Data Mining and Knowledge Discovery (2021) 35:321–371
https://doi.org/10.1007/s10618-020-00721-9

Mining explainable local and global subgraph patterns with surprising densities

Junning Deng[1] · Bo Kang[1] · Jefrey Lijffijt[1] · Tijl De Bie[1]

Received: 13 February 2020 / Accepted: 21 October 2020 / Published online: 10 November 2020
© The Author(s) 2020

Abstract

The connectivity structure of graphs is typically related to the attributes of the vertices. 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 certain properties X are often (or rarely) friends with individuals in another subgroup defined by properties Y', ideally relative to their expected connectivity. Such rules present potentially actionable and generalizable insight into the graph. Prior work has already considered the search for dense subgraphs ('communities') with homogeneous attributes. The first contribution in this paper is to generalize this type of pattern to densities between a *pair of subgroups*, as well as between *all pairs from a set of subgroups that partition the vertices*. Second, we develop a novel information-theoretic approach for quantifying the subjective interestingness of such patterns, by contrasting them with prior information an analyst may have about the graph's connectivity. We demonstrate empirically that in the special case of dense subgraphs, this approach yields results that are superior to the state-of-the-art. Finally, we propose algorithms for efficiently finding interesting patterns of these different types.

Responsible editor: M. J. Zaki

✉ Junning Deng
Junning.Deng@ugent.be

Bo Kang
Bo.Kang@ugent.be

Jefrey Lijffijt
Jefrey.Lijffijt@ugent.be

Tijl De Bie
Tijl.DeBie@ugent.be

1 IDLab, Ghent University, Technologiepark-Zwijnaarde 122, Ghent, Belgium

Keywords Graph mining · Graph summarization · Graph clustering · Community detection · Subjective interestingness · Subgroup discovery

1 Introduction

Real-life graphs (also known as networks) often contain attributes for the vertices. In social networks for example, where vertices correspond to individuals, vertex attributes can include the individuals' interests, education, residency, and more. The connectivity of the network is usually highly related to those attributes (Fond and Neville 2010; McPherson et al. 2001; Aral et al. 2009; Li et al. 2017). The attributes of individuals affect the likelihood of them meeting in the first place, and, if they meet, of becoming friends. Hence, it appears likely it should be possible to understand the connectivity of a graph in terms of those attributes, at least to a certain extent.

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 of a vertex pair, predicting the edge as present or absent (Gong et al. 2014; Yin et al. 2010; Barbieri et al. 2014; Wei et al. 2017). Such global models often fail to provide insight though . To address this, the local pattern mining community 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 (Herrera et al. 2011; Atzmueller 2015). Such subgroup rules 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 vertex-induced subgraphs, dubbed *communities*, that are coherent also in terms of attributes. There are two complementary approaches, as stated in Atzmueller et al. (2016). The first explores the space of communities that meet certain criteria in terms of density, in search for those that are also homogeneous with respect to some of the attributes (Moser et al. 2009; Mougel et al. 2010). The second explores the space of rules over the attributes, in search for those that define subgroups (of vertices) that form a dense community (Pool et al. 2014; Galbrun et al. 2014; Atzmueller et al. 2016). This is effectively a subgroup discovery approach to dense subgraph mining.

Limitations of the state-of-the-art Both these approaches hinge on the existence of attribute homophily in the network: the tendency of links to exist between vertices with similar attributes (McPherson et al. 2001). Yet, while the assumption of homophily is often reasonable, it limits the scope of application of prior work . 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 vertices, 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 a single group and not between two potentially distinct groups. 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, 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 (De Bie 2011a, 2013). 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. Finally, apart from a local pattern mining strategy which is used to identify interesting patterns one by one, we also propose a strategy to mine patterns globally, that is, to summarize the whole graph in a meaningful way such that all the interesting patterns can immediately be seen. The resulting summarization can be considered as a type of global pattern. Our specific contributions are:

– Novel definitions of single-subgroup patterns and bi-subgroup patterns, as well as patterns that are global summaries for attributed graphs. (Sect. 3)
– 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. 4)
– An algorithm to mine bi-subgroup patterns based on beam search. (Sect. 5)
– An algorithm to mine global (or summarization) patterns from which a series of interesting single-subgroup and bi-subgroup patterns can be revealed. (Sect. 5)
– An empirical evaluation of our method on real-world data, to investigate its ability to encode the analyst's prior beliefs and identify subjective interesting patterns. (Sect. 6)

This manuscript is a significant extension of Deng et al. (2020). The main additions include the further generalization of the single-subgroup and bi-subgroup patterns (both types are local patterns) to global patterns, the quantification of the SI as well as the search algorithm for global patterns. Moreover, we substantially extend the experiment section by analyzing the parameter sensitivity of our beam search methods to the beam width, further investigating research questions already proposed in Deng et al. (2020) on more real-world datasets, as well as evaluating the performance of our global pattern mining method.

2 Related work

In this section, we first briefly review some graph modelling work (Sect. 2.1), more specifically, those based on formulating a *statistical ensemble of networks* (i.e., the collection of all possible realizations into which the considered network may reasonably evolve with a probability (Fronczak 2012)). The numerical and analytical study of such ensembles provides the foundation of model fitting, model selection, for various applications including the pattern mining (Casiraghi et al. 2016). We then review related work dedicated to pattern mining in attributed graphs. This review is along two dimensions, concerning local patterns (Sect. 2.2.1) and global patterns (Sect. 2.2.2) respectively.

2.1 Graph modelling

Graph modelling typically considers a given network (i.e., the one we observe) as merely a realization among a large number of possibilities. All possible realizations including the observed one that are consistent with some given aggregate statistics, forms the so-called *statistical ensemble of networks*.

A well-founded probabilistic framework to such graph modelling is provided by exponential random graph models (ERGMs) (Holland and Leinhardt 1981; Harris 2013). In ERGMs, each graph has a probability that depends on a number of chosen statistics of the network. Such models allow one to sample random graphs that match certain graph properties as closely as possible, without the need to know the underlying network generation process (Fronczak 2012). Nevertheless, a downside of ERGMs is their intractable fitting on large, finite networks. Recently, Casiraghi et al. introduce a broad class of analytically tractable statistical ensembles of finite, directed and weighted networks, referred to as *generalized hypergeometric ensembles* (Casiraghi et al. 2016).

Unlike ERGMs that aim to be an accurate and objective probabilistic model for the data, the aim of our method is to provide the data analyst with subjectively interesting insights into the data. To do that, intelligible pattern syntaxes need to be designed to represent the data‘s local or global information. Secondly, the found patterns must be contrasted with a model of the data analyst’s belief state about the data (called the background distribution) to quantify their interestingness to the data analyst (this makes our approach a subjective one). A further distinction from ERGMs is that our method is naturally an iterative method, allowing the data analyst to gain new insights from one or a few patterns at a time.

2.2 Pattern mining in attributed graphs

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

2.2.1 Local pattern mining

The problem of mining cohesive patterns was introduced by Moser et al. (2009). 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. Mougel et al. (2010) computes all maximal homogeneous clique sets that satisfy some user-defined constraints. All these works emphasize the graph structure and consider attributes as complementary information. Rather than assuming attributes to be complementary, descriptive community mining, introduced by Pool et al. (2014) 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., 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, and more specifically, it is a preliminary description or a set of vertices that are expected to be part of a community. Then the search is triggered by those seed candidates. Our proposed SI, in contrast, is more versatile in a sense that allows incorporating more general background knowledge. Galbrun et al. (2014) 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. (2016) introduces description-oriented community detection. In this work, a subgroup discovery approach is applied 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 cannot 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 vertices.

2.2.2 Global pattern mining by summarizing or clustering

Discovering global patterns that can uncover useful insights in attributed graphs are typically tailored to a graph summarization or a clustering task. Although these two tasks can both output graph summary, their goals (even when solely considering the structural aspect) are fundamentally different. Graph summarization seeks to group together vertices that connect with the rest of the graph in a similar way, while clustering simply group vertices that are densely connected to each other and are well separated from other groups (Liu et al. 2018).

Graph summarization Tian et al. (2008) proposes *SNAP* and *k-SNAP* for controlled and intuitive graph summarization. These methods can produce customized summaries based on user-selected attributes and relationships that are of interest. Furthermore, the resolutions of the resulting summaries can also be controlled by users. Then Zhang et al. (2010) further builds on this work by addressing two key limitations. First, they allow automatic categorization of numeric attributes (which is a common scenario). Second, they propose a measure to access the interestingness of summaries so that the user does not have to manually inspect a large number of summaries to find the interesting ones. However, their interestingness measure is not subjective, simply considering the tradeoff among diversity, coverage and conciseness. Chen et al. (2009) proposes *SUMMARIZE-MINE*, a framework that performs the detection of frequent subgraphs on randomised summaries for multiple iterations, so that a lossy compression can be effectively turned into a virtually lossless one. In addition to pattern discovery, graph summarization on attributed graphs can serve for several applications including compression (Hassanlou et al. 2013; Wu et al. 2014), influence analysis (Shi et al. 2015; Adhikari et al. 2017) and so on. For a more comprehensive review of existing publications regarding these goals, we refer the interested readers to a survey paper by Liu et al. (2018).

Graph clustering Prior methods of clustering attributed graphs seek to partition the given graph into clusters with cohesive intra-cluster structures and homogeneous attribute values. Some enforce homogeneity in all attributes (Akoglu et al. 2012; Zhou et al. 2009; Xu et al. 2012; Cheng et al. 2011). However, they are not guaranteed to reveal meaningful patterns in datasets without efforts of attribute selection, since irrelevant attributes can strongly obfuscate clusters. More recently, subspace clustering is used to loosen this constraint (Günnemann et al. 2010; Günnemann et al. 2011). Perozzi et al. (2014) detects *focused* clusters and outliers based on user preferences, allowing the user to control the relevance of attributes and as a consequence, the graph mining results. Wang et al. (2016) proposes a novel nonnegative matrix factorization (NMF) model in which sparsity penalty is introduced to select the most related attributes for each cluster.

Unlike all previous graph summarization or clustering methods where the resulting vertex groups are forced to satisfy some pre-specified topologies or edges structures (e.g., being more densely connected within the group), patterns revealed in our summarization approach are not limited to that, as their interestingness is quantified by a subjective measure depending on the user's prior expectation.

3 Subgroup pattern and summary syntaxes for graphs

In this section we introduce both single subgroup and bi-subgroup patterns along with summaries 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, $E \subseteq \binom{V}{2}$ 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 categorical Dom_a and for each $y \in \text{Dom}_a$, we introduce a Boolean function $s_{a,y} : V \to \{\text{true}, \text{false}\}$, with $s_{a,y}(v) \triangleq \text{true}$ for $v \in V$ iff $a(v) = y$. Analogously, for each $a \in A$ with $\text{Dom}_a \subseteq \mathbb{R}$ and for each $l, u \in \text{Dom}_a$ such that $l < u$, we define $s_{a,[l,u]} : V \to \{\text{true}, \text{false}\}$, with $s_{a,[l,u]}(v) \triangleq \text{true}$ iff $a(v) \in [l, u]$. We call these Boolean 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

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

subset of vertices that satisfy it: $\varepsilon(W) \triangleq \{v \in V | W(v) = \text{true}\}$. We also informally 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,

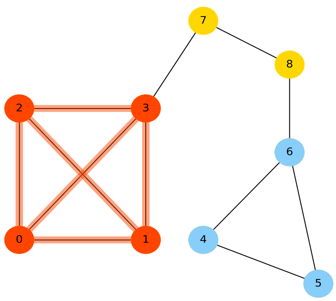
(i) $V_W = \varepsilon(W)$, i.e., the set of vertices from V that is the extension of the description W;
(ii) $E_W = \binom{V_W}{2} \cap E$, i.e., the set of edges from E that have both endpoints in V_W.

Example 1 Figure 1 displays an example attributed graph $G = (V, E, A)$ with $n = 9$ vertices, $m = 12$ edges (Graph in Fig. 1a, vertex attributes in Fig. 1b). Each vertex is annotated with one real-valued attribute (i.e., a) and three nominal (or for simplicity, binary) attributes (i.e., b,c,d). Consider a description $W = s_{a,[0,3]} \wedge s_{b,1}$. The extension of this description is the set of vertices with attribute a value from 0 to 3 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 (dark in greyscale) in Fig. 1a).

3.1 Local pattern

3.1.1 Single-subgroup pattern

A first pattern syntax we consider, and which has already been studied in prior work, 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



(a) Graph

Vertex	a	b	c	d
0	0.6	1	0	1
1	1.5	1	0	0
2	2.2	1	0	1
3	1.4	1	1	1
4	2.5	0	0	0
5	4.7	0	0	0
6	4.3	0	0	0
7	3.6	1	1	1
8	3.8	1	1	0

(b) Vertex attributes

Fig. 1 Example attributed graph with 9 vertices (0–8) and 4 associated attributes (**a**–**d**). The subgraph induced by the description ($W = s_{a,[0,3]} \wedge s_{b,1}$) is highlighted in red (dark in greyscale) the number of edges E_W in subgraph $G[W]$ induced by W is greater (or less) than k_W. Thus, $I = 0$ indicates the induced subgraph is dense, whereas $I = 1$ characterizes a sparse 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. One example of a single-subgroup pattern in Fig. 1 can be $(s_{a,[0,3]} \wedge s_{b,1}, 0, 6)$, corresponding to the dense subgraph highlighted in red (dark in greyscale).

Remark 1 (*Difference to dense subgraph pattern in* van Leeuwen et al. (2016)) Though the syntax for our single-subgroup pattern seems similar to that of the dense subgraph pattern (i.e., (W, k_W)) proposed by van Leeuwen et al. (2016), they are essentially different definitions serving for different data mining tasks. In van Leeuwen et al. (2016), the aim is to identify subjectively interesting subgraphs based on merely link information. For this aim, W in the dense subgraph pattern syntax represents the set of vertices in the subgraph, which has no association with node attributes. Moreover, an indicator I is included in our pattern syntax. This allows to regard not only surprisingly dense subgraphs but also surprisingly sparse ones as interesting. In contrast, van Leeuwen et al. (2016) focuses on those surprisingly dense subgraphs. Because of these differences in W and I, k_W is different accordingly.

3.1.2 Bi-subgroup pattern

We also define a pattern syntax informing the analyst about the edge density between two potentially 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 $n_W \triangleq |\varepsilon(W_1)||\varepsilon(W_2)| - \frac{1}{2}|\varepsilon(W_1 \wedge W_2)|(|\varepsilon(W_1 \wedge W_2)| + 1)$ for undirected graphs without self-edges. For example, the bi-subgroup pattern $(s_{a,[0,3]} \wedge s_{b,1}, s_{b,0}, 1, 0)$ in Fig. 1, expresses sparse (or more precisely, zero) connection between the red vertex group (i.e., $\{0, 1, 2, 3\}$) and the blue one (i.e., $\{4, 5, 6\}$) . Note that single-subgroup patterns are a special case of bi-subgroup patterns when $W_1 \equiv W_2$.

