Explainable Subgraphs with Surprising Densities: A Subgroup Discovery Approach

Junning Deng, Bo Kang, Jefrey Lijffijt, Tijl De Bie
firstname.lastname@ugent.be
IDLab, Ghent University
Ghent, Belgium

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 any pair of people depends on a range of attributes, such as their age, residence location, workplace, and hobbies. The high-level structure of a graph can thus possibly be described well by means of patterns of the form 'the subgroup of all individuals with a certain properties X are often (or rarely) friends with individuals in another subgroup defined by properties Y', in comparison to what is expected. Such rules present potentially actionable and generalizable insight into the graph.

We present a method that finds node subgroup pairs between which the edge density is interestingly high or low, using an information-theoretic definition of interestingness. Additionally, the interestingness is quantified subjectively, to contrast with prior information an analyst may have about the connectivity. This view immediatly enables iterative mining of such patterns. This is the first method aimed at graph connectivity relations between different subgroups. Our method generalizes prior work on dense subgraphs induced by a subgroup description. Although this setting has been studied already, we demonstrate for this special case considerable practical advantages of our subjective interestingness measure with respect to a wide range of (objective) interestingness measures.

KEYWORDS

Subgroup discovery, Graph mining, Subjective interestingness, Community detection

ACM Reference Format:
Junning Deng, Bo Kang, Jefrey Lijffijt, Tijl De Bie. 2019. Explainable Subgraphs with Surprising Densities: A Subgroup Discovery Approach. In *MLG '19: 15th International Workshop On Mining and Learning with Graphs, August 05, 2019, Anchorage, Alaska USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Real-life graphs (also known as networks) often contain attributes for the nodes. In social networks for example, where nodes correspond to individuals, node attributes can include the individuals' interests, education, residency, and more. The connectivity of the network is usually highly related to those attributes. For example, the attributes of individuals affect the likelihood of them meeting in the first place, and, if they meet, of becoming friends. Hence, it must be possible to understand the connectivity of a graph in terms of those attributes, at least to a certain extent.

An obvious 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 can fail to provide insight in other classification problems. The local pattern mining community has therefore introduced the concept of *subgroup discovery*, where the aim is to identify subgroups of data points for which a target attribute has homogeneous and/or outstanding values. Such subgroup rules are local patterns, in that they provide information about a certain part of the input space only.

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 in the network: the tendency of links to exist between nodes sharing similar attributes. While the assumption of homophily 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 the fact that the interestingness of such patterns has invariably been quantified using objective measures—i.e. measures that do not depend on 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). Not only may uninteresting patterns appear interesting if prior knowledge is ignored, also interesting patterns may appear uninteresting and are hence not found. E.g., a pattern in a student friendship network that indicates tennis lovers are rarely connected may be due to the lack of suitable facilities or a tennis club.

A *third limitation* of prior work is that the patterns describe only the connectivity within communities and not between groups. As

MLG '19, August 05, 2019, Anchorage, Alaska USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

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, but as we will show in the experiments real-life networks contain many less obvious examples.

Contributions. We depart from the existing literature in formalizing a subjective interestingness measure, rather than an objective one, 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. More specifically, we build on the ideas from the exploratory data mining framework FORSIED [5, 7]. This framework stipulates in abstract terms how to formalize the subjective interestingness of patterns. Basically, a *background distribution* is constructed to model prior beliefs the analyst holds about the data. Given that, one can identify patterns which strongly contrast to this background knowledge and are highly surprising to the analyst. Moreover, this interestingness measure is naturally applicable for patterns describing 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:

- Novel definitions of single-subgroup patterns and bi-subgroup patterns. [Sect. 2]
- A quantification of their Subjective Interestingness (SI), based on what prior beliefs an analyst holds, or what information an analyst gains when observing a pattern. [Sect. 3]
- We propose an algorithm to mine bi-subgroup patterns based on the classical beam search. [Sect. 4]
- We empirically evaluate our method on three real-world data, to investigate its ability to encode the analyst's prior beliefs and identify subjective interesting patterns. [Sect. 5]

2 SUBGROUP PATTERN SYNTAXES FOR GRAPHS

In this section we introduce both single subgroup and bi-subgroup patterns for graphs. Here, we first introduce 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 $m = |E|$ edges, and A is a set of attributes $a \in A$ defined as functions $a : V \rightarrow \text{Dom}_a$, 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 $s_{a,y} : V \rightarrow \{\text{true}, \text{false}\}$, 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 $s_{a,[l,u]} : V \rightarrow \{\text{true}, \text{false}\}$, 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 *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 the following two properties hold,*

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

(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.

