributes expressing a user's music interests.

***DblpAffs*.** A DBLP[4] citation network based on a random subset of publications from 20 conferences[5] selected to cover 4 research areas: Machine Learning, Database, Information Retrieval, and Data Mining. Only papers for which the authors' country (or state, in the USA) of affiliation is available are included. The resulting 116 countries/states are included as binary node attributes, set to 1 iff one of the paper's authors is affiliated to an institute in that country/state.

5.2 Results on single-subgroup patterns First, we analyzed single-subgroup patterns on *Lastfm* using beam search with beam width 20 and search depth 2.

5.2.1 Evaluation of the identified subgroups When using the SI measure to perform the pattern discovery, the prior belief is on the individual vertex degrees. As a result, single-subgroup patterns' density will not be explainable merely from the individual degrees of the constituent vertices. For *Lastfm*, given its sparsity, incorporating this prior leads to a background distribution with a small average connection probability. In this case, our algorithm tends to identify dense clusters (i.e. $I = 0$), as these are more informative.

[4] https://aminer.org/citation

[5] IJCAI, AAAI, ICML, NIPS, ICLR, ICDE, VLDB, SIGMOD, ICDT, PODS, SIGIR, WWW, CIKM, ECIR, KDD, ECML-PKDD, WSDM, PAKDD, ICDM, SDM

There exist numerous measures objectively quantifying the interestingness of a dense subgraph community. We make a comparison between our SI measure and some of these objective ones, including the edge density, the average degree, Pool's community score [17], the edge surplus [19], the segregation index [7], the modularity of a single community [15, 16], the inverse average-ODF (out-degree fraction) [21] and the inverse conductance. For space limitations, tables with the most interesting patterns w.r.t these measures are put in the supplement [6]. The main findings are summarized here.

Each of those objective measures exhibits a particular bias that arguably makes the attained patterns less useful in practice. The edge density is easily maximized to a value of 1 simply by considering very small subgraph, and thus top patterns w.r.t this measure are all those composed of only 2 vertices with 1 connecting edges. In contrast, using the average degree tends to find very large communities, because in a large community there are many other vertices for each vertex to be possibly connected to. Although Pool argued that their measure may be larger for larger communities than for smaller ones, in their own experiments on *Lastfm* as well as in our own results, it yields relatively small communities. As they explained, the reason was *Lastfm*'s attribute data is extremely sparse with a density of merely 0.15%. Note the most interesting patterns w.r.t the edge surplus are the same as those w.r.t the Pool's measure. Although these two measures are defined in different ways, Pool's measure can be further simplified to a form essentially the same as the edge surplus (shown in the supplement [6]). Pursuing a larger segregation index essentially targets communities which have less cross-community links than expected. This measure emphasizes more strongly the number of cross-community links, and yields extremely small or large communities with few inter-edges on *Lastfm*. Using the modularity of a single community tends to find rather large communities representing audiences of mainstream music. The results for the inverse average-ODF and the inverse conductance are not displayed, because the largest values for these two measures can be easily achieved by a community with no edges leaving this community, for which a trivial example is the whole network.

We argue that the attained patterns by applying our SI measure are most insightful, striking the right balance between coverage (sufficiently large) and specificity (not conveying too generic or trivial information). The top one characterises a group of 78 idm (i.e., intelligent dance music) fans. Audiences in this group are connected more frequently than expected, and they altogether only have 496 connections to those people not into idm, a small number compared to the number of people outside the group (i.e., $1892 - 78 = 1814$).

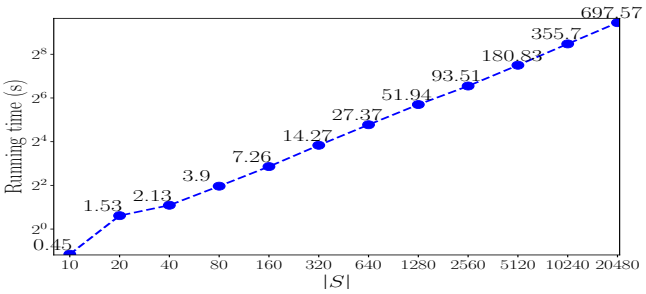


Figure 2: Run time on *Lastfm* for various $|S|$

This sort of qualitative comparison was also made on *DblpAffs* (results in the supplement [6]), for which the same conclusion as above can be reached.