Remark 2 (*Setting of* k_W) 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 focuses 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.

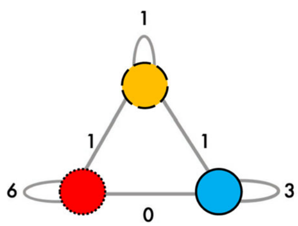
3.2 Global pattern: summarization for graphs

Here we define a global pattern syntax, which describes the edge density between any pair of subgroups selected from a set of subgroups that form a partition of the vertices. We first define the notion of a *summarization rule*, before introducing the global pattern syntax itself.

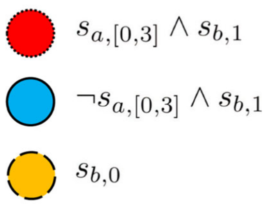
Definition 2 *(Summarization rule for an attributed graph)* Given an attributed graph $G = (V, E, A)$, *the summarization rule* $\mathbb{S}$ of G is a set of descriptions such that their extensions are vertex-clusters that form a partition of the whole vertex set. That is, $\mathbb{S} = \{W_i | i = 1, 2, \ldots, c\}$ where $c \in \mathbb{N}$ is the number of disjoint vertex-clusters, where $\cup_{i=1}^{c} \varepsilon(W_i) = V$, $\forall W_i \in \mathbb{S}$ it holds that $\varepsilon(W_i) \neq \emptyset$, and $\forall W_i, W_j \in \mathbb{S}, i \neq j$ it holds that $\varepsilon(W_i) \cap \varepsilon(W_j) = \emptyset$.

Definition 3 (*Summary for an attributed graph based on a summarization rule*) A summary $\mathcal{S}$ for an attributed graph $G = (V, E, A)$ based on a summarization rule $\mathbb{S} = \{W_i | i = 1, 2, \ldots, c\}$ is *a complete weighted graph* $\mathcal{S} = (V^{\mathbb{S}}, E^{\mathbb{S}}, w)$ with weight function $w : E^{\mathbb{S}} \to \mathbb{R}$, whereby $V^{\mathbb{S}} = \{\varepsilon(W) | W \in \mathbb{S}\}$ is the set of vertices (referred to as *supervertices* of the original graph G, i.e. each vertex from $\mathcal{S}$ is a set of vertices from G), $E^{\mathbb{S}} = \binom{V^{\mathbb{S}}}{2} \cup V^{\mathbb{S}}$ is the set of edges (to which we refer as *superedges*; the superedges in $\binom{V^{\mathbb{S}}}{2}$ represent the undirected edges between distinct supervertices, and the superedges in $V^{\mathbb{S}}$ represent the self-loops). The weight $w(\{\varepsilon(W_i), \varepsilon(W_j)\})$ for each superedge $\{\varepsilon(W_i), \varepsilon(W_j)\} \in E^{\mathbb{S}}$ will be denoted shorthand by $d_{i,j}$, and is defined as the number of edges between vertices from $\varepsilon(W_i)$ and those from $\varepsilon(W_j)$.

We define a global pattern syntax informing the analyst about the summarization for an attributed graph $G = (V, E, A)$ with c disjoint vertex-clusters. More formally, we define a *summarization pattern* as a tuple $(\mathbb{S}, \mathcal{S})$ where $\mathbb{S}$ is the summarization rule, and $\mathcal{S}$ is the corresponding summary. Note that when revealing a summarization pattern $(\mathbb{S}, \mathcal{S})$ to an analyst, she or he gets access to its related local subgroup patterns: c single-subgroup patterns and $c(c-1)/2$ bi-subgroup patterns. An example of the global pattern for Fig. 1 can be $(\{s_{a,[0,3]} \wedge s_{b,1}, \neg s_{a,[0,3]} \wedge s_{b,1}, s_{b,0}\}, \mathcal{S}^*)$ where $\mathcal{S}^*$ represents the corresponding summary (see Fig. 2).



(a) Summary $\mathcal{S}^*$.



(b) Descriptions in the summarization rule.

Fig. 2 Example summarization pattern for Fig. 1 with the summarization rule $\{s_{a,[0,3]} \wedge s_{b,1}, \neg s_{a,[0,3]} \wedge s_{b,1}, s_{b,0}\}$ and the summary $\mathcal{S}^*$. This summary $\mathcal{S}^*$ is composed of three supervertices each of which corresponds to a set of vertices satisfying $s_{a,[0,3]} \wedge s_{b,1}$ (red circle with dotted line), $\neg s_{a,[0,3]} \wedge s_{b,1}$ (blue circle with solid line), $s_{b,0}$ (yellow circle with dashed line) respectively, and superedges each of which connects one supervetex to the other with a weight representing the number of edges between them

4 Formalizing the subjective interestingness

4.1 General approach

We follow the approach as outlined by De Bie (2011b) to quantify the subjective interestingness of a pattern, which enables us to account for prior beliefs a data analyst may hold about the data. In this framework, the analyst's belief state is modeled by a *background distribution* P 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 probability for any given pattern to be present in the data, and thus to assess the surprise of 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. 4.2.

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 the 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 length of the code needed to communicate the pattern to the analyst. These are discussed in more detail in Sect. 4.3, but first we further explain the background distribution (Sect. 4.2).

Remark 3 (*Positioning with respect to directly related literature*) Here we clarify how previous work is leveraged, and what concepts are newly introduced in our work. We define single/bi-subgroup patterns and global patterns in an attributed graph. To quantify the SI measure for such patterns, we follow the framework outlined by De Bie (2011b). As mentioned above, in this framework, the SI is computed as the ratio of the IC and the DL w.r.t. the background distribution which models the analyst's belief state. This framework also provides the general idea for deriving the initial background distribution and updating it to reflect newly acquired knowledge. Adriaens et al. (2017) later introduced a new type of graph-related prior that the background distribution can incorporate, and this prior is considered in our work. In van Leeuwen et al. (2016), this framework was used to identify subjectively interesting dense subgraphs, merely based on link information. In our work, we leverage some computational results from van Leeuwen et al. (2016) (i.e., in updating the background distribution, approximating the IC), and made further adaptions such that the framework proposed by De Bie (2011b) can serve for our newly proposed patterns based on attribute information (i.e., single-subgroup patterns, bi-subgroup patterns and global patterns).

4.2 The background distribution

4.2.1 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 which are common in practice: (1) on individual vertex degrees; (2) on the overall graph density; (3) on densities between bins (particular subsets of vertices).

(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, De Bie (2011b) showed that the maximum entropy distribution is a product of independent Bernoulli distributions, one for each of the random variable $b_{u,v}$, which equals to 1 if $(u, v) \in E$ and 0 otherwise. Denoting the probability that $b_{u,v} = 1$ by $p_{u,v}$, this distribution is of the form:

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

This can be conveniently expressed as:

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

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.

(3) *Additional prior beliefs on densities between bins.* We can partition vertices in an attributed graph into bins according to their value for a particular attribute. For example, vertices 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 are 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, a discovered pattern would be more informative if it contrasts more with this kind of belief, i.e. if it reveals a high density between two sets of people from different class years. As shown in Adriaens et al. (2017), the resulting background distribution is also a product of Bernoulli distributions, one for each of the random variables $b_{u,v} \in \{0, 1\}$:

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

where $k_{u,v}$ is the index for the block corresponding to the intersecting part of two bins which vertex u and vertex v belongs to correspondingly. λ_u^r, λ_v^c and $\gamma_{k_{u,v}}$ are parameters and can be computed efficiently. Note our model is not limited to incorporate this type of belief related to a single attribute. Vertices 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[2].

4.2.2 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 subgroup 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 $\varepsilon(W_2)$. De Bie (2011a) presented an argumentation for choosing 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. (2016) showed that the resulting P' is again a product of Bernoulli distributions:

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

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

How to compute λ_W is also given in van Leeuwen et al. (2016).

Remark 4 (*Updating P if a summarization pattern is presented*) In the case that a summarization pattern $(\mathbb{S}, \mathcal{S})$ is presented to the analyst, we simply update the background distribution as if all the subgroup patterns related to $(\mathbb{S}, \mathcal{S})$ were presented, and we denote such updated background distribution by $P_{(\mathbb{S}, \mathcal{S})}$.

4.3 The subjective interestingness measure

We now discuss how the SI measure can be formalized by relying on the background distribution, first for local and then for global patterns.

4.3.1 The SI measure for a local pattern

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

[2] Simply by replacing $\gamma_{k_{u,v}}$ in Eq. 1 with $\sum_{i=1}^{i=h} \gamma_{k^i_{u,v}}$ where h is the number of attributes considered (also the number of partitions).

of getting more than k_W (for $I = 0$) or $n_W - k_W$ (for $I = 1$) successes in n_W trials with possibly different 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 $1 - p_{u,v}$. The work of van Leeuwen et al. (2016) proposed to tightly upper bound the probability of a similar dense subgraph pattern by applying the general Chernoff/Hoeffding bound (Chernoff 1952; Hoeffding 1963). Here, we can use the same approach, which gives:

$$\mathbf{Pr}[(W_1, W_2, I = 0, k_W)] \leq \exp\Big(-n_W \mathbf{KL}\Big(\frac{k_W}{n_W} \parallel p_W\Big)\Big),$$
$$\mathbf{Pr}[(W_1, W_2, I = 1, k_W)] \leq \exp\Big(-n_W \mathbf{KL}\Big(1 - \frac{k_W}{n_W} \parallel 1 - p_W\Big)\Big),$$

where

$$p_W = \frac{1}{n_W} \sum_{u \in \varepsilon(W_1), v \in \varepsilon(W_2)} p_{u,v}. \tag{2}$$

$\mathbf{KL}\Big(\frac{k_W}{n_W} \parallel p_W\Big)$ is the Kullback-Leibler divergence between two Bernoulli distributions 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, regardless of I:

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

The information content is the negative log probability of the pattern being present under the background distribution. Thus, using the above:

$$\begin{aligned}\mathrm{IC}[(W_1, W_2, I, k_W)] &= -\log(\mathbf{Pr}[(W_1, W_2, I, k_W)]),\\ &\geq n_W \mathbf{KL}\Big(\frac{k_W}{n_W} \parallel p_W\Big).\end{aligned} \tag{3}$$

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 then be written as

$$\mathrm{DL}[(W_1, W_2, I, k_W)] = \alpha(|W_1| + |W_2|) + \beta. \tag{4}$$

The subjective interestingness (SI) In summary, we obtain:

$$\begin{aligned}\mathrm{SI}[(\mathbb{S}, \mathcal{S})] &= \frac{\mathrm{IC}[(W_1, W_2, I, k_W)]}{\mathrm{DL}[(W_1, W_2, I, k_W)]}, \\ &= \frac{n_W \mathbf{KL}\left(\frac{k_W}{n_W} \parallel p_W\right)}{\alpha(|W_1| + |W_2|) + \beta}.\end{aligned} \tag{5}$$

Remark 5 (*Justification about choices of* α *and* β) In all our experiments for use cases, we apply $\alpha = 0.6$, $\beta = 1$. We here state the reason for this choice.

In practice, the absolute value of the SI from Eq. 5 is largely irrelevant, as it is only used for ranking the patterns, or even just for finding a single pattern (i.e., the most interesting one to the analyst). Thus, we can set $\beta = 1$ without losing generality, such that the only remaining parameters is α.

Tuning α biases the results toward more or fewer selectors to describe the subgroup pattern. Notice an optimal extent of such kind of bias cannot be determined by doing model selection in the statistical sense, but rather should be chosen based on aspects of human cognition (e.g., larger α should be used when the analyst prefers patterns in a more succinct form). In this work, we set $\alpha = 0.6$ throughout all use cases which gives qualitative results. However, α can be flexibly tuned for adapting to the analyst' preferences.

4.3.2 The SI measure for a global pattern

The information content (IC) The probability of a global summarization pattern turns out to be harder to formulate analytically, and thus also the negative log probability of the pattern – which is the subjective amount of information gained by observing the pattern. However, it is relatively straightforward to quantify the (subjective) amount of information in the connectivity in the graph prior to observing the pattern, and after observing the pattern. The difference between these two is thus the information gained. More formally, we thus mathematically define the IC of a summarization pattern $(\mathbb{S}, \mathcal{S})$ as the difference between the log probability for the connectivity in the graph (i.e., the edge set E) under $P_{(\mathbb{S},\mathcal{S})}$ and that under P:

$$\mathrm{IC}[(\mathbb{S}, \mathcal{S})] = \log P_{(\mathbb{S},\mathcal{S})}(E) - \log P(E). \tag{6}$$

This quantity is straightforward to compute where $P_{(\mathbb{S},\mathcal{S})}$ is computed as the updated background distribution as if all the subgroup patterns related to $(\mathbb{S}, \mathcal{S})$ were presented (previously mentioned in Remark 4 in Sect. 4.2.2).

The description length (DL). We search for optimal $\mathbb{S}$ by a strategy that is based on splitting a binary search tree (for details see Sect. 5.2.1). Thus, the cost for the data analyst to assimilate $\mathbb{S}$ is linear to the number of descriptions in $\mathbb{S}$, i.e. c. As for $\mathcal{S}$, assimilating it costs quadratically to c, because $\mathcal{S}$ is essentially a complete graph with c vertices and $c(c+1)/2$ edges. The total description length of a pattern $(\mathbb{S}, \mathcal{S})$ can be written as

$$\mathrm{DL}[(\mathbb{S}, \mathcal{S})] = \zeta \cdot c(c+1)/2 + \eta \cdot c + \theta. \tag{7}$$

where θ is a constant term for mitigating the quadratically increasing drop in SI value given by an increasing c, and this helps to avoid early stopping.
The subjective interestingness (SI) In summary, we obtain:

$$\begin{aligned}\mathrm{SI}[(\mathbb{S}, \mathcal{S})] &= \frac{\mathrm{IC}[(\mathbb{S}, \mathcal{S})]}{\mathrm{DL}[(\mathbb{S}, \mathcal{S})]},\\ &= \frac{\log P_{(\mathbb{S},\mathcal{S})}(E) - \log P(E)}{\zeta \cdot c(c+1)/2 + \eta \cdot c + \theta}.\end{aligned} \tag{8}$$

Remark 6 (*Justification about choices of* ζ, η *and* θ) In all our experiments, we use $\zeta = 0.02, \eta = 0.02, \theta = 1$. As stated in Remark 5 in Sect. 4.3.1, parameters of the DL indicate how much the data analyst prefers patterns that can be described succinctly, and thus should be determined based on aspects of human cognition instead of statistical model selection. We here follow the similar sense to choose the DL parameters for global patterns (i.e.,ζ, η and θ in Eq. 8). Notice we set a high value for θ (i.e., 1) in comparison with ζ (i.e., 0.02) and η (i.e., 0.02). This is a safe choice to avoid early stopping (i.e., the iterating stops before the analyst observes a suitable global pattern).