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 $\frac{1}{2}|\varepsilon(W)|(|\varepsilon(W)| - 1)$ for undirected graphs without self-edges.

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 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 $n_W \triangleq \varepsilon(W_1) * \varepsilon(W_2) - \varepsilon(W_1 \wedge W_2)$ for undirected graphs without self-edges. Note that single-subgroup patterns are a special case of bi-subgroup patterns when $W_1 \equiv W_2$.

Remark 1. *While single-subgroup patterns and bi-subgroup patterns have been defined for undirected graphs without self-edges, all our results are easily extended to directed graphs and graphs with self-edges by adapting the definitions of k_W and n_W accordingly.*

3 FORMALIZING THE SUBJECTIVE INTERESTINGNESS

Previous work on mining patterns in attributed graphs tended to identify dense communities, of which the interestingness was quantified in an objective way (see Sect. 6). However, for a data analyst with prior information about the data (a situation we argue is common), the resulting patterns may be trivial, containing limited information that is novel to them. Tackling this necessitates the use of subjective measures of interestingness.

3.1 The FORSIED framework

Here, we follow the so-called FORSIED[1] framework [6] to quantify the subjective interestingness of a pattern, which enables to account for prior beliefs the data analyst holds about the data. In this framework, the analyst's belief state is modeled by a so-called *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 evaluate the

[1] An acronym for 'Formalizing subjective interestingness in exploratory data mining'.

probability for any given pattern to be present in the data, and thus to assess the surprise in the analyst when informed about its presence. 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 Sect. 3.2.

Given a background distribution, the Subjective Interestingness (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, which can be computed as the negative log probability of the pattern w.r.t. the background distribution P. The DL is quantified as the code length needed to communicate the pattern to the analyst. For details see Sect. 3.3.

3.2 The background distribution

3.2.1 The initial background distribution. To derive the initial background distribution, we need to assume what prior beliefs the data analyst may have. Here we discuss three types of prior beliefs: (1) on individual vertex degrees; (2) on the overall graph density; (3) on densities between bins.

(1–2) *Prior beliefs on individual vertex degrees and on the overall graph density.* Given the analyst's prior beliefs about the degree of each vertex, [6] showed that the maximum entropy distribution is a product of independent Bernoulli distribution, one for each of the random variable $b_{u,v}$:

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

where $b_{u,v}$ equals to 1 if $(u,v) \in E$ and 0 otherwise. The parameters λ_u^r and λ_v^c can be computed efficiently. For a prior belief on the overall density, every edge probability simply equals the assumed density.

(3) *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 in [1], the resulting background distribution is also a product of Bernoulli distributions, one for each of the random variable $b_{u,v} \in \{0,1\}$):

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

where λ_u^r ,λ_v^c and γ_k are parameters and can be computed efficiently. Note our model are not limited to incorporate this type of belief related to a single attribute. Nodes can be partitioned differently by another attribute. Our model can consider multiple attributes so that analysts could express prior beliefs regarding the edge densities between bins resulting from multiple partitions.

3.2.2 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 $\varepsilon(W_2)$. Again in [5], 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 [20] showed that the resulting P' is again a product of Bernoulli distribution:

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

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

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

3.3 The subjective interestingness measure

3.3.1 The information content. 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 more than k_W (for $I = 0$) or $n_W - k_W$ (for $I = 1$) successes in n_W trials with possibly various success probability $p_{u,v}$ (for $I = 0$) or $1 - p_{u,v}$ (for $I = 1$). More specifically, we consider a success for the case $I = 0$ to be the presence of an edge between some pair of vertices (u,v) for $u \in \varepsilon(W_1)$, $v \in \varepsilon(W_2)$, and $p_{u,v}$ is the corresponding success probability. In contrast, the absence of an edge between some vertices (u,v) is deemed to be a success for the case $I = 1$, with the probability as $1 - p_{u,v}$. [20] proposed to tightly upper bound the probability of a sort of dense subgraph pattern by applying the general Chernoff/Hoeffding bound [4, 11]. Here, we can use the same methodology, which gives:

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

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

Here, $p_W = \frac{1}{n_W} \sum_{u \in \varepsilon(W_1), v \in \varepsilon(W_2)} p_{u,v}$, $\mathbf{KL}\Big(\frac{k_W}{n_W} \parallel p_W\Big)$ is the Kullback-Leibler divergence between two Bernoulli distribution with success probabilities $\frac{k_W}{n_W}$ and p_W respectively. Note that:

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

We can thus write:

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

The information content is the negative log probability of the pattern being present under the background distribution. That is,

$$\begin{aligned} \mathrm{IC}[(W_1, W_2, I, k_W)] &= -\log(\mathbf{Pr}[(W_1, W_2, I, k_W)]), \\ &\geq n_W \mathbf{KL}\left(\frac{k_W}{n_W} \parallel p_W\right). \end{aligned} \tag{1}$$

3.3.2 The description length. The DL should capture the length of the description needed to communicate the pattern (W_1, W_2, I, k_W). 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 costs constantly as 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

$$\mathrm{DL}[(W_1, W_2, I, k_W)] = \alpha(|W_1| + |W_2|) + \beta. \tag{2}$$

3.3.3 The subjective interestingness. In summary, we get:

$$\begin{aligned} \mathrm{SI}[(W_1, W_2, I, k_W)] &= \frac{\mathrm{IC}[(W_1, W_2, I, k_W)]}{\mathrm{DL}[(W_1, W_2, I, k_W)]}, \\ &= \frac{n_W \mathbf{KL}\left(\frac{k_W}{n_W} \parallel p_W\right)}{\alpha(|W_1| + |W_2|) + \beta}. \end{aligned} \tag{3}$$

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, we consider applying some heuristic search strategies to help maintain the tractability. For searching single-subgroup patterns, we used beam search (see Sect. 4.1). To search for the bi-subgroup patterns, however, a traditional beam over both W_1 and W_2 simultaneously turned out to be more difficult to apply effectively. We thus propose a nested beam search strategy to handle this case. More details about this strategy are covered by Sect. 4.2, followed by a brief introduction to the implementation in Sect. 4.3.

4.1 The beam search

In the case of 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.

4.2 The nested beam search

The basic idea is to nest one beam search into the other one where the outer search branches based on a 'beam' of promising selectors for the description W_1 , and the inner 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 interesting patterns identified by our algorithm is $x_1 \cdot x_2$. To maintain a sufficient diversity among the discovered patterns, we deliberately constrain the outer beam to contain at least x_1 different W_1 descriptions. Due to the space limitation, the pseudo code for this search and related notations are put in online supplement[2] (see Algorithm 1 and Table S1).

Table 1: Dataset statistics summary

Dataset	$\|V\|$	$\|E\|$	#Attributes	$\|S\|$
Caltech36	762	16651	7	602
Reed98	962	18812	7	748
Lastfm	1892	12717	11946	23892

4.3 Implementation

Pysubgroup [12], a Python package for subgroup discovery implementation written by Florian Lemmerich, was used as a base to be built upon. 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 at https://www.dropbox.com/sh/z782s8ohuo3jfee/AAC9bxrfN_wqCLGU4DR49RDDa?dl=0. All experiments were conducted on a PC with Ubuntu OS, Intel(R) Core(TM) i7-7700K 4.20GHz CPUs, and 32 GB of RAM.

5 EXPERIMENTS

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

5.1 Data

Of the three datasets used in the experiments, two are obtained from the Facebook100 dataset [18], and the other is from the online music platform Lastfm[3]. Data size statistics are given Table 1.

Facebook100 contains a set of 100 Facebook networks of different American college and universities from a single-day snapshot in September 2005. Each network consists of the complete set of users (nodes) from Facebook at one particular university and all the friendship links (edges) between those users. Each node is annotated with additional information including the user's student/faculty status flag, gender, major, minor, dorm/house, graduation year, and high school. We select the networks at Caltech (i.e., Caltech36) and Reed university (i.e. Reed98) to experiment on.

Lastfm is a publicly available dataset from the HetRec 2011 workshop [3]. The social network is generated from friend relations between users in Lastfm. A list of most-listened musical artists and tag assignments for each user is given in a tuple form as [user, tag, artist]. We took all the tags that a user ever assigned to any artist and assigned those to the user. Then we transformed those tags for each user into a binary vector to serve as the attributes. Those attributes express a user's music interests to some extent.