5.2.2 Scalability Fig. 2 illustrates how the algorithm scales w.r.t the number of selectors in the search space (i.e., $|S|$). Both axes are assigned with logarithmic scales with base 2. It is clear that the run time experiences a linear growth as we double the $|S|$ except a tiny disagreement from the second implementation.

5.3 Results on bi-subgroup patterns To identify bi-subgroup patterns, we applied the nested beam search with $x_1 = 8, x_2 = 6$, and $D = 2$. Moreover, we constrain the target descriptions W_1 and W_2 to include at least one common attribute but with various values, so that the corresponding pair of subgroups $\varepsilon(W_1)$ and $\varepsilon(W_2)$ do not overlap with each other. Under this setting, the attained patterns are more explainable, and the results are easier to evaluate.

5.3.1 Evaluation of the SI measure The evaluation of the SI measure addresses two questions:

- Is the SI truly subjective, in the sense of being able to consider data analyst's prior beliefs? (Task 1)
- How can optimizing SI help avoid redundancy in the resulting patterns from an iterative mining? (Task 2)

Task 1: The effects of different prior beliefs, and a subjective evaluation. We consider different prior beliefs, in search for bi-subgroup patterns w.r.t the SI on *Caltech36* and *Reed98*. The top 4 patterns under each prior are presented in Table 2 (for *Caltech36*) and Table 3 (for *Reed98*). For each pattern, the expected number of edges between $\varepsilon(W_1)$ and $\varepsilon(W_2)$ w.r.t the background distribution (i.e., $p_W \cdot n_W$) is also shown.

Prior beliefs on the individual vertex degrees. We first incorporated prior belief on the individual vertex degree (i.e. Prior 1). In general, the identified patterns belong to knowledge commonly held by people, and are not useful. The top 4 patterns on *Caltech36* all

Table 2: Varying prior beliefs in *Caltech36* network

	Rank	W_1	W_2	$\lvert\varepsilon(W_1)\rvert$	$\lvert\varepsilon(W_2)\rvert$	I	k_W	$p_W \cdot n_W$
Prior 1	1	year = 2006	year = 2008	153	173	1	1346	2379.10
	2	status = student ∧ year = 2008	status = alumni	167	159	1	842	1783.26
	3	status = student ∧ year = 2008	year = 2006	167	153	1	1330	2367.96
	4	status = student ∧ year = 2006	year = 2008	152	173	1	1346	2377.53
Prior 1 + Prior 2	1	dorm/house = 169	dorm/house = 171	99	67	1	194	569.56
	2	dorm/house = 169	dorm/house = 166	99	70	1	237	620.42
	3	dorm/house = 169	dorm/house = 172	99	91	1	319	706.65
	4	dorm/house = 169	dorm/house = 170	99	87	1	300	646.04
Prior 1 + Prior 2 + Prior 3	1	status = student ∧ year = 2004	year = 2008	3	173	0	108	25.23
	2	status = student ∧ year = 2004	year = 2008 ∧ minor = 0	3	114	0	71	15.67
	3	status = student ∧ year = 2004	year = 2008 ∧ gender = male	3	116	0	71	16.97
	4	status = student ∧ dorm/house = 166	status = alumni ∧ high school = 19445	53	1	0	51	17.52

reveal people graduating in different years rarely know each other (rows for Prior 1 in Table 2), in particular between ones in class of 2006 and ones in class of 2008 (indicated by the most interesting pattern). Although W_2 of the second pattern (i.e., *status = alumni*) does not contain the attribute graduation year, it implicitly represents people who had graduated in former year. For *Reed98*, the discovered patterns under Prior 1 also express the negative influence of different graduation years on connections (rows for Prior 1 in Table 3).