5 Algorithms

This section describes the algorithms for mining interesting patterns locally and globally, in Sects. 5.1 and 5.2 respectively, followed by an outline to the implementation in Sect. 5.3.

5.1 Local pattern mining

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. 5.1.1). 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. 5.1.2.

5.1.1 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. This approach is standard practice in subgroup discovery, being the search algorithm implemented in popular packages such as Cortana (Meeng and Knobbe 2011), One Click Miner (Boley et al. 2013), and pysubgroup (Lemmerich and Becker 2018).

5.1.2 Nested beam search

The basic idea of this approach is to nest one beam search into the other one where the outer search branches based on a 'beam' of promising selector candidates for the description W_1 , and the inner search expands those for W_2. The detailed procedure for this nested beam search is shown in Algorithm 1, and related notation displayed in Table 1.

The total number of interesting patterns identified by Algorithm 1 is $x_1 \cdot x_2$. Note that we deliberately constrain the beam to contain at least x_1 different W_1 descriptions so that a sufficient diversity among all the discovered patterns is guaranteed (see lines 22–23 in Algorithm 1).

5.2 Global pattern mining

To identify the most interesting global (or summarization) pattern, a greedy search strategy (see Sect. 5.2.1) equipped with some speedup strategies (see Sect. 5.2.2) are adopted.

Table 1 Notations for Algorithm 1

Notation	Description
OuterBeam	The outer beam storing best description pairs (W_1, W_2) during the search
InnerBeam	The inner beam only storing best descriptions W_2
x_1	The outer beam width (i.e., the minimum number of different descriptions W_1 contained in the outer beam
x_2	The inner beam width
D	The search depth (i.e., maximum number of selectors combined in a description)

Algorithm 1: Subjectively Interesting BiSubgroup Pattern Mining

```
input : Graph $G = \{V, E, A\}, x_1, x_2, D$
output: Top $x_1 \cdot x_2$ bi-subgroup patterns contained in OuterBeam

$S \leftarrow$ the set of all selectors to build descriptions from;
OuterBeam $\leftarrow \{\emptyset\}$ ;
$d_1 \leftarrow 0$;
$d_2 \leftarrow 0$;
while $d_1 < D$ do // The outer search
    $\mathbb{C}_1 \leftarrow$ all the $W_1$ candidates in OuterBeam;
    for $C_1 \in \mathbb{C}_1$ do // Expand on $W_1$ candidates
        for $s_1 \in S$ do
            $Z_1 \leftarrow C_1 \wedge s_1$;
            InnerBeam $\leftarrow \{\emptyset\}$;
            while $d_2 < D$ do // The inner search
                $\mathbb{C}_2 \leftarrow$ all the $W_2$ candidates in InnerBeam;
                for $C_2 \in \mathbb{C}_2$ do // Expand $W_2$ candidates
                    for $s_2 \in S$ do
                        $Z_2 \leftarrow C_2 \wedge s_2$;
                        $k_W \leftarrow$ the number of edges between vertices $\varepsilon(Z_1)$ and $\varepsilon(Z_2)$;
                        // compute SI of the pattern $(Z_1, Z_2, I, k_W)$
                           using Eq. 5
                        si$' \leftarrow$ SI$[(Z_1, Z_2, I, k_W)]$;
                        // Add (si$'$, $Z_2$) to the InnerBeam if InnerBeam
                           contains less than $x_2$ elements or replace
                           the tuple with the smallest SI in
                           InnerBeam if si$'$ is larger than that
                           value
                        InnerBeam $\leftarrow$ UpdateBeam(InnerBeam, (si$'$, $Z_2$), $x_2$);
                $d_2 \leftarrow d_2 + 1$
            for (si, $Z$) $\in$ InnerBeam do
                // Add (si, $Z_1$, $Z$) to the OuterBeam if the number of
                   various $W_1$ descriptions in OuterBeam is less
                   than $x_1$ or replace the tuple with the smallest
                   SI if si is larger than that value
                OuterBeam $\leftarrow$ UpdateBeam(OuterBeam, (si, $Z_1$, $Z$), $x_1$);
    $d_1 \leftarrow d_1 + 1$
```

5.2.1 The basic search strategy

The algorithm begins by checking each possible summarization rule only containing a single-selector description and its negation. Applying such a rule at the beginning means cutting the whole vertex set into two non-overlapping clusters, each of which satisfies a description in this rule correspondingly. The rule whose corresponding summarizaiton pattern has the maximal SI value is selected as a seed set for $\mathbb{S}$. Then the algorithm iterates in the following way to greedily grow that set: for each existing description in the set, the algorithm again checks the application of an additional single-selector description and its negation. This further separates a particular vertex cluster into two sub-clusters, one of which additionally satisfies this description and the other does not. The optimal combination of the existing description to further specify and the additional single-selector description are selected. The search stops when reaching some search budget (e.g. the maximum number of iterations). The detailed procedure for this search is displayed in Algorithm 2.

Algorithm 2: Interesting Summarization Pattern Mining

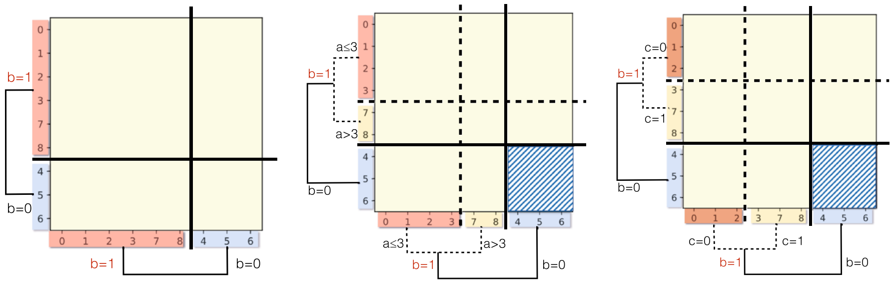
input : Graph $G = \{V, E, A\}$, Search Iteration budget D
output: $(\mathbb{S}, \mathcal{S})$

1 $\mathbb{S} \leftarrow \{\emptyset\}$;
2 VertexClusters $\leftarrow \{V\}$ `// A set of vertex-clusters each of which is formed by the extension of a description in` $\mathbb{S}$`. Initially, it is a set only containing one member, the whole vertex set;`
3 $i \leftarrow 0$ `// The number tracking the iteration round;`
4 **while** $i < D$ **do**
5   si$\leftarrow -\infty$;
6     $\mathbb{S}' \leftarrow \mathbb{S}$;
7     $\mathbb{S}'' \leftarrow \mathbb{S}$;
8   VertexClusters$'$ $\leftarrow$ VertexClusters;
9   **for** $W \in \mathbb{S}$ **do** `// Iterate over each description rule currently in` $\mathbb{S}$
10   **for** $a \in A$ **do** `// Iterate over each attribute`
11             $S_a \leftarrow$ the set of all selectors associated with the attribute a;
12   **for** $s \in S_a$ **do** `// Iterate over each selector of the attribute` a
13   `// Update` $\mathbb{S}'$ `by replacing` W `with two more specific descriptions such that one additionally satisfies` s`, and the other does not`
14                 $\mathbb{S}' \leftarrow \mathbb{S}' \setminus \{W\} \cup \{W \wedge s, W \wedge \neg s\}$;
15   `// Update VertexClusters' correspondingly`
16   VertexClusters$'$ $\leftarrow$ VertexClusters$'$ $\setminus \{\varepsilon(W)\} \cup \{\varepsilon(W \wedge s), \varepsilon(W \wedge \neg s)\}$;
17                 $\mathcal{S}' \leftarrow$ A summary of $G = \{V, E, A\}$ based on the summarization rule $\mathbb{S}$;
18   si$' \leftarrow$ SI$[(\mathbb{S}', \mathcal{S}')]$;
19   **if** si$'$ $>$si **then**
20   si$\leftarrow$si$'$;
21                     $\mathbb{S} \leftarrow \mathbb{S}'$;
22   VertexClusters$\leftarrow$VertexClusters$'$;
23                     $\mathcal{S} \leftarrow \mathcal{S}'$
24                 $\mathbb{S}' \leftarrow \mathbb{S}''$ `// Revert to` $\mathbb{S}''$`;`
25     $i++$;

5.2.2 Speedup strategies

Parallel processing Our search strategy is trivially parallelizable. To gain some speedup, the search process for each attribute and its related selectors (lines 10–24 in Algorithm 2) is executed simultaneously in multiple processors.

Reusing some computations We further speedup the search by circumventing some redundant computations when computing the SI for each candidate of summarization



(a) The adjacency matrix before branching.

(b) Branching from $b = 1$ in a way.

(c) Branching from $b = 1$ in another way.

Fig. 3 Illustration of the existence of a common subgroup pattern when branching in two different ways

pattern. As mentioned above in Sect. 4.2.2, $P_{(\mathbb{S},\mathcal{S})}$ is computed as an updated background distribution as if all the subgroup patterns related to $(\mathbb{S}, \mathcal{S})$ were presented, which requires to determine λ_W for each related subgroup pattern. Nevertheless, when branching in different ways during the search (i.e., using different pairs of a selector and its negation to extend a given description), extensions do not interfere with subgroup patterns whose descriptions are not extended. Hence, their λ_W do not need to be recomputed, providing a speed up.

Here we illustrate that, by taking the attributed network in Fig. 1 as the example (see Fig. 3 which visualizes the corresponding adjacency matrix with arranged vertex indices in left and in bottom; Entries are not indicated for simplicity). Assume the network is currently divided into two vertex subgroups each respectively satisfying $b = 1$ and $b = 0$, and the search is in the step of finding the optimal selector to specify the description $b = 1$ (indices of corresponding vertices are highlighed in red (dark in greyscale) in Fig. 3a). Though the adjacency matrix is cut in two different ways, refining the description $b = 1$ into two more specific ones by adding $a \leq 3$ and $a > 3$ (in Fig. 3b), or adding $c = 0$ and $c = 1$ (in Fig. 3c), both do not interfere with the subgroup satisfying $b = 0$ (the blue striped area).

5.3 Implementation

For mining pattern locally, *Pysubgroup* (Lemmerich and Becker 2018), 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 (along with other state-of-the-art interestingness measures for comparison) into this original interface. A Python implementation of all the algorithms and the experiments is available at https://bitbucket.org/ghentdatascience/globalessd_public. All experiments were conducted on a PC with Ubuntu OS, Intel(R) Core(TM) i7-7700K 4.20GHz CPUs, and 32 GB of RAM.

6 Experiments

We evaluate our methods on six real-world networks. In the following, we first describe the datasets (Sect. 6.1). Then we present the conducted experiments and discuss the results with a purpose to address the following questions:

RQ1 Are our local pattern mining algorithms sensitive to the beam width? (Sect. 6.2)
RQ2 Does our SI measure outperform state-of-the-art objective interestingness measures? (Sect. 6.3)
RQ3 Is the SI truly subjective, in the sense of being able to consider a data analyst's prior beliefs? (Sect. 6.4)
RQ4 How can optimizing SI help avoid redundancy between iteratively mined patterns? (Sect. 6.5)
RQ5 Is our global pattern mining approach able to summarize the whole graph in a meaningful way such that all the interesting patterns can be revealed? (Sect. 6.6)
RQ6 How do the algorithms scale? (Sect. 6.7)

6.1 Data

Basic data information is summarized in Table 2.
Caltech36 and Reed98 Two Facebook social networks from the Facebook100 (Traud et al. 2012) data set, gathered in September 2005: one for Caltech Facebook users, and one for Reed University. Vertex attributes describe the person's status (faculty or student), gender, major, minor, dorm/house, graduation year, and high school.
Lastfm A social network of friendships between `Lastfm.com` users, generated from the publicly available dataset (Cantador et al. 2011) in the HetRec 2011 workshop. In this dataset, tag assignments of a list of most-listened musical artists provided by each user are given in [user, tag, artist] tuples, where those tags are unstructured text labels that users used to express songs of artists. We then took tags that a user ever assigned to any artist and assigned those to the user as binary attributes expressing a user's music interests. This dataset has been used in many publications to evaluate local pattern mining methods (Pool et al. 2014; Atzmueller et al. 2016; Galbrun et al. 2014).
DBLPtopics A citation network generated from the DBLP citation data V11[3] (Tang et al. 2008; Sinha et al. 2015) by choosing a random subset of publications from 20 conferences[4] selected to cover 4 research areas: Machine Learning, Database, Information Retrieval, and Data Mining. Vertices represent publications, and directed edges represent citation relationships. Each publication is annotated with 50 attributes (denoted by $a_1, a_2, \ldots, a_{50}$) whose value indicates the relevance of this paper to a certain topic. These attributes are obtained by computing the first 50 *latent semantic indexing (LSI)* components for the original paper-topic matrix (of size 10837×9074)

[3] This citation dataset are extracted from DBLP website: https://dblp.uni-trier.de/, containing 4107340 publications (from unknown year till May 2019) and 36624464 citation relationships. It can be accessed by: https://aminer.org/citation.

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

Table 2 Dataset statistics summary

Dataset	Type	$\|V\|$	$\|E\|$	Attribute type	#Attributes	$\|S\|$
Caltech36	Undirected	762	16, 651	Nominal	7	602
Reed98	Undirected	962	18, 812	Nominal	7	748
Lastfm	Undirected	1892	12, 717	Binary	11, 946	21, 695
DBLPtopics	Directed	10, 837	6883	Numerical	50	300
DBLPaffs	Directed	6472	3066	Binary	116	232
MPvotes	Undirected	650	49, 631	Binary	39	78

where each entry value indicates the relevance of a paper (represented by row) to a field of study (represented by column) and this value is provided by the original DBLP data. In our work, the selector space on which the search is carried does not include every attribute value pair. A discretization is applied here: values for each attribute are sorted and discretized into 4 partitions of equal size by 3 quartiles. This gives $3 \times 2 = 6$ selectors for each attribute ($6 \times 50 = 300$ selectors in total) three of which respectively assign *true* to vertices with value smaller than the first, second, third quartile of the total values for this attribute, and the other three are the corresponding negations. We denote the i-th quartile of values for the attribute a by Q_i^a.

DBLPaffs A DBLP citation network based on a random subset of publications same as the one for the above task. Only papers for which the authors' country (or state, in the USA) of affiliation is available are included as vertices. The resulting 116 countries/states are included as binary vertex attributes, set to 1 iff one of the paper's authors is affiliated to an institute in that country/state.