[2] https://www.dropbox.com/s/pc0w4uwniwy5u3m/supplementaryMLG.pdf?dl=0
[3] http://www.lastfm.com

5.2 Results on single-subgroup patterns

First, we analyzed single-subgroup patterns on the Lastfm dataset 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. 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 [8], the modularity of a single community [15, 16], the inverse average-ODF (out-degree fraction) [21] and the inverse conductance. For space limitations, the tables presenting the most interesting patterns w.r.t these measures are put in the online supplement[4], Table S2 for SI, Table S4 - S8 for other measures. Also, Table S3 gives a description for each of those objective measures. 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. That's why the patterns identified by using this measure are all those composed of only 2 vertices with 1 connecting edge. 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 the Lastfm network 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 that patterns with the top 10 edge surplus values are the same as those for 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 Table S3). Pursuing a larger segregation index essentially targets communities which have much 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 in the supplement, 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., $1892 - 78 = 1814$).

[4]https://www.dropbox.com/s/pc0w4uwniwy5u3m/supplementaryMLG.pdf?dl=0

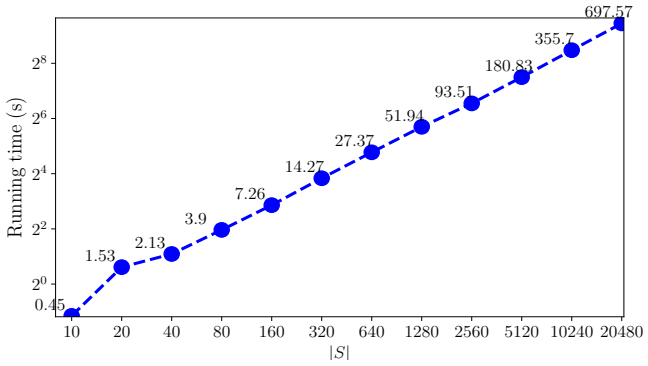

Figure 1: Running time on Lastfm data for various $|S|$

5.2.2 The scalability evaluation. Fig. 1 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 running 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 search depth 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 on the SI measure. We steered tasks of evaluating the SI measure to answer the following questions:

- **Q1** Is our SI measure truly subjective, in the sense of being able to consider data analyst's prior beliefs? (Task 1)
- **Q2** Does our SI measure embody the effects of different description lengths? (Task 2)
- **Q3** How can optimizing SI help avoid redundancy in the resulting patterns from an iterative mining? (Task 3)

Task 1: The effects of different prior beliefs, and a subjective evaluation.

We consider different prior beliefs, in search for bi-subgroup patterns according to our SI measure on Caltech36 and Reed98 dataset. The 4 most interesting patterns under each prior belief are presented in the following (Table 2 for Caltech36, Table 3 for Reed98). For each pattern, we also display its value of $p_W * n_W$, the expected number of edges between $\varepsilon(W_1)$ and $\varepsilon(W_2)$ w.r.t the background distribution. Comparing $p_W * n_W$ to k_W gives a direct sense how much the analyst's expectation differs from the truth.

Prior beliefs on the individual vertex degrees. The first prior belief is on the individual vertex degree (i.e. Prior 1). In general,

Table 2: Varying prior beliefs in Caltech36 network

	Rank	W_1	W_2	$\|\varepsilon(W_1)\|$	$\|\varepsilon(W_2)\|$	I	k_W	$p_W * n_W$
Prior 1	1	year = 2006	year = 2008	153	173	1	1346	2379.095
	2	student status = student ∧year = 2008	student status= alumni	167	159	1	842	1783.259
	3	student status = student ∧year = 2008	year= 2006	167	153	1	1330	2367.963
	4	student status = student ∧year = 2006	year= 2008	152	173	1	1346	2377.526
Prior 1+Prior 2	1	dorm/house = 169	dorm/house = 171	99	67	1	194	569.558
	2	dorm/house = 169	dorm/house = 166	99	70	1	237	620.424
	3	dorm/house = 169	dorm/house = 172	99	91	1	319	706.645
	4	dorm/house = 169	dorm/house = 170	99	87	1	300	646.044
Prior 1	1	student status = student ∧year =2004	year = 2008	3	173	0	108	25.232
	2	student status = student ∧year =2004	year= 2008∧minor = 0	3	114	0	71	15.671
+ Prior 2	3	student status = student ∧year =2004	year = 2008∧gender = male	3	116	0	71	16.967
+ Prior 3	4	student status = student ∧ dorm/house = 166	student status = alumni ∧ high school = 19445	53	1	0	51	17.523