Prior beliefs on particular attribute knowledge. We then incorporated the prior on the densities between bins for different graduation years (i.e., Prior 2). All the top 4 patterns on *Caltech 36* indicate rare connections between people living in different dormitories, and this is also not surprising. By additionally incorporating prior beliefs on the dependency of the connectivity probability on different dormitories (i.e., Prior 3), patterns characterizing some interesting dense connections are attained. For instance, the top one reveals three people in class of 2004 connect with many in class of 2008. In fact, these three people's graduation had been postponed, as their status is 'student' rather than 'alumni' in year 2005. These two groups are possible to become friends, as the starting year for those 2008 cohort is exactly 2004. The forth pattern indicates a certain alumni knew almost all the students living in dormitory 166. The reason for that might be worth investigating, which could be, e.g., this alumni worked in this dormitory. For *Reed98*, incorporating Prior 1 and Prior 2 provides interesting patterns. The top one indicates people living in dormitory 88 are friends with many in dormitory 89. For an analyst who has preconceived notion such that people living in different dormitories are less likely to know each other (which we believe is common), this pattern is surprising. Both the fourth and the seventh patterns reveal a certain person knew many people in class of 2009.

Summary. As the results show, incorporating different prior belifs leads to different patterns that strongly contrast with these beliefs. The SI can quantify the interestingness subjectively.

Task 2: Evaluation on the iterative pattern mining. Our method is naturally suited for iterative pattern mining, in a way to incorporate the newly obtained pattern into the background distribution for subsequent iterations. For this task, we used *DblpAffs* and *Lastfm* dataset. Results for *Lastfm* are displayed and discussed in the supplement [6]. Here we only analyze the results on *DblpAffs*. Table 4 displays top 3 patterns found in each of the four iterations on *DblpAffs*.

Iteration 1. Initially, we incorporated prior on the overall graph density. The resulting top pattern indicates papers from institutes in USA seldom cite those from other countries.

Iteration 2. After incorporating the top pattern in iteration 1, a set of dense patterns were identified. All the top 3 patterns reveal a highly-cited subgroup of papers whose authors are affiliated to institutes in California and New Jersey. This is possible as many of the world's largest high-tech corporations and reputable universities are located in this region. Examples include Silicon valley, Stanford university in CA, NEC Laboratories, AT&T Laboratories in NJ, among others.

Iteration 3. The top 3 patterns in iteration 3 reveal that papers from authors with Chinese affiliations are rarely cited by papers with authors from other countries. However, they are frequently cited by papers with Chinese authors, as indicated by our identified top single-subgroup pattern in *DblpAffs* (see supplement [6]). This indicates researchers with Chinese affiliations are surprisingly isolated, the reason of which might be interesting to investigate.

Iteration 4. The top patterns in iteration 4 reveal that papers from institutions in Washington state are highly cited by others, in particular by papers from

Table 3: Varying prior beliefs in *Reed98* network

	Rank	W_1	W_2	$\lvert\varepsilon(W_1)\rvert$	$\lvert\varepsilon(W_2)\rvert$	I	k_W	$p_W \cdot n_W$
Prior 1	1	year = 2008	year = 2005	209	117	1	495	1401.97
	2	year = 2007	year = 2009	165	158	1	112	661.41
	3	status = student ∧ year = 2008	year = 2005	209	117	1	495	1401.97
	4	year = 2008	year = 2006	209	131	1	765	1643.38
Prior 1 +Prior 2	1	dorm/house = 89	dorm/house = 88	23	37	0	188	68.80
	2	dorm/house = 89 ∧ status = student	dorm/house = 88	22	37	0	188	68.45
	3	dorm/house = 88 ∧ status = student	dorm/house = 89	36	23	0	183	65.47
	4	dorm/house = 111 ∧ year = 0	year = 2009	1	158	0	24	0.66
	7	dorm/house = 96 ∧ year = 2005	year = 2009	1	158	0	12	0.07

California. Closer inspection revealed that the majority of these papers are written by authors from Microsoft Corporation and the University of Washington.