MPvotes The Twitter social network generated from friendships between Members of Parliament (MPs) in UK (Chen et al. 2020). Their voting records on Brexit from 12th June 2018 to 3rd April 2019 are included as 39 binary vertex attributes, set to be 1, or -1 iff this MP vote for/abstain or, against/abstain respectively. Note we include abstain on both positive and negative sides rather than make abstain (or not abstain) alone being a value, because a selector that describes a subgroup of MPs abstaining (or not abstaining) in a particular vote is not very meaningful in practice.

6.2 Parameter sensitivity (RQ1)

For mining local patterns, we used the standard beam search for single-subgroup patterns, and the nested beam search for bi-subgroup patterns. In all experiments, we set the search depth $D = 2$ (because patterns that are described by more than 2 selectors often appear less interesting in practice, and they would add unnecessary difficulty for interpretation). Then the performance of those beam search methods ultimately depends on the beam width.

6.2.1 Experimental setup

Choice of datasets We used *Lastfm* to investigate the effect of the beam width on the performance of single-subgroup pattern mining, as it involves the largest search space (given by the largest number of selectors i.e., 21695). With regard to that on bi-subgroup pattern mining, because the search is more time-consuming, we used *Lastfm* while only considering 100 most frequently used tags as attributes (i.e., giving 200 selectors as the search space). We also used *Reed98* as it involves the largest search space among datasets that were used in our experiments on bi-subgroup pattern mining.

Other settings Though we applied the SI measure with $\alpha = 0.6$, $\beta = 1$ in all use cases of local pattern mining (as previously mentioned in Remark 5 in Sect. 4.3.1), to more meaningfully investigate the parameter sensitivity in this experiment, we set α to be smaller, i.e., $\alpha = 0.1$.[5]

6.2.2 Results

Effect of the beam width on single-subgroup pattern mining First, we analyze the sensitivity of the standard beam search w.r.t. the beam width for single-subgroup pattern mining. How the search performance changes with the beam width (denoted by x) is illustrated below (see Fig. 4a for the SI value of the identified best pattern and Fig. 4b for the run time).

Clearly, increasing x from 1 to 40 results in the same best pattern (with the SI value as 258.7, the description as 'IDM = 1') along with a gentle increase in the run time. Though it shows a greedy search (i.e., $x = 1$) can already perform well, this is not guaranteed.

As indicated in a further investigation, increasing the beam width is rendered useless by the existence of a dominant pattern with a single selector (i.e., 'IDM = 1') such that there are no other patterns that have higher SI value than it and its children. Once our method incorporates this dominant pattern into the background distribution for one subsequent iteration to reflect the data analyst's newly acquired knowledge, the advantage of a lager beam width appears as the best pattern is identified when x increases to be 3 (see Fig. 5a). The run time grows linearly as x increases (see Fig. 5b).

Effect of the beam width on bi-subgroup pattern mining To study the effects of the beam width, we implemented all cases with x_1 and x_2 being 1, 2, 3, 4, or 7.

In *Lastfm*, clearly from Fig. 6a, small beam widths (e.g., when $x_1 = 1$ with $x_2 = 3$) are sufficient for our algorithm to identify the best bi-subgroup pattern (i.e., the one with SI as 194.8). This is even more the case for *Reed98* network, as our method of bi-subgroup pattern mining always identify the same best bi-subgroup pattern (i.e., the one with SI as 728) when gradually increasing x_1 and x_2.

5 In this sensitivity investigation, applying a relatively larger α (e.g., $\alpha = 0.6$) can more possibly lead to positive results (i.e., showing the insensitivity as the same best pattern is always identified while varying the beam width) but by a fluke: setting α *larger* in the SI measure penalizes more complex patterns *more heavily*, and this makes the best pattern found before further branching in a beam search more easily dominate, giving less credible positive results. We thus safely chose α to be 0.1 in this experiment.

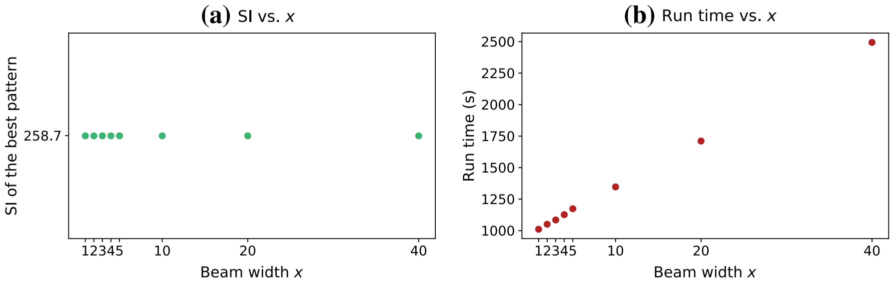


Fig. 4 Varying the beam width x in the search for single-subgroup patterns in *Lastfm*

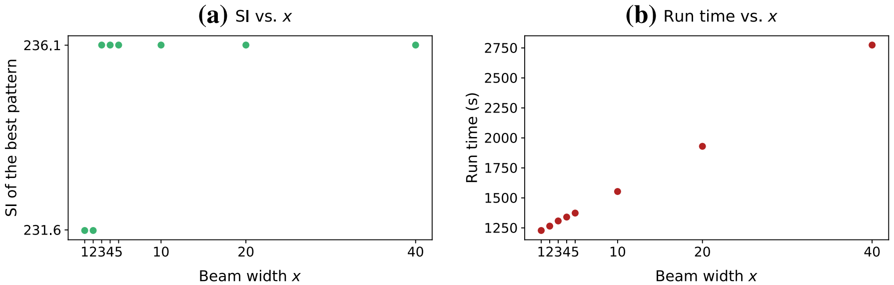


Fig. 5 Varying the beam width x in the search for single-subgroup patterns in *Lastfm* after incorporating the dominant pattern described by 'IDM = 1'

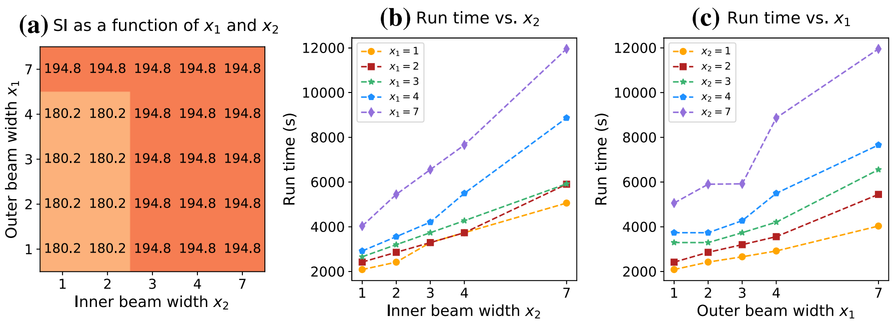


Fig. 6 Varying the outer/inner beam width x_1/x_2 in the search for bi-subgroup patterns in *Lastfm*

For bi-subgroup pattern mining in either *Lastfm* or *Reed98*, the run time experiences an approximately linear growth as x_1 or x_2 increases with the other beam width is fixed (see Fig. 6b and c for *Lastfm*, Fig. 7b and c for *Reed98*).

Summary This empirical analysis suggests that overall our algorithms are not sensitive to the beam width. A small beam width is usually sufficient, particularly if there is a dominant pattern. When that is not the case, slightly increasing the beam width was sufficient in our experiments.

We recommend an initial setting with $x = 5$ for single-subgroup pattern discovery and $x_1 = 2$, $x_2 = 3$ for bi-subgroup pattern discovery, which is usually more than

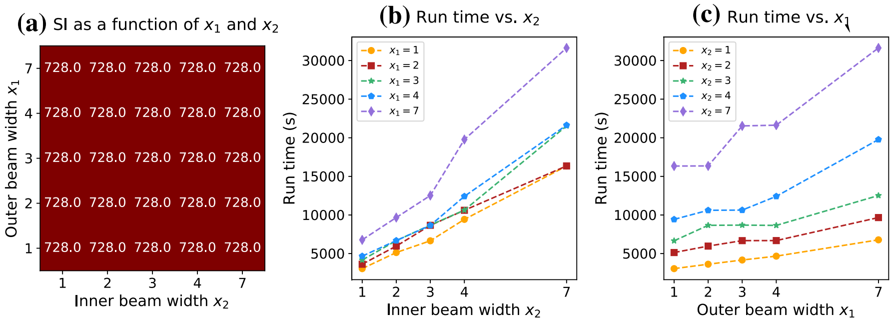


Fig. 7 Varying the outer/inner beam width x_1/x_2 in the search for bi-subgroup patterns in *Reed98*

Table 3 Top 4 single-subgroup patterns w.r.t. the SI in *Lastfm* network

Rank	W	I	k_W	$\lvert\varepsilon(W)\rvert$	$p_W \cdot n_W$	#inter-edges
1	idm = 1	0	96	78	8.93	496
2	heavy metal = 1	0	220	165	60.04	1322
3	synthpop = 1	0	208	131	57.32	1307
4	new wave = 1	0	292	191	104.01	1731

For each pattern (each row), we display values for elements that constitute the pattern syntax including W, I, k_W, and also other statistics including its rank, $|\varepsilon(W)|$, $p_w \cdot n_W$ and #inter-edges (each column). k_W is the number of observed edges within $\varepsilon(W)$ (i.e., the set of vertices satisfying the description W), and $p_W \cdot n_W$ is the expected number of edges within $\varepsilon(W)$ w.r.t. the background distribution. I is the indicator equal to 0 if the observed pattern is dense for the analyst (i.e., $k_W > p_W \cdot n_W$) or 1 otherwise (i.e., $k_W < p_W \cdot n_W$). #inter-edges is the number of connections between $\varepsilon(W)$ and $V \setminus \varepsilon(W)$

sufficient. If it is not sufficient, the analyst can increment x, either x_1 or x_2 by 1 iteratively until satisfying results are yielded.

6.3 Comparative evaluation (RQ2)

6.3.1 Experimental setup

A comparison between the SI and other objective interestingness measures can only be made on their performances on single-subgroup pattern discovery (or more precisely, dense subgraph mining), because those existing objective measures are limited to quantify the interestingness of a dense subgraph community.

Choice of datasets and prior beliefs To constrain the search that uses our SI measure to only identify dense subgraphs, we applied individual vertex degrees as the prior beliefs, and chose sparse networks (i.e, *Lastfm* and *DBLPaffs*) for this comparative task. When using the individual vertex degree as priors, single-subgroup patterns' density will not be explainable merely from the individual degrees of the constituent vertices. For real-world networks, given its sparsity (which is common), incorporating this prior leads to a background distribution with a low average connection probability.

Table 4 Top 4 single-subgroup patterns w.r.t. baselines in *Lastfm* network

Measure	W	I	k_W	$\|\varepsilon(W)\|$	#inter-edges
Edge density	1981 songs = 1	0	1	2	21
	africa = 1	0	1	2	76
	40s = 1	0	1	2	22
	early reggae = 1	0	1	2	10
Average degree	post rock = 0 ∧ post-rock = 0	0	12181	1783	498
	post-rock = 0 ∧ dark ambient = 0	0	12092	1770	573
	post-rock = 0 ∧ grindcore = 0	0	12032	1762	634
	post-rock = 0 ∧ technical death metal = 0	0	12106	1773	560
Pool's community score	bionic = 1 ∧ 30 seconds to mars = 0	0	8	6	343
	bionic = 1 ∧ taylor swift = 0	0	8	6	343
or Edge surplus	bionic = 1 ∧ latin = 0	0	8	6	343
	bionic = 1 ∧ spanish = 0	0	8	6	343
Segregation index	gluhie 90e = 0 ∧ lithuanian black metal = 1	0	3	3	1
	goddesses = 0 ∧ pagan black metal = 1	0	3	3	1
	gluhie 90e = 0 ∧ pagan black metal = 1	0	3	3	1
	heartbroke = 0 ∧ lithuanian black metal = 1	0	3	3	1
Modularity of a single community	pop = 1 ∧ new wave = 0	0	2689	475	4913
	pop = 1 ∧ progressive rock = 0	0	2943	514	5083
	pop = 1 ∧ experimental = 0	0	2844	497	5083
	pop = 1 ∧ metal = 0	0	2761	496	5067

For each pattern, we display values for elements that constitute the pattern syntax including W, I, k_W, and also other statistics including $|\varepsilon(W)|$, and #inter-edges. k_W is the number of observed connections within $\varepsilon(W)$ (i.e., the set of vertices satisfying the description W). As all other measures are only for quantifying the interestingness of dense subgraphs, the indicator I is always equal to 0. #inter-edges is the number of connections between $\varepsilon(W)$ and $V \setminus \varepsilon(W)$

In this case, our algorithm identify mostly dense clusters (i.e. $I = 0$), as these are more informative in the sense of strongly contrasting with the expectation which is towards sparsity. *Lastfm*, *DBLPtopics* and *DBLPaffs* are all evidently sparse networks. Among them, *Lastfm* and *DBLPaffs* were chosen as their attributes and the discovered patterns are more readily understood.

Baselines For this comparative evaluation, we consider the following baselines:

– *Edge density.* The number of edges divided by the maximal number of edges.
– *Average degree.* The degree sum for all vertices divided by the number of vertices.
– *Pool's community score* (Pool et al. 2014). The reduction in the number of erroneous links between treating each vertex as a single community and treating all vertices as a whole.
– *Edge surplus* (Tsourakakis et al. 2013). The number of edges exceeding the expected number of edges assuming each edge is present at the same probability α.
– *Segregation index* (Freeman 1978). The difference between the number of expected inter-edges to the number of observed inter-edges, normalized by the expectation.
– *Modularity of a single community* (Newman 2006; Nicosia et al. 2009). The modularity measure of a single community based on transforming the definition of modularity to a local measure.
– *Inverse average-ODF (out-degree fraction)* (Yang and Leskovec 2015). 1 minus the average fraction of vertices' out-degrees to degrees.
– *Inverse conductance.* The number of edges inside the cluster divided by the number of edges leaving the cluster.

More detailed descriptions along with mathematical definitions for these baselines can be found in Table 11 in "Appendix A".

Other settings For single-subgroup pattern discovery on both *Lastfm* and *DBLPaffs* networks, we use beam search with beam width 5 and search depth 2.

6.3.2 Results

Four most interesting patterns w.r.t. the SI and these baseline measures on *Lastfm* are presented in Tables 3 and 4 respectively. For each pattern, we display values for elements that constitute the pattern syntax including W, I, k_W, and also other statistics including its rank, $|\varepsilon(W)|$, and #inter-edges. #inter-edges is the number of connections between $\varepsilon(W)$ and $V \setminus \varepsilon(W)$, telling how isolated a particular group of members is. Particularly for patterns discovered using the SI, we also display $p_W \cdot n_W$, the expected number of connections within $\varepsilon(W)$ w.r.t. the background distribution. Comparing $p_W \cdot n_W$ to k_W gives a direct sense of how much the analyst's expectation differs from the truth (Recall p_W from Eq. 2).