patterns found based on Prior 1 belong to knowledge commonly held by people, and are not useful. The top 4 patterns on the Caltech36 data all suggest inactive connections particularly between people graduating from different years (rows for Prior 1 in Table 2). The most interesting pattern states that ones graduating in year 2008 rarely know those graduating in year 2006. Either the third or the fourth pattern can be regarded as a more strict version of the top one, compared to which an additional feature is satisfied. Although the second description of the second pattern (i.e., *student status = alumni*) does not contain the attribute graduation year, it implicitly represents people who had already graduated previously. For the Reed98 network, the discovered patterns under the 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 further incorporate this commonly believed knowledge by encoding it as as prior beliefs on densities between bins for different graduation years (i.e., Prior 2). The yielded top 4 patterns on Caltech 36 all indicate sparse connections between people living in different dormitories. Again, knowing this is not surprising. By incorporating prior beliefs on the dependency of the connecting probability on the difference in dormitories (i.e., Prior 3), plus Prior 1 and Prior 2, truly interesting patterns describing some dense connections are discovered. For instance, the top pattern states three people graduating in 2004 are friends with many graduating in 2008. Notice these three people's status is student rather than alumni in year 2005. A probable reason for their graduation delay can be, for example, a failure in the exam. Also, the starting year for those 2008 cohort is exactly when these three people should had graduated. Therefore, these two groups had opportunities to become friends with each other. The forth pattern indicates an alumni who had studied in a high school connected with almost all the students living in a certain dormitory. Knowing this can simulate analyst's curiosity to discern the reason, which for instance, could be that the alumni may work in this dormitory. As for Reed98 network, incorporating Prior 1 and Prior 2 is sufficient to gain some useful information. The top pattern in this case indicates people living in dormitory 88 are often friends with those in dormitory 89. What people commonly believe, by contrast, is that people living in different dormitory are less likely to know each other. For an analyst who has such preconceived notion, this top pattern indeed conveys useful information. The fourth and the seventh patterns are also very interesting. Either of them describes a certain person who knows much more people graduating in year 2009 than being expected.

Summary. As we can see, incorporating different prior beliefs leads to the attain of different patterns that contrast with these beliefs. Our measure can indeed quantify the interestingness subjectively.

Task 2: The effect of the description length.

By comparing the fourth pattern to the top one in rows for Prior 1 in Table 2, we can notice the effect of the DL. The information contents of these two patterns must be very similar, because they have exactly the same bound about the connection counting (i.e., k_W), and nearly the same expected value for that (i.e., $p_W * n_W$). We can deduce what drags the fourth pattern to a lower rank is its longer DL, as one more selector is contained in W_1. This shows our SI measure can take DL into account.

Task 3: 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. Table 4 displays the top 3 patterns found in each of the five iterations on the Lastfm dataset. The description search space is built based on only 100 most frequently used tags, that means, $|S| = 100 * 2$.

Iteration 1. Initially, we incorporate Prior 1. In this case, the most interesting pattern reflects a conflict between aggressive heavy metal fans and mainstream pop lovers who do not listen to heavy metal at all.

Iteration 2. After incorporating the top pattern identified in iteration 1, what comes top is the one expressing again a conflict between mainstream and non-mainstream music preference, but another kind (i.e., pop with no indie, and experimental with no pop). Also, we can notice only the second pattern for the iteration 1 is remained in the iteration 2 top list but with a lower rank as third.

Table 3: Varying prior beliefs in Reed98 network

	Rank	W_1	W_2	$\|\varepsilon(W_1)\|$	$\|\varepsilon(W_2)\|$	I	k_W	$p_W * n_W$
Prior 1	1	year = 2008	year = 2005	209	117	1	495	1401.969
	2	year = 2007	year = 2009	165	158	1	112	661.411
	3	student status = student ∧year =2008	year = 2005	209	117	1	495	1401.969
	4	year = 2008	year = 2006	209	131	1	765	1643.375
Prior 1+Prior2	1	dorm/house= 89	dorm/house= 88	23	37	0	188	68.803
	2	dorm/house= 89∧ student status = student	dorm/house= 88	22	37	0	188	68.454
	3	dorm/house= 88∧ student status = student	dorm/house= 89	36	23	0	183	65.465
	4	dorm/house = 111∧ year = 0	year = 2009	1	158	0	24	0.661
	7	dorm/house = 96∧ year = 2005	year = 2009	1	158	0	12	0.067

The interestingness of any sparse pattern associated with the newly incorporated one under the updated background distribution is expected to decrease, as the data analyst's would not feel surprised about such pattern.