Summary. By incorporating the newly attained patterns into the background distribution for subsequent iterations, our method can identify patterns which strongly contrast to this knowledge. This results in a set of patterns that are not redundant and highly surprising to the data analyst. Note this does not means we restrict patterns in different iterations not to be associated with each other. In fact, overlapping could happen when this is informative.

5.3.2 Evaluation on the run time The run time of the nested beam search on each dataset, as well as the $|S|$ and $|V|$ statistics are listed in Table 5. The influence of the $|S|$ and $|V|$ on the run time is evident.

6 Related work

Real-life graphs often have attributes on the vertices. Pattern mining considering both structures and attribute information promises more meaningful results, and thus has received increasing research attention. The problem of mining cohesive patterns was introduced by Moser et al. [13]. They define a cohesive pattern as a connected subgraph whose edge density exceeds a given threshold, and vertices exhibit sufficient homogeneity in the attribute space. Gunnemann et al. [9] propose to combine subspace clustering and dense subgraph mining. The former technique is to determine set of nodes that are highly similar according to their attribute values, and the latter is to pursue the cohesiveness of the attained subgraph. Mougel et al. [14] compute all maximal homogeneous clique sets that satisfy some user-defined constraints. All these work emphasizes on the graph structure and consider attributes complementary.

Rather than assuming attributes to be complementary, descriptive community mining, introduced by Pool et al. [17] aims to identify cohesive communities that have a concise description in the vertices' attribute space. They propose a cohesiveness measure based on counting erroneous links (i.e., connections that are either missing or obsolete w.r.t the 'ideal' community given the induced subgraph). To a limited extent, their method can be driven by user's domain-specific background knowledge, which is a preliminary description or a set of nodes that are expected to be part of a community. The search is then triggered by those seed candidates. Our proposed SI is more versatile in a sense that can incorporate more general background knowledge. Galbrun et al. [8] proposes a similar target to Pool et al.'s, but relies on a different density measure, which is essentially the average degree. Atzmueller et al. [2] introduce description-oriented community detection. They apply a subgroup discovery approach to mine patterns in the description space so it comes naturally that the identified communities have a succinct description.

All previous works quantify the interestingness in an objective manner, in the sense that they can not consider a data analyst's prior beliefs and thus operate regardless of context. Also, all previous works focus on a set of communities or dense subgraphs, overlooking other meaningful structures such as a sparse or dense subgraph between two different subgroups of nodes.

7 Conclusion

We presented a method to identify patterns in the form of (pairs of) subgroups of nodes in a graph, such that the density of (the graph between) those node subgroups is interesting. Here, 'interesting' is quantified in a subjective manner, with respect to a flexible type of prior knowledge about the graph the analyst may have, including insights gained from previous patterns.

Our approach improves upon the interestingness measures used in prior work on subgroup discovery for dense subgraph mining in attributed graphs, and generalizes it in two ways: in identifying not only dense but also sparse subgraphs, and in describing the density

Table 4: Top 3 discovered bi-subgroup patterns of each iteration in *DblpAffs* network

	Rank	W_1	W_2	$\lvert\varepsilon(W_1)\rvert$	$\lvert\varepsilon(W_2)\rvert$	I	k_W	$p_W \cdot n_W$
Iteration 1	1	USA = 1	USA = 0	3132	3340	1	335	765.827
	2	USA= 1 $\wedge$ China = 0	USA = 0	2969	3340	1	288	725.970
	3	USA= 1 $\wedge$ Australia = 0	USA = 0	3092	3340	1	320	756.046
Iteration 2	1	NJ (New Jersey) = 0	NJ = 1 $\wedge$ CA (California) = 1	6262	15	0	93	6.909
	2	CA = 0	NJ= 1 $\wedge$ CA = 1	5584	15	0	86	6.132
	3	NJ= 1 $\wedge$ Israel = 0	NJ= 1 $\wedge$ CA = 1	6153	15	0	93	6.757
Iteration 3	1	China = 0	China = 1	5599	873	1	144	271.022
	2	China = 0	China = 1 $\wedge$ IL (Illinois) = 0	5599	861	1	128	266.103
	3	China = 0 $\wedge$ USA = 0	China = 1	2630	873	1	64	168.086
Iteration 4	1	CA = 1	CA = 0 $\wedge$ WA = 1	888	184	0	55	11.726
	2	WA = 0	WA = 1	6254	218	0	182	97.776
	3	CA = 1 $\wedge$ TX (Texas) = 0	CA = 0 $\wedge$ WA = 1	876	184	0	55	11.568