Here, we summarize the main findings.

Using baselines Each of those objective measures exhibits a particular bias that arguably makes the obtained patterns less useful in practice. The edge density is easily maximized to a value of 1 simply by considering very small subgraphs. 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 (Pool et al. 2014). 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. 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.

Using the SI We argue that the patterns extracted using 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 (96 vs. 8.93), and they altogether only have 496 connections to those people not into IDM, which is much sparser than connections within the IDM group (as the connectivity density across the group and that within the group are respectively $496/(78\times1814) \approx 0.0035$ and $96/(78\times(78-1)/2) \approx 0.0320$).

Remark 7 (*Results on DBLPaffs*) For *DBLPaffs*, the same conclusion as above can also be reached. See top 4 single-subgroup patterns on *DBLPaffs* w.r.t. our SI and other measures in Tables 12 and 13 respectively in "Appendix A".

Summary Unlike state-of-the-art objective interestingness measures, each of which exhibits a particular bias, the proposed SI measure achieves a natural balance between coverage and specificity, arguably leading to more insightful patterns.

6.4 The effects of different prior beliefs: a subjective evaluation (RQ3)

6.4.1 Experimental setup

To demonstrate the SI's subjectiveness, we consider different prior beliefs, in search for patterns w.r.t. the SI. We deliberately perform this evaluation on bi-subgroup pattern discovery for a more generic and interesting setting.

Choice of datasets In the following, we analyze results on *Caltech36* and *Reed98*. These two networks are chosen, because their straightforward domain knowledge provides us the ease for prior belief settings. People, even those that are not social scientists, normally hold prior beliefs about this sort of friendship network (e.g., they commonly believe that students of different class years are less likely to know each other than students from the same class year).

Other settings For bi-subgroup pattern discovery, we applied the nested beam search with $x_1 = 2, x_2 = 3$, 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 obtained patterns are more explainable, and the results are easier to evaluate.

6.4.2 Results

The 4 most subjectively interesting patterns under each prior belief are presented in Table 6 (for *Caltech36*) and Table 7 (for *Reed98*), with their associated notations are summarized in Table 5.

Incorporating Prior 1 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 reveal people graduating in different years rarely know each other (rows for Prior 1 in Table 6), 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 7).

Incorporating Prior 1 and Prior 2 We then incorporated prior beliefs on the densities between bins for different graduation years (i.e., Prior 2). All the extracted top 4 patterns on *Caltech 36* indicate rare connections between people living in different dormitories, and this is also not surprising (rows for Prior 1 + Prior 2 in Table 6).

For *Reed98*, incorporating Prior 1 and Prior 2 provides interesting patterns (rows for Prior 1 + Prior 2 in Table 7). The top one indicates people living in dormitory 88 are friends with many in dormitory 89. In contrast, what people commonly believe is that people living in different dormitories are less likely to know each other. For an analyst who has such preconceived notion, this pattern is interesting. Both the fourth and the seventh patterns reveal a certain person knew more people in class of 2009 than expected.

Table 5 Notations in Tables 6, 7 and 8

Notation	Description
W_1/W_2	The description of the first/second subgroup
$\lvert\varepsilon(W_1)\rvert/\lvert\varepsilon(W_2)\rvert$	The subgroup of vertices satisfying the description W_1/W_2
k_W	The number of observed edges between $\varepsilon(W_1)$ and $\varepsilon(W_2)$
$p_W \cdot n_W$	The expected number of edges between $\varepsilon(W_1)$ and $\varepsilon(W_2)$ w.r.t. the background distribution
I	The indicator equal to 0 if the observed pattern is dense for the analyst (i.e., $k_W > p_W \cdot n_W$) or 1 otherwise (i.e., $k_W < p_W \cdot n_W$)

Table 6 Varying prior beliefs in *Caltech36* network

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

Prior 1 represents the prior belief on the individual vertex degree. Both Prior 2 and Prior 3 regard particular attribute knowledge. More specifically, Prior 2 expresses the data analyst's knowledge on the edge densities between bins for different graduation years, and Prior 3 expresses that for different dormitories. For each pattern (each row), we display values for elements that constitute the pattern syntax including W_1, W_2, I, k_W, and also other statistics including its rank, $\lvert\varepsilon(W_1)\rvert$, $\lvert\varepsilon(W_1)\rvert$, and $p_w \cdot n_W$ (each column)

Table 7 Varying prior beliefs in *Reed98* network

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

Prior 1 represents the prior belief on the individual vertex degree. Prior 2 is on the edge densities between bins for different graduation years. For each pattern (each row), we display values for elements that constitute the pattern syntax including W_1, W_2, I, k_W, and also other statistics including its rank, $|\varepsilon(W_1)|$, $|\varepsilon(W_1)|$, and $p_w \cdot n_W$ (each column)

Incorporating Prior 1, Prior 2 and Prior 3 For *Caltech 36*, by additionally incorporating prior beliefs on the dependency of the connectivity probability on the difference in dormitories (i.e., Prior 3), patterns characterizing some interesting dense connections are discovered (rows for Prior 1 + Prior 2 + Prior 3 in Table 7). For instance, the top pattern indicates 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. Furthermore, the starting year for those 2008 cohort is exactly when these three people should have graduated. Therefore, these two groups had opportunities to become friends. The fourth pattern indicates an alumnus who had studied in a high school knew almost all the students living in a certain dormitory. The reason behind this pattern might be worth investigating, which could be for instance, this alumni worked in this dormitory.

Summary As the results show, incorporating different prior beliefs leads to discovering different patterns that strongly contrast with these beliefs. The proposed SI measure thus succeeds in quantifying the interestingness in a subjective manner.

6.5 Evaluation on iterative pattern mining (RQ4)

6.5.1 Experimental setup

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. We show this on searching for bi-subgroup patterns because they are more generic.

Choice of datasets Dataset *DBLPaffs* and *Lastfm* are used, as the meanings of their attributes are clear and straightforward, giving an ease to explain the discovered patterns.

Other settings Other settings for this task are the same as for addressing RQ2. The nested beam search with $x_1 = 2, x_2 = 3$, and $D = 2$ was applied. The target descriptions W_1 and W_2 are constrained to include at least one common attribute but with various values, making the corresponding pair of subgroups $\varepsilon(W_1)$ and $\varepsilon(W_2)$ not overlap with each other.

6.5.2 Results

Results for *Lastfm* are displayed and discussed in "Appendix B". Here we only analyze the results on *DBLPaffs*. Table 8 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 agrees with fact that many of the world's largest high-tech corporations and reputable universities are

Table 8 Top 3 discovered bi-subgroup patterns of each iteration in *DBLPaffs* network

	Rank	W_1	W_2	I	k_W	$\lvert\varepsilon(W_1)\rvert$	$\lvert\varepsilon(W_2)\rvert$	$p_W \cdot n_W$
Iteration 1	1	USA = 1	USA = 0	1	335	3132	3340	765.83
	2	USA = 1 ∧ China = 0	USA = 0	1	288	2969	3340	725.97
	3	USA = 1 ∧ Australia = 0	USA = 0	1	320	3092	3340	756.05
Iteration 2	1	NJ (New Jersey) = 0	NJ = 1 ∧ CA (California) = 1	0	93	6262	15	6.91
	2	CA = 0	NJ = 1 ∧ CA = 1	0	86	5584	15	6.13
	3	NJ = 1 ∧ Israel = 0	NJ = 1 ∧ CA = 1	0	93	6153	15	6.76
Iteration 3	1	China = 0	China = 1	1	144	5599	873	271.02
	2	China = 0	China = 1 ∧ IL (Illinois) = 0	1	128	5599	861	266.10
	3	China = 0 ∧ USA = 0	China = 1	1	64	2630	873	168.09
Iteration 4	1	CA = 1	CA = 0 ∧ WA = 1	0	55	888	184	11.73
	2	WA = 0	WA = 1	0	182	6254	218	97.78
	3	CA = 1 ∧ TX (Texas) = 0	CA = 0 ∧ WA = 1	0	55	876	184	11.57

For each pattern (each row), we display values for elements that constitute the pattern syntax including W_1, W_2, I, k_W, and also other statistics including its rank, $\lvert\varepsilon(W_1)\rvert$, $\lvert\varepsilon(W_1)\rvert$, and $p_w \cdot n_W$ (each column). See Table 5 for descriptions of these statistics located in these regions. 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 Table 12 in "Appendix A"). 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 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 obtained patterns into the background distribution for subsequent iterations, our method can identify patterns which strongly contrast with this knowledge. This results in a set of patterns that are not redundant and highly surprising to the data analyst. Note that the lack of redundancy arises naturally, without the need for explicitly constraining the overlap between the patterns in consecutive iterations. In fact, some amount of overlap may still occur, as long as the non-redundant part of the information is sufficiently large.

6.6 Empirical results on the discovered global patterns (RQ5)

To demonstrate the use of our method for mining interesting global patterns, we illustrate and analyze the experimental results on *DBLPaffs* (in Sect. 6.6.1), *DBLPtopics* (in Sect. 6.6.2) and *MP* (in "Appendix C"). Each of these datasets serves an interesting case study for us to evaluate our method on.

6.6.1 Case study on *DBLPaffs*

Task Paper citations relate to authors' affiliations to some extent. For example, institutions in some particular countries or regions are reputable, and often produce highly-cited research. Also, collaborations and mutual citations may frequently occur in institutions from some certain countries or regions. Thus, of particular interest could be patterns that describe a subgroup of papers from affiliations A frequently (or rarely) cite papers in another subgroup from affiliations B. We show such patterns can be revealed by a summarization yielded by our approach.

The resulting summarization By running our algorithm for 6 iterations, this citation network is summarized into 7 subgroups each consisting of papers satisfying a particular description about their authors' affiliations. These 7 subgroups are respectively defined by

1. USA = 1 and WA (Washington) = 1;
2. USA = 1 and WA = 0 and China = 1;
3. USA = 1 and WA = 0 and China = 0 and CA (California) = 1 and NJ (New Jersey) = 1;
4. USA = 1 and WA = 0 and China = 0 and CA = 1 and NJ = 0;
5. USA = 1 and WA = 0 and China = 0 and CA = 0;

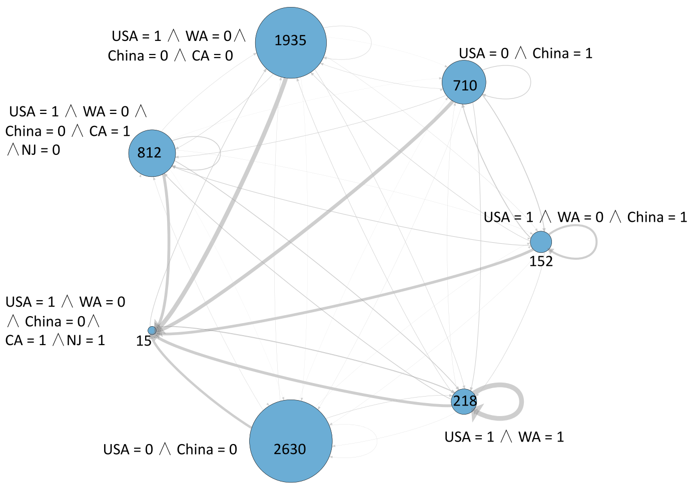


Fig. 8 The resulting summary of *DBLPaffs*. Each supervertex (representing a paper subgroup) is labelled by its number of members (in the centre of the blue circle) and its description (near the blue circle). Each directed edge connects one supervertex to the other, and its linewidth indicates the connectivty density from a subgroup (e.g. $\varepsilon(W_1)$) to the other one (e.g., $\varepsilon(W_2)$). A thicker edge means the citations from $\varepsilon(W_1)$ to $\varepsilon(W_2)$ are more frequent) (Color figure online)

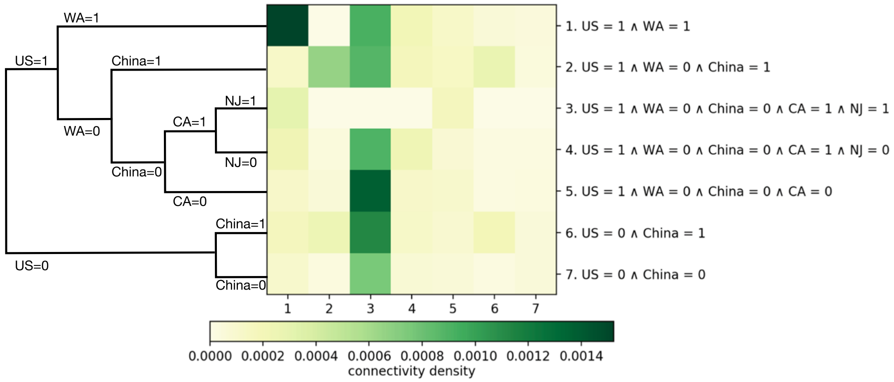


Fig. 9 The heatmap representation of the density matrix for *DBLPaffs*, aligned with a dendrogram illustration of the splitting hierarchy on the left. A deeper color of each square indicates a higher connectivity density from a subgroup (represented by row) to another one (represented by column) (Color figure online)

6. USA = 0 and China = 1;
7. USA = 0 and China = 0.

The summary is displayed in Fig. 8. In the following, we discuss properties of local subgroup patterns revealed in our summarization to access its validity.

Remark 8 (*Redundancy in the descriptions*) One may notice that some subgroup descriptions can be more concise. For example, the first subgroup pattern "USA = 1 and WA = 1" should induce the same extension as only"WA = 1". There is no mechanism in our approach for the global pattern mining that would prefer the alternative shorter description of the same subgroup. Yet, such redundancy can be easily identified and adjusted in post-processing. Moreover, this issue does not affect our single/bi-subgroup pattern mining approach where each iteration of the search essentially identifies an optimal pattern rather than a split (in global pattern mining approach), and shorter description of the same subgroup would have a larger SI value given by its smaller DL value.

Discussion A series of interesting local subgroup patterns emerge from the resulting summarization. The density matrix where its entry at the i-th row and the j-th column is the citation density from papers in the i-th subgroup to the j-th is visualized by a heatmap, of which the left side is lined up with a dendrogram illustrating the splitting hierachy (see Fig. 9).

Obviously, the most cohesive subgroup are papers from institutions in Washington state in USA, as they cite those within this subgroup most frequently (indicated by the darkest green square in the top left). Closer inspection revealed that the majority of these papers are written by authors from Microsoft Corporation and the University of Washington.

The most highly-cited subgroup is the third one (indicated by the dark color of all the squares along the third column except the one in the third row). This subgroup only contains 15 papers, and their authors are affiliated to institutes in California and New Jersey, neither in Washington nor China. Note this also agrees with bi-subgroup patterns found in previous experiment for addressing RQ3 (Iteration 2 in Sect. 6.5). As already been pointed out, 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.