Iteration 3. In iteration 3, our method tends to identify some interesting dense patterns, mainly related to synth pop and new wave genres. The top one states synth pop fans frequently connect with many people listening to new wave but not synth pop. This pattern appears fallacious at the first glance. Nevertheless, synth pop is a subgenre of new wave music. Also, the latter group may listen to synth pop but they use a different tag 'synthpop' instead of 'synth pop', as there are even 102 audience only tag synth pop as 'synthpop' (see the third patten). Hence, this pattern makes sense as it describes dense connections between two groups which resemble each other.

Iteration 4. The top 3 patterns in iteration 4 all express negative associations between new wave and some sort of catchy mainstream music (eg. pop, rnb, or hip-hop, among several others).

Iteration 5. Once we incorporate the most interesting one, patterns characterizing some positively associated genres stand out. For example, the top one in iteration 5 indicates instrumental audience are friends with many ambient audience who doesn't listen to instrumental music. These two genres are not opposite concepts and share many in common (e.g., recordings for both do not include lyrics). Actually, ambient music can be regarded as a slow form of instrumental music.

Summary. By incorporating the newly attained patterns into the background distribution for subsequent iterations, our method can identify another patterns which strongly contrast to this knowledge, resulting a set of patterns that are not redundant and are 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 running time. The running 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 $|V|$ to the running time is evident.

6 RELATED WORK

Real-life graphs often have attributes on the vertices. Pattern mining taking into account both structural aspect and attribute information promises more meaningful results, and has received increasing research attention.

The problem of mining cohesive patterns is 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. [10] 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. In [14], Mougel et al. compute all maximal homogeneous clique sets that satisfy some user-defined constraints. All these work emphasizes on the graph structure and consider attributes as complementary information.

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 cohesiveness measure, which is based on counting erroneous links (i.e., user connections that are either missing or obsolete with respect to the 'ideal' community given the induced subgraph). Their method can be driven by user's domain-specific background knowledge, but very limitedly. Specifically, the background knowledge is only allowed to be either a preliminary description or a set of nodes that are expected to be part of a community. Then the search is triggered by those seed candidates. Compared to that, our SI measure is more versatile in a sense that allows incorporating more general background knowledge. Galbrun et al.'s work [9] shares similar target to Pool et al.'s, but relies on a different density measure, which is essentially the average degree. A comparison from our SI measure to Pool's measure and the average degree are included in our experimental evaluation. 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 work quantify the interestingness in an objective manner, in the sense that they can not consider a data analyst's beliefs or expectations and thus operate regardless of context. The novelty of our algorithm is in modelling and using the analyst's beliefs, and inserting the subjective informativeness into the targeted patterns. Also, previous work focus on a set of single communities or dense subgraphs, overlooking other meaningful structures, e.g.,

Table 4: Top 4 discovered bi-subgroup patterns of each iteration in Lastfm network

	Rank	W_1	W_2	$\|\varepsilon(W_1)\|$	$\|\varepsilon(W_2)\|$	I	k_W	$p_W * n_W$
Iteration 1	1	heavy mental = 1	heavy mental = 0∧ pop = 1	165	529	1	349	769.175
	2	pop = 1∧experimental = 0	rnb = 0∧experimental = 1	497	230	1	360	812.781
	3	pop = 1∧experimental = 0	experimental = 1	497	247	1	495	943.964
Iteration 2	1	pop = 1∧ indie = 0	pop = 0∧ experimental = 1	366	159	1	103	369.443
	2	pop = 1∧alternative = 0	pop = 0∧experimental = 1	325	159	1	84	334.766
	3	pop = 1∧experimental = 0	rnb = 0∧ experimental = 1	497	230	1	360	750.771
Iteration 3	1	synth pop = 1	synth pop = 0∧ new wave = 1	54	150	0	163	43.009
	2	synth pop = 1∧ british = 1	new wave = 1∧british = 0	26	113	0	116	20.710
	3	synth pop = 1	synth pop = 0∧ synthpop = 1	54	102	0	125	29.643
Iteration 4	1	new wave = 1∧ hip-hop = 0	new wave = 0∧ pop = 1	160	475	1	343	670.739
	2	new wave = 1∧ rnb = 0	new wave = 0∧pop = 1	170	475	1	379	705.432
	3	new wave = 1∧soul = 0	new wave = 0∧ pop = 1	150	475	1	323	624.411
Iteration 5	1	instrumental = 1	instrumental = 0∧ ambient = 1	195	144	0	273	114.619
	2	electronic = 1	electronic = 0∧ambient = 1	167	160	0	268	113.664
	3	progressive metal = 1	progressive metal = 0∧ heavy metal = 1	99	111	0	128	34.807