Table 5: Run time of bi-subgroup pattern mining

Dataset	$\lvert S\rvert$	$\lvert V\rvert$	Run time (s)
Caltech36	602	762	6855.52
Reed98	748	962	10692.83
Lastfm	200	1892	5954.50
DblpAffs	232	6472	10015.70

between subgroups that may differ from each other.

The empirical results show that the method succeeds in taking into account prior knowledge in a meaningful way, and is able to identify patterns that provide genuine insight into the high-level network's structure.

References

[1] F. Adriaens, J. Lijffijt, and T. De Bie. Subjectively interesting connecting trees. In *Proc. of ECML-PKDD*, pages 53–69, 2017.

[2] M. Atzmueller, S. Doerfel, and F. Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Inf. Sc.*, 329:965 – 984, 2016.

[3] I. Cantador, P. Brusilovsky, and T. Kuflik. Hetrec workshop. In *Proc. of RecSys*, 2011.

[4] T. De Bie. An information-theoretic framework for data mining. In *Proc. of KDD*, pages 564–572, 2011.

[5] T. De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *DMKD*, 23(3):407–446, 2011.

[6] Junning Deng, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. Explainable Subgraphs with Surprising Densities: A Subgroup Discovery Approach. *arXiv e-prints*, page arXiv:2002.00793, Jan 2020.

[7] L.C. Freeman. Segregation in social networks. *Soc. Meth. & Res.*, 6(4):411–429, 1978.

[8] E. Galbrun, A. Gionis, and N. Tatti. Overlapping community detection in labeled graphs. *DMKD*, 28(5):1586–1610, 2014.

[9] S. Gunnemann, I. Farber, B. Boden, and T. Seidl. Subspace clustering meets dense subgraph mining: A synthesis of two paradigms. In *Proc. of ICDM*, pages 845–850, 2010.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *JASA*, 58(301):13–30, 1963.

[11] F. Lemmerich. Pysubgroup, 2018.

[12] M. Meeng and A. Knobbe. Flexible enrichment with cortana (software demo), 2011.

[13] F. Moser, R. Colak, A. Rafiey, and M. Ester. Mining cohesive patterns from graphs with feature vectors. In *Proc. of SDM*, pages 593–604, 2009.

[14] P.-N. Mougel, M. Plantevit, C. Rigotti, O. Gandrillon, and J.-F. Boulicaut. Constraint-based mining of sets of cliques sharing vertex properties. In *ACNE Workshop @ ECML-PKDD*, pages 48–62, 2010.

[15] M.E.J. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[16] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech.*, (3):P03024, 2009.

[17] S. Pool, F. Bonchi, and M. van Leeuwen. Description-driven community detection. *ACM TIST*, 5(2), 2014.

[18] A.L. Traud, P.J. Mucha, and M.A. Porter. Social structure of Facebook networks. *Phys. A: Stat. Mech. Appl.*, 391(16):4165–4180, 2012.

[19] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli. Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees. In *Proc. of KDD*, pages 104–112, 2013.

[20] M. van Leeuwen, T. De Bie, E. Spyropoulou, and C. Mesnage. Subjective interestingness of subgraph patterns. *MLJ*, 105(1):41–75, 2016.

[21] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *KAIS*, 42(1):181–213, 2015.