Another interesting subgroup is the second one of which authors are with affiliations in China and USA (except Washington). Researchers related to this subgroup are surprisingly isolated, as their papers are seldom cited by those from other subgroups but very frequently (or to be more precise, the second most frequently) within this subgroup (indicated by the shallow color of all the squares along the second column except the one in the second row). In fact, Chinese affiliated with research organisations in China and Chinese affiliated with organisations in USA, have coauthored most papers in this subgroup. The reason of their isolation might be interesting for data analysts to investigate. Again, this coincides with what we found in experiment for addressing RQ3 (Iteration 3 in Sect. 6.5). The difference is the identified subgroup here is more specified (i.e., also being with affiliation in USA except Washington).

A follow-up experiment The rest subgroup defined by USA = 0 and China = 0 (i.e., the 7th one) contains a considerable number of members (indicated by the largest circle in Fig. 8). Continuing to run our algorithm for subsequent iterations tends to split this subgroup up such that some cohesive groups affiliated with organisations in other countries are revealed. For example, subgroups related to affiliations in Singapore, Canada, the Netherlands emerge respectively in the first 3 subsequent iterations (see the corresponding splitting hierarchy highlighted by red dashed lines in Fig. 10). They all cite papers within the same subgroup or those from the third subgroup (i.e., the overall most highly-cited one) very frequently (see rows 7, 8, 9 of the heatmap in Fig. 10).

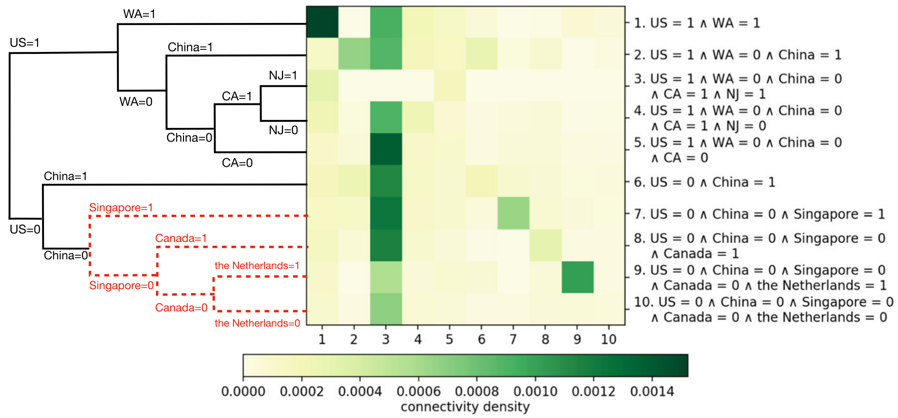


Fig. 10 The heatmap representation of the density matrix among subgroups obtained by running our algorithm for another 3 subsequent iterations on *DBLPaffs*, with a dendrogram illustration of the splitting hierarchy on the left. A deeper color of each square indicates a higher connectivity density from a subgroup (represented by row) to another one (represented by column). The splitting hierarchy for the 3 new iterations are in red dashed lines (Color figure online)

6.6.2 Case study on *DBLPtopics*

Task A data analyst working for an academic organization may want to obtain a high-level view of citation vitality among different research fields. Given *DBLPtopics* dataset, we here show the global pattern identified by our summarization approach can provide such high-level view, revealing interesting local subgroup patterns of the form 'papers of study field A frequently (or rarely) cite those of field B'. We also show the obtained global pattern can provide the data analyst further insights by linking with information about paper distribution among different conferences.

The resulting summarization The summarization of *DBLPtopics* is generated by running our algorithm for 4 iterations, and the resulting summarization rule means to divide all papers into the following 5 subgroups:

1. $a_1 < Q_2^{a_1} \wedge a_8 \geq Q_1^{a_8}$ (Theoretical machine learning);
2. $a_1 < Q_2^{a_1} \wedge a_8 < Q_1^{a_8}$ (Practical machine learning);
3. $a_1 \geq Q_2^{a_1} \wedge a_5 < Q_3^{a_5} \wedge a_3 < Q_3^{a_3}$ (Data mining);
4. $a_1 \geq Q_2^{a_1} \wedge a_5 < Q_3^{a_5} \wedge a_3 \geq Q_3^{a_3}$ (Information retrieval);
5. $a_1 \geq Q_2^{a_1} \wedge a_5 \geq Q_3^{a_5}$ (Database).

For each subgroup, we list its original description and a corresponding short interpretation (in brackets) based on summarizing attributes' meaning. As mentioned previously (in Sect. 6.1), an attribute is essentially one of the first 50 LSI components for the original paper-topic matrix. Its meaning can thus be described by its 5 subcomponents with highest absolute weights (shown in Table 9). A higher weight means this attribute's meaning is closer (positive sign) or more contrasting (negative sign) to this research field. We will use these short interpretations rather than original descriptions in the following part, because these are more straightforward. Generally, this summarization not only successfully captures those 4 research areas that publications in *DBLPtopics*

Table 9 The meaning of attributes related to the resulting summarization

Attribute	Meaning (Top 5 most strongly associated fields of study by absolute weight)
a_1	Data mining (0.55)
	Machine Learning (– 0.49)
	Database (0.32)
	Computer Science (0.28)
	Information retrieval (0.25)
a_3	Data mining (0.41)
	Computer science (– 0.40)
	Mathematics (0.39)
	Information retrieval (0.30)
	Pattern recognition (0.24)
a_5	Database (0.61)
	Information retrieval (– 0.49)
	Query optimization (0.21)
	World Wide Web (– 0.18)
	Mathematics (0.15)
a_8	Mathematical optimization (0.45)
	Information retrieval (0.44)
	Database (0.37)
	Data mining (– 0.25)
	Computer science (0.22)

are intended to cover (i.e., Machine Learning, Database, Information Retrieval, and Data Mining), but also identifies a deeper-level structure (i.e., the partition of machine leaning papers into two subgroups according to different aspects they emphasize: more practical or more theoretical).

The summary of *DBLPtopics* based on the resulting summarization rule is displayed in Fig. 11. To highlight the citation vitality between each pair of subgroups, the corresponding citation density matrix is visualized by a heatmap, lined up with a dendrogram on the left illustrating the splitting hierarchy (see Fig. 12).

Discussion As shown in Fig. 12, the citation density within the same subgroup is often high, indicating papers of similar research field often cite each other.

Exceptions are the second (practical machine learning) subgroup and the third one (data mining) which respectively cite the fifth (database) and the fourth (information retrieval) most frequently. This accords with the fact that solving data mining or practical machine learning research questions often necessitates database techniques or information retrieval to solve some subtasks.

Clearly, the fourth and the fifth subgroup are most cohesive (indicated by those two evidently dark green squares in the fourth and the fifth place of the diagonal). Also, these two groups cite each other and the data mining subgroup very frequently.

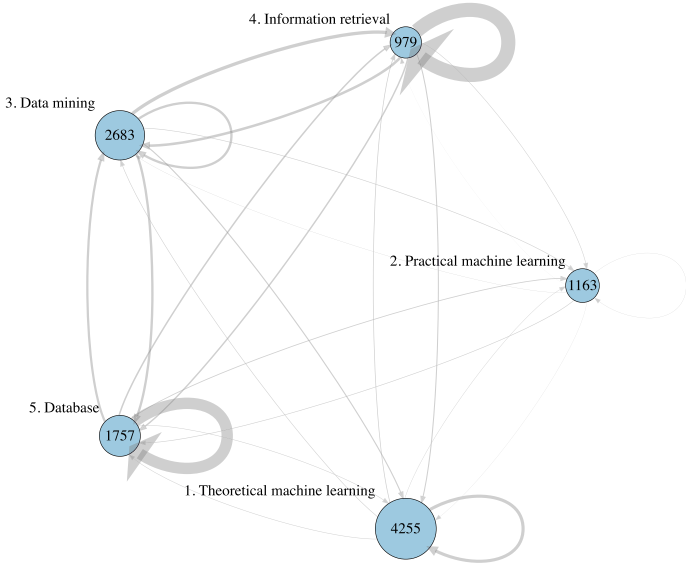


Fig. 11 The resulting summary of *DBLPtopics*. Each supervertex (representing a paper subgroup) is labelled by its number of members (in the centre of the blue circle) and its description (near the blue circle). Each directed edge connects one supervertex to the other, and its linewidth indicates the connectivty density from a subgroup (e.g. $\varepsilon(W_1)$) to the other one (e.g., $\varepsilon(W_2)$). A thicker edge means the citations from $\varepsilon(W_1)$ to $\varepsilon(W_2)$ are more frequent) (Color figure online)

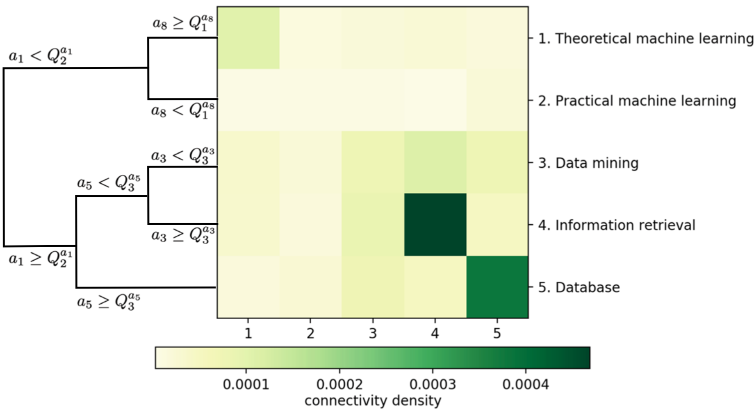


Fig. 12 The heatmap representation of the density matrix for *DBLPtopics*, aligned with a dendrogram illustration of the splitting hierarchy on the left (Recall Q_i^a denotes the i-th quartile of values for the attribute a). A deeper color of each square indicates a higher connectivity density from a subgroup (represented by row) to another one (represented by column) (Color figure online)

One downstream task: knowing more about conferences The summarization generated by our approach can be useful in some downstream analysis tasks. Here we show an example of utilizing it to know more about conferences, simply by linking with

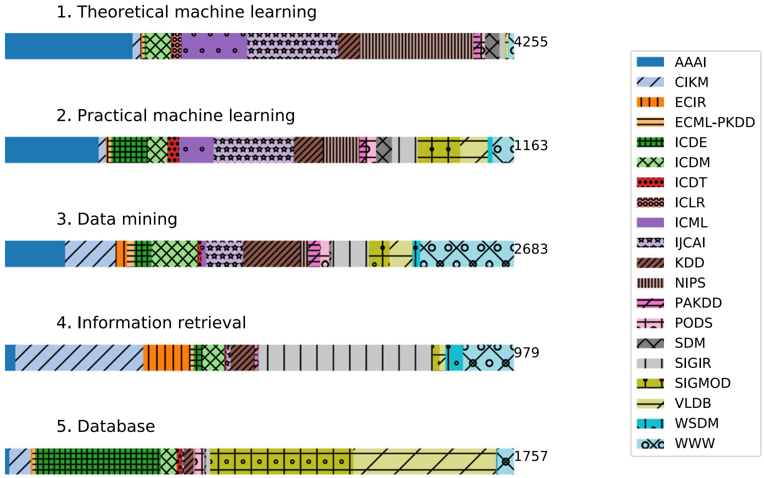


Fig. 13 The distribution publications in 20 selected conferences within each subgroup. For each bin representing a subgroup, the subgroup description is placed on the top, and the number of papers in this subgroup is placed on the right end. The length of a rectangular in a certain color and hatch inside a bin is proportional to the percentage of publications in a certain conference in a subgroup. Conferences are in alphabetical order (Color figure online)

the distribution of publications in those 20 selected conferences within each subgroup (displayed in Fig. 13).

First, by merely looking at the distribution for each subgroup, the data analysts can learn the relationship between research fields and conferences, e.g., answering questions like which research field is dominated by which conference. As can be seen, a noticeable large proportion of publications in regard to the information retrieval (the fourth subgroup) are in SIGIR and CIKM. and the database publications (the fifth subgroup) are mostly in ICDE, VLDB, SIGMOD. The data mining subgroup (the third one) is special in a sense that their publications are distributed quite evenly. WWW only holds a slim majority, and publications from KDD, AAAI, ICDM, CIKM are a little bit more than those from another venue (except WWW). Moreover, it is interesting to notice KDD and ICDM appear to be more interdisciplinary, accepting papers surprisingly evenly from these research areas compared to other conferences (as there is no noticeably longer dark brown or light green rectangular in either one of these 5 horizontal bins in Fig. 13).

Also, the data analyst can combine Figs. 12 and 13 to deduce the citation vitality among different conferences. For example, publications in SIGIR and CIKM often cite those also in these two conferences (as the fourth subgroup is very cohesive), and they also often cite publications in WWW, AAAI, KDD,CIKM (those dominating the third subgroup).

Summary As shown by these case studies on different datasets, global patterns identified by our method can not only directly provide insights by revealing a series of interesting single-subgroup and bi-subgroup patterns, but also be utilized to facilitate some downstream analysis tasks.

6.7 Scalability evaluation (RQ6)

6.7.1 Experimental setup

Choice of datasets We used *Lastfm* to investigate the scalability to the number of selectors, because it can give a largest number of selectors (i.e., 21695) as the search space.

Other settings Same as for other experiments, in the scalability evaluation, we applied the beam search with $x = 5$ (for single-subgroup pattern discovery), the nested beam search with $x_1 = 2$, $x_2 = 3$, and $D = 2$ (for bi-subgroup pattern discovery), 8 processors running in parallel (for global pattern mining).

6.7.2 Results

Effect of $|S|$. Figure 14 displays run time on *Lastfm* w.r.t. the number of selectors in the search space (i.e., $|S|$). It is clear that, in either single-subgroup or global pattern mining, the run time experiences a linear growth as we gradually double the $|S|$ (from 10 to 20,480), whereas the run time for bi-subgroup pattern mining increases more than linearly, and exceeds 1 day when $|S|$ is larger than 2560.

Run time The run time of our experiments for addressing RQ2 to RQ5, as well as the $|S|$ and $|V|$ statistics are listed in Table 10. The influence of the $|S|$ and $|V|$ on the run time is evident.

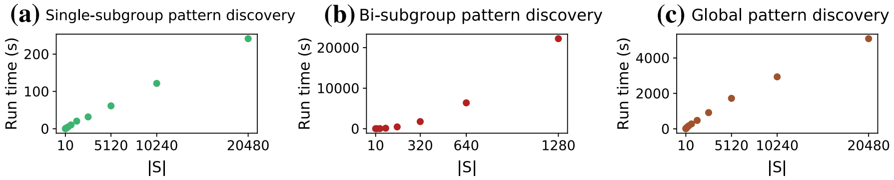


Fig. 14 Run time (s) parametrized by $|S|$ on *Lastfm*

Table 10 Run time

	Dataset	$\|S\|$	$\|V\|$	Run time (s)
Single-subgroup pattern mining (RQ2)	*Lastfm*	21, 695	1892	278.49
	DBLPaffs	232	6472	32.40
Bi-subgroup pattern mining (RQ3 and RQ4)	*Caltech36*	602	762	1312.57
	Reed98	748	962	1965.41
	Lastfm	200	1892	679.85
	DBLPaffs	232	6472	3114.78
Global pattern mining (RQ5)	*DBLPaffs*	232	6472	830.69
	DBLPtopics	150	10, 837	1570.90
	MPvotes	78	650	12.73

Summary The run time grows linearly in the number of attributes in both single-subgroup and global pattern mining, whereas it grows faster than linearly in bi-subgroup pattern mining.

7 Conclusion

Prior work of pattern mining in attributed graphs typically only search for dense subgraphs ('communities') with homogenous attributes. We generalized this type of pattern to densities within this subgraph (no matter whether dense or sparse, which we refer as *single-subgroup pattern*), between a pair of different subgroups (which we refer as *bi-subgroup pattern*), as well as between all pairs from a set of subgroups that partition the whole vertex set (which we refer as *global pattern*).

We developed a novel information-theoretic approach for quantifying interestingness of such patterns in a subjective manner, with respect to a flexible type of prior knowledge the analyst may have about the graph, including insights gained from previous patterns.

The empirical results show that our method can efficiently find interesting patterns of these new different types. In the standard problem of dense subgraph mining, our method can yield results that are superior to the state-of-the-art. We also demonstrated empirically that our method succeeds in taking in account prior knowledge in a meaningful way.

The proposed SI interestingness measure has considerable advantages, but a price to pay for this is in terms of computational time. To help maintain the tractability, we succumb to some accurate heuristic search strategies. It would be useful for the future work to discover a search strategy with performance guarantee and to further speed up the search (e.g., by branch and bounds).

Acknowledgements The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement No. 615517, from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme, from the FWO (Project Nos. G091017N, G0F9816N, 3G042220), and from the European Union's Horizon 2020 research and innovation programme and the FWO under the Marie Sklodowska-Curie Grant Agreement No. 665501.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix

A For Section 6.3: A comparative evaluation on *DBLPaffs* network (RQ2)

Some objective interestingness measures we used for comparison, as well as their explanations are listed in Table 11.

We consider undirected graphs for the sake of presentation and consistency with most literature. However, we note that the generalization to directed graphs is straightforward.

B For Section 6.5: Evaluation on the iterative pattern mining on *Lastfm* dataset (RQ4)

Table 14 displays the top 3 patterns found in each of the five iterations on the *Lastfm*. The description search space is built based on only 100 most frequently used tags, that means, $|S| = 100 \times 2$.

Iteration 1 Initially, we incorporate prior belief on individual vertex degree. The extracted 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. 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.

Table 11 Existing measures for a comparison

Measure	Description	Mathematical definition
Edge density	The ratio of the number of edges to the number of possible edges in the cluster	$\frac{2*k_W}{\lvert\varepsilon(W)\rvert*(\lvert\varepsilon(W)\rvert-1)}$
Average degree	The ratio of the degree sum for all vertices to the number of vertices in the cluster	$\frac{2*k_W}{\lvert\varepsilon(W)\rvert}$
Pool's measure (Pool et al. 2014)	The reduction in the number of erroneous links between treating each vertex as a single community and treating all the vertices as a whole	$\sum_{u\in\varepsilon(W)} d(u) - \left(\frac{\lvert\varepsilon(W)\rvert*(\lvert\varepsilon(W)\rvert-1)}{2} - k_W\right) - \#\text{inter-edges} = -\frac{\lvert\varepsilon(W)\rvert*(\lvert\varepsilon(W)\rvert-1)}{2} + 3*k_W$
Edge Surplus (Tsourakakis et al. 2013)	The number of edges exceeding the expected number of edges within the cluster assuming each edge is present with the same probability α	$k_W - \alpha * \lvert\varepsilon(W)\rvert * (\lvert\varepsilon(W)\rvert - 1)$
Segregation index (Freeman 1978)	The difference between the number of expected inter-edges to the number of the observed inter-edges, normalized by the expectation	$1 - \frac{\#\text{inter}-\text{edges}*\lvert V\rvert*(\lvert V\rvert-1)}{2*\lvert E\rvert*\lvert\varepsilon(W)\rvert*(\lvert V\rvert-\lvert\varepsilon(W)\rvert)}$
Modularity of a single community (Newman 2006; Nicosia et al. 2009)	The measure quantifying the modularity contribution of a single community based on transforming the definition of modularity to a local measure	$\frac{1}{2*\lvert E\rvert}\sum_{u,v\in\varepsilon(W)}\left(a_{u,v} - \frac{d(u)*d(v)}{2*\lvert E\rvert}\right)$
Inverse Average-ODF (out-degree fraction) (Yang and Leskovec 2015)	The inverse of the Average-ODF which is based on averaging the fraction of inter-degree and the degree for each vertex in the cluster	$1 - \frac{1}{\lvert\varepsilon(W)\rvert}\sum_{u\in\varepsilon(W)}\frac{\overline{d}_W(u)}{d(u)}$
Inverse Conductance	The ratio of the number of edges inside the cluster to the number of edges leaving the cluster	$\frac{k_W}{\#\text{inter-edges}}$

For a given attributed graph $G = \{V, E, \Lambda\}$, and a community induced by a description W such that $\varepsilon(W) \in V$, $d(u)$ denotes the degree of vertex $u \in V$; $\overline{d}_W(u)$ denotes the inter-degree of vertex $u \in \varepsilon(W)$, specfically, $\overline{d}_W(u) := \lvert\{(u, v) \in E : v \in V \setminus \varepsilon(W)\}\rvert$; and #inter-edges denotes the number of connections between $\varepsilon(W)$ and $V \setminus \varepsilon(W)$

Summary By incorporating the newly obtained 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 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.

Table 12 Top 4 single-subgroup patterns w.r.t. our SI in *DBLPaffs* network

Rank	W	I	k_W	$\lvert\varepsilon(W)\rvert$	$p_W \cdot n_W$	#inter-edges
1	China = 1	0	179	873	63.20	566
2	China = 1 $\wedge$ IN (Indiana) = 0	0	179	869	62.58	561
3	China = 1 $\wedge$ Italy = 0	0	179	870	62.67	561
4	China = 1 $\wedge$ Denmark = 0	0	179	870	62.69	562

For each pattern (each row), we display values for elements that constitute the pattern syntax including W, I, k_W and also other statistics including its rank, $|\varepsilon(W)|$, $p_w \cdot n_W$ and #inter-edges (each column). k_W is the number of observed edges within $\varepsilon(W)$ (i.e., the set of vertices satisfying the description W), and $p_W \cdot n_W$ is the expected number of edges within $\varepsilon(W)$ w.r.t. the background distribution. I is the indicator equal to 0 if the observed pattern is dense for the analyst (i.e., $k_W > p_W \cdot n_W$) or 1 otherwise (i.e., $k_W < p_W \cdot n_W$). #inter-edges is the number of connections between $\varepsilon(W)$ and $V \setminus \varepsilon(W)$

C For Section 6.6: One more case study on *MPvotes* for the evaluation of global pattern mining

Task Brexit is a hot topic of debate in UK. MPs' voting behaviours on Brexit might affect the likelihood of their connections. Using this information to summarize MPs friendship network is thus potential to provide insights on the Brexit saga. We here investigate whether our approach can achieve this.

The resulting summarization The summarization of *MPvotes* generated from running our algorithm for 4 iterations splits all MPs into 5 subgroups, and they are respectively defined by

1. I1 $= -1$ or 0 $\wedge$ I10 V3 $= -1$ or 0 $\wedge$ I10 V4 $= -1$ or 0;
2. I1 $= -1$ or 0 $\wedge$ I10 V3 $= -1$ or 0 $\wedge$ I10 V4 = 1;
3. I1 $= -1$ or 0 $\wedge$ I10 V3 = 1;
4. I1 = 1 $\wedge$ I7 V4 = 1 or 0;
5. I1 = 1 $\wedge$ I7 V4 $= -1$.

where 'Ii Vj' represents the j-th vote in the i-th issue. For an issue around which there exists only one vote, say the 1st issue, it is simply represented as I1. Detailed interpretation of all voting issues related to our summarization are displayed in Table 15. The summary of *MPvotes* is illustrated in Fig. 15. For a dedicated view of the connectivity density between each subgroup pair, the corresponding density matrix is visualized by a heatmap, aligned with an dendrogram illustration of the splitting hierachy on the left (see Fig. 16).

Discussion Clearly in Fig. 16, our summarization identifies several crucial votings that partition MPs into cohesive subgroups. That is, MPs taking the same sides in these votings connect more frequently to each other (i.e., those within the same subgroup) than MPs voting differently (i.e., those in other subgroups). The only exception is the 2nd subgroup who connect most frequently to the 3rd subgroup. More interpretations of these patterns are provided in the following.

Combining with political parties The data analyst can utilize our summarization of *MPvotes* to obtain insights about Brexit saga. Here, we provide one example. More

Table 13 Top 4 single-subgroup patterns w.r.t. other measures in *DBLPaffs* network

Measure	W	I	k_W	$\lvert\varepsilon(W)\rvert$	#inter-edges
Edge density	DE (Delaware) = 1 ∧ MD(Maryland) = 1	0	1	2	2
	DC (District of Columbia) = 1 ∧ TX (Texas) = 1	0	1	2	6
	Netherlands = 1 ∧ MA(Massachusetts) = 1	0	1	2	3
	Netherlands = 1 ∧ WA = 1	0	1	2	5
Average degree	UK = 0 ∧ Japan = 0	0	2882	6038	161
	UK = 0 ∧ Ireland = 0	0	2975	6234	79
	Japan = 0 ∧ Ireland = 0	0	2952	6191	106
	Sweden = 0 ∧ Ireland = 0	0	3044	6391	22
Pool's community score	DE = 1 ∧ MD = 1	0	1	2	2
	DC = 1 ∧ TX = 1	0	1	2	6
or Edge surplus	Netherlands = 1 ∧ MA = 1	0	1	2	3
	Netherlands = 1 ∧ WA = 1	0	1	2	5
Segregation index	AL (Alabama) = 0	0	3066	6470	0
	AL = 1	0	0	2	0
	Bulgaria = 0	0	3066	6471	0
	AS (American Samoa) = 0	0	3066	6471	0
Modularity of a single community	China = 0 ∧ United States = 1	0	1173	2969	1203
	NY(New York) = 0 ∧ United States = 1	0	1067	2757	1224
	Singapore = 0 ∧ United States = 1	0	1247	3088	1194
	Germany = 0 ∧ United States = 1	0	1262	3077	1191

For each pattern (each row), we display values for elements that constitute the pattern syntax including W, I, k_W and also other statistics including $|\varepsilon(W)|$ and #inter-edges (each column). k_W is the number of observed edges within $\varepsilon(W)$ (i.e., the set of vertices satisfying the description W), and $p_W \cdot n_W$ is the expected number of edges within $\varepsilon(W)$ w.r.t. the background distribution. $I = 0$ in all cases as other measures can only quantify the interestingness of dense subgraphs. #inter-edges is the number of connections between $\varepsilon(W)$ and $V \setminus \varepsilon(W)$

Table 14 Top 3 discovered bi-subgroup patterns of each iteration in *Lastfm* network

	Rank	W_1	W_2	I	k_W	$\lvert\varepsilon(W_1)\rvert$	$\lvert\varepsilon(W_2)\rvert$	$p_W \cdot n_W$
Iteration 1	1	heavy mental = 1	heavy mental = 0 ∧ pop = 1	1	349	165	529	769.18
	2	pop = 1 ∧experimental = 0	rnb = 0 ∧experimental = 1	1	360	497	230	812.78
	3	pop = 1 ∧experimental = 0	experimental = 1	1	495	497	247	943.96
Iteration 2	1	pop = 1 ∧ indie = 0	pop = 0 ∧experimental = 1	1	103	366	159	369.44
	2	pop = 1 ∧ alternative = 0	pop = 0 ∧ experimental = 1	1	84	325	159	334.77
	3	pop = 1 ∧ experimental = 0	rnb = 0 ∧ experimental = 1	1	360	497	230	750.77
Iteration 3	1	synth pop = 1	synth pop = 0 ∧ new wave = 1	0	163	54	150	43.10
	2	synth pop = 1 ∧ british = 1	new wave = 1 ∧british = 0	0	116	26	113	20.71
	3	synth pop = 1	synth pop = 0 ∧ synthpop = 1	0	125	54	102	29.64
Iteration 4	1	new wave = 1 ∧ hip-hop = 0	new wave = 0 ∧ pop = 1	1	160	475	343	670.74
	2	new wave = 1 ∧ rnb = 0	new wave = 0 ∧ pop = 1	1	379	170	475	705.43
	3	new wave = 1 ∧ soul = 0	new wave = 0 ∧ pop = 1	1	323	150	475	624.41
Iteration 5	1	instrumental = 1	instrumental = 0 ∧ ambient = 1	0	273	195	144	114.62
	2	electronic = 1	electronic = 0 ∧ ambient = 1	0	268	167	160	113.66
	3	progressive metal = 1	progressive metal = 0 ∧ heavy metal = 1	0	128	99	111	34.81

For each pattern (each row), we display values for elements that constitute the pattern syntax including W_1, W_2, I, k_W, and also other statistics including its rank, $|\varepsilon(W_1)|$, $|\varepsilon(W_1)|$, and $p_w \cdot n_W$ (each column). k_W is the number of observed connections between $\varepsilon(W_1)$ (i.e., vertices satisfying the description W_1) and $\varepsilon(W_2)$ (i.e., vertices satisfying the description W_2), and $p_W \cdot n_W$ is the expected number of connections between $\varepsilon(W_1)$ to $\varepsilon(W_2)$ w.r.t. the background distribution. I is the indicator equal to 0 if the observed pattern is dense for the analyst (i.e., $k_W > p_W \cdot n_W$) or 1 otherwise (i.e., $k_W < p_W \cdot n_W$)

Table 15 The description of voting issues related to the resulting summarization in the order of spliting

Vote Notation	Description
I1	Government in rejecting an amendment that would have given MPs the power to stop the UK from leaving the EU without a deal
I10 V3	Labour's plan for a close economic relationship with the EU
I10 V4	UK membership of the European Free Trade Association (Efta) and European Economic Area (EEA)
I7 V4	Government in contempt of parliament

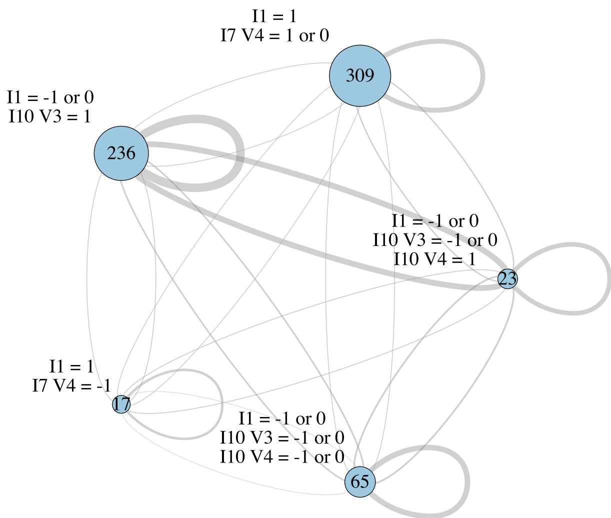


Fig. 15 The resulting summary of *MPvotes*. Each supervertex (representing a subgroup of MPs) is labelled by its number of members (in the centre of the blue circle) and its description (near the blue circle). Each undirected edge connects between one supervertex and the other, with its linewidth indicating the connectivity density between these two corresponding subgroups (The thicker the edge, the higher the connectivity density) (Color figure online)

specifically, we show, by combining with the distribution of MPs' party affiliations within each subgroup (illustrated in Fig. 17), our summarization can:

(a) reveal crucial voting issues over which MPs from different parties take different sides;
(b) provide a high-level view of connectivity densities among different political parties.

Now we trace the partition process based on our summarization in order to show (a). The first split is a vote on I1 of which 'ayes' side with the government to keep no-deal Brexit on the table as a possibility (see the dendrogram in Fig. 17). A clear opinion conflict between different parties can be observed. More specifically, all the MPs from Scottish National Party (SNP), Liberal Democrat (LD), Sinn Fein (SF), Plaid

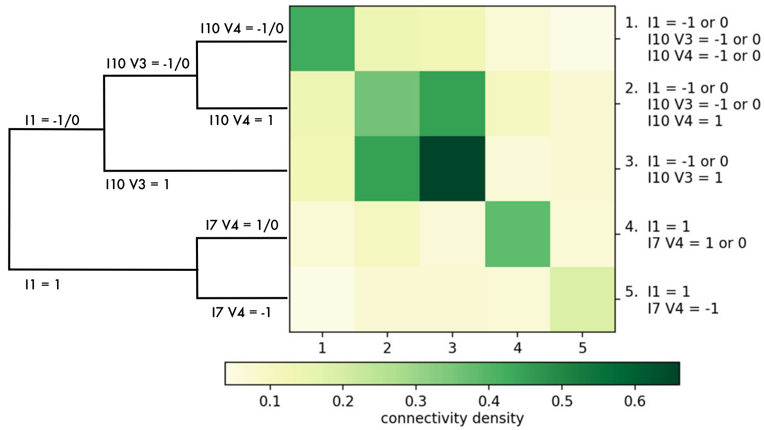


Fig. 16 The heatmap representation of the density matrix among subgroups obtained by running our algorithm for 4 iterations on *MPvotes*, aligned with a dendrogram illustration of the splitting hierarchy on the left. A darker color of each square indicates a higher connectivity density between a subgroup (represented by row) and another one (represented by column) (Color figure online)

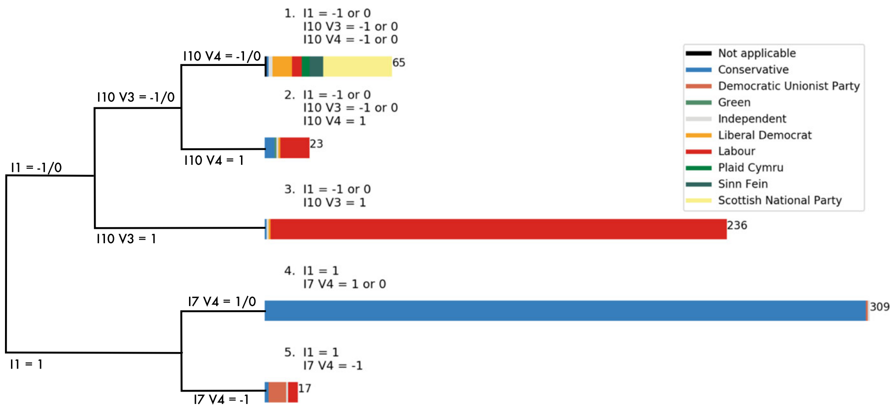


Fig. 17 The distribution of party affiliations of MPs in each subgroup, aligned with a dendrogram illustrating the splitting hierarchy on the left. For each bin corresponding to a subgroup, the subgroup description is placed on the top, and the number of MPs in this subgroup is placed on the right end. The rectangular length of a particular color inside a bin is proportional to the number of MPs affiliated with a particular party in this subgroup (Color figure online)

Cymru (PC), Green (Grn) and the majority of MPs in Labour (Lab) voted against I1 or abstained (the aggregation of the first, second and third subgroup). All except two MPs from Conservative (Con) and all from Democratic Unionist Party (DUP) were in favour (the aggregation of the fourth and fifth subgroup). Then those ‘Noes’ and abstainers of I1 are divided according to their stances on Lab’s plan for a close economic relationship with the EU (i.e., I10 V3). ‘Ayes’ of I10 V3 (i.e., the third subgroup) are dominated by most MPs from Lab. The others are further split over their votes on UK membership of Efta and Eea (i.e., I10 V4), in which MPs from some non-mainstream parties voted for or abstained (i.e., the firstst subgroup) and 15 MPs from Lab voted against. In the fourth split of vote on I7 V4, MPs affiliated with

Con and those with DUP are clearly separated from each other, leading to the fourth and fifth subgroup respectively.

Then we show (b) by combining our summarization (Fig. 16) and the party affiliation distribution (Fig. 17). Here we show some interesting findings. As mentioned previously, one bi-subgroup pattern reveals frequent connections between the second subgroup and the third one. The second subgroup can be interpreted as a group of unrepresentative Lab MPs, whereases the third subgroup corresponds to a representative group, as closer inspection shows MPs in either of these two subgroups are mostly affiliated with Lab, though the population of the second subgroup is much smaller. Also, MPs affiliated with some non-mainstream parties (e.g., SNP, LD,SF,PC) connect much more to those affiliated with Lab than those with Con, especially those with Lab belonging to the second subgroup. Although the fourth subgroup is almost made up with purely MPs that are from Con, its relatively small self-connectivity in comparison with that to the first and the third subgroup indicates not many MPs from Con build friendship with each other.

References

Adhikari B, Zhang Y, Bharadwaj A, Prakash BA (2017) Condensing temporal networks using propagation, pp 417–425. https://doi.org/10.1137/1.9781611974973.47

Adriaens F, Lijffijt J, De Bie T (2017) Subjectively interesting connecting trees. In: Ceci M, Hollmén J, Todorovski L, Vens C (eds) Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2017, Skopje, Macedonia, Sept 18–22, 2017, Proceedings, Part II, Springer, vol 10535, pp 53–69. https://doi.org/10.1007/978-3-319-71246-8_4

Akoglu L, Tong H, Meeder B, Faloutsos C (2012) PICS: parameter-free identification of cohesive subgroups in large attributed graphs, pp 439–450. https://doi.org/10.1137/1.9781611972825.38

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc Natl Acad Sci 106(51):21544–21549. https://doi.org/10.1073/pnas.0908800106

Atzmueller M (2015) Subgroup discovery. WIREs Data Min Knowl Discov 5(1):35–49. https://doi.org/10.1002/widm.1144

Atzmueller M, Doerfel S, Mitzlaff F (2016) Description-oriented community detection using exhaustive subgroup discovery. Inf Sci 329:965–984. https://doi.org/10.1016/j.ins.2015.05.008

Barbieri N, Bonchi F, Manco G (2014) Who to follow and why: link prediction with explanations. In: The 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD ’14, New York, NY, USA, Aug 24–27, 2014, pp 1266–1275. https://doi.org/10.1145/2623330.2623733

Boley M, Mampaey M, Kang B, Tokmakov P, Wrobel S (2013) One click mining: interactive local pattern discovery through implicit preference and performance learning. In: IDEA ’13 proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics, ACM, New York, NY, USA 2013, pp 27–35. https://doi.org/10.1145/2501511.2501517

Cantador I, Brusilovsky P, Kuflik T (2011) 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011) In: Proceedings of the 5th ACM conference on recommender systems. ACM, New York, NY, USA, RecSys 2011

Casiraghi G, Nanumyan V, Scholtes I, Schweitzer F (2016) Generalized hypergeometric ensembles: statistical hypothesis testing in complex networks. arXiv:1607.02441

Chen C, Lin CX, Fredrikson M, Christodorescu M, Yan X, Han J (2009) Mining graph patterns efficiently via randomized summaries. Proc VLDB Endow 2:742–753

Chen X, Kang B, Lijffijt J, De Bie T (2020) ALPINE: active link prediction using network embedding. arXiv e-prints arXiv:2002.01227

Cheng H, Zhou Y, Yu JX (2011) Clustering large attributed graphs: a balance between structural and attribute similarities. ACM Trans Knowl Discov Data (TKDD) 5(2):12:1–12:33. https://doi.org/10.1145/1921632.1921638

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Stat 23(4):493–507. https://doi.org/10.1214/aoms/1177729330

De Bie T (2011a) An information theoretic framework for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '11, pp 564–572. https://doi.org/10.1145/2020408.2020497

De Bie T (2011b) Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Min Knowl Discov 23(3):407–446. https://doi.org/10.1007/s10618-010-0209-3

De Bie T (2013) Subjective interestingness in exploratory data mining. In: Proceedings of the 12th international symposium on advances in intelligent data analysis XII—volume 8207, Springer, Berlin, IDA 2013, pp 19–31. https://doi.org/10.1007/978-3-642-41398-8_3

Deng J, Kang B, Lijffijt J, Bie TD (2020) Explainable subgraphs with surprising densities: a subgroup discovery approach. In: Proceedings of the 2020 SIAM international conference on data mining, Cincinnati, Ohio, USA

Fond TL, Neville J (2010) Randomization tests for distinguishing social influence and homophily effects. In: Proceedings of the 19th international conference on world wide web, WWW '10, ACM, pp 601–610

Freeman LC (1978) Segregation in social networks. Sociol Methods Res 6(4):411–429. https://doi.org/10.1177/004912417800600401

Fronczak A (2012) Exponential random graph models. arxiv:1210.7828

Galbrun E, Gionis A, Tatti N (2014) Overlapping community detection in labeled graphs. Data Min Knowl Discov 28(5–6):1586–1610. https://doi.org/10.1007/s10618-014-0373-y

Gong NZ, Talwalkar A, Mackey L, Huang L, Shin ECR, Stefanov E, Shi ER, Song D (2014) Joint link prediction and attribute inference using a social-attribute network. ACM Trans Intell Syst Technol 5(2):1–20. https://doi.org/10.1145/2594455

Günnemann S, Färber I, Boden B, Seidl T (2010) Subspace clustering meets dense subgraph mining: a synthesis of two paradigms. In: 2010 IEEE international conference on data mining, pp 845–850. https://doi.org/10.1109/ICDM.2010.95

Günnemann S, Boden B, Seidl T (2011) DB-CSC: a density-based approach for subspace clustering in graphs with feature vectors. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M (eds) Machine learning and knowledge discovery in databases. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 565–580

Harris JK (2013) An introduction to exponential random graph modeling, vol 173. Sage Publications, Beverly Hills

Hassanlou N, Shoaran M, Thomo A (2013) Probabilistic graph summarization. In: Wang J, Xiong H, Ishikawa Y, Xu J, Zhou J (eds) Web-age information management. Springer, Berlin, pp 545–556

Herrera F, Carmona CJ, González P, del Jesus MJ (2011) An overview on subgroup discovery: foundations and applications. Knowl Inf Syst 29(3):495–525. https://doi.org/10.1007/s10115-010-0356-2

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. J Ame Stat Assoc 58(301):13–30. https://doi.org/10.1080/01621459.1963.10500830

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76(373):33–50. https://doi.org/10.1080/01621459.1981.10477598

Lemmerich F, Becker M (2018) pysubgroup: easy-to-use subgroup discovery in python. In: Joint European conference on machine learning and knowledge discovery in databases, pp 658–662

Li J, Wu L, Zaïane O, Liu H (2017) Toward personalized relational learning. In: Proceedings of the 17th SIAM international conference on data mining, SDM 2017, Society for Industrial and Applied Mathematics Publications, United States, pp 444–452

Liu Y, Safavi T, Dighe A, Koutra D (2018) Graph summarization methods and applications: a survey. ACM Comput Surv (CSUR) 51(3):62:1–62:34. https://doi.org/10.1145/3186727

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annu Rev Sociol 27(1):415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Meeng M, Knobbe A (2011) Flexible enrichment with cortana–software demo. In: Proceedings of BeneLearn, pp 117–119

Moser F, Colak R, Rafiey A, Ester M (2009) Mining cohesive patterns from graphs with feature vectors. In: Proceedings of the 2009 SIAM international conference on data mining, pp 593–604. https://doi.org/10.1137/1.9781611972795.51

Mougel PN, Plantevit M, Rigotti C, Gandrillon O, Boulicaut JF (2010) Constraint-based mining of sets of cliques sharing vertex properties. In: Workshop on analysis of complex networks ACNE'10 co-located with ECML PKDD 2010, Barcelona, Spain, pp 48–62. https://hal.archives-ouvertes.fr/hal-01381539

Newman M (2006) Modularity and community structure in networks. Proce Natl Acad Sci 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103

Nicosia V, Mangioni G, Carchiolo V, Malgeri M (2009) Extending the definition of modularity to directed graphs with overlapping communities. J Stat Mech Theory Exp 03:P03024. https://doi.org/10.1088/1742-5468/2009/03/p03024

Perozzi B, Akoglu L, Iglesias Sánchez P, Müller E (2014) Focused clustering and outlier detection in large attributed graphs. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '14, pp 1346–1355. https://doi.org/10.1145/2623330.2623682

Pool S, Bonchi F, Leeuwen M (2014) Description-driven community detection. ACM Trans Intell Syst Technol (TIST) 5(2):28:1–28:28. https://doi.org/10.1145/2517088

Shi L, Tong H, Tang J, Lin C (2015) Vegas: visual influence graph summarization on citation networks. IEEE Trans Knowl Data Eng 27(12):3417–3431. https://doi.org/10.1109/TKDE.2015.2453957

Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BP, Wang K (2015) An overview of Microsoft