Table 5: Running time of bi-subgroup pattern mining

Dataset	$\|S\|$	$\|V\|$	Running time (s)
Caltech36	602	762	6855.522
Reed98	748	962	10692.833
Lastfm	200	1892	5954.501

a sparse subgraph and a pair of different subgroups, which our method can handle.

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 subgraphs, and generalizes it in two ways: in identifying not only dense but also sparse subgraphs, and in describing the density between subgroups that may be different 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.

ACKNOWLEDGMENTS

This work was supported by the ERC under the EU's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement no. 615517, FWO (project no. G091017N, G0F9816N), and the EU's Horizon 2020 research and innovation programme with the FWO under the MSC Grant Agreement no. 665501.

REFERENCES

[1] F. Adriaens, J. Lijffijt, and T. De Bie. 2017. Subjectively Interesting Connecting Trees. In *Proc. of ECML-PKDD*. 53–69.
[2] M. Atzmueller, S. Doerfel, and F. Mitzlaff. 2016. Description-oriented community detection using exhaustive subgroup discovery. *Inf. Sc.* 329 (2016), 965 – 984.
[3] I. Cantador, P. Brusilovsky, and T. Kuflik. 2011. HetRec Workshop. In *Proc. of RecSys*.
[4] Herman Chernoff. 1952. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *Ann. Math. Stat.* 23, 4 (1952), 493–507.
[5] T. De Bie. 2011. An information-theoretic framework for data mining. In *Proc. of KDD*. 564–572.
[6] T. De Bie. 2011. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *DMKD* 23, 3 (2011), 407–446.
[7] T. De Bie. 2013. Subjective interestingness in exploratory data mining. In *Proc. of IDA*. 19–31.
[8] L.C. Freeman. 1978. Segregation in Social Networks. *Soc. Meth. & Res.* 6, 4 (1978), 411–429.
[9] E. Galbrun, A. Gionis, and N. Tatti. 2014. Overlapping community detection in labeled graphs. *DMKD* 28, 5 (2014), 1586–1610.
[10] S. Gunnemann, I. Farber, B. Boden, and T. Seidl. 2010. Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms. In *Proc. of ICDM*. 845–850.
[11] W. Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. *JASA* 58, 301 (1963), 13–30.
[12] F. Lemmerich. 2018. PySubgroup.
[13] F. Moser, R. Colak, A. Rafiey, and M. Ester. 2009. Mining Cohesive Patterns from Graphs with Feature Vectors. In *Proc. of SDM*. 593–604.
[14] P.-N. Mougel, M. Plantevit, C. Rigotti, O. Gandrillon, and J.-F. Boulicaut. 2010. Constraint-Based Mining of Sets of Cliques Sharing Vertex Properties. In *ACNE Workshop @ ECML-PKDD*. 48–62.
[15] M.E.J. Newman. 2006. Modularity and community structure in networks. *PNAS* 103, 23 (2006), 8577–8582.
[16] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. 2009. Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech.* 3 (2009), P03024.
[17] S. Pool, F. Bonchi, and M. van Leeuwen. 2014. Description-Driven Community Detection. *ACM TIST* 5, 2, Article 28 (2014), 28 pages.
[18] A.L. Traud, P.J. Mucha, and M.A. Porter. 2012. Social structure of Facebook networks. *Phys. A: Stat. Mech. Appl.* 391, 16 (2012), 4165–4180.
[19] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli. 2013. Denser Than the Densest Subgraph: Extracting Optimal Quasi-cliques with Quality Guarantees. In *Proc. of KDD*. 104–112.
[20] M. van Leeuwen, T. De Bie, E. Spyropoulou, and C. Mesnage. 2016. Subjective interestingness of subgraph patterns. *MLJ* 105, 1 (2016), 41–75.
[21] J. Yang and J. Leskovec. 2015. Defining and Evaluating Network Communities Based on Ground-truth. *KAIS* 42, 1 (2015), 181–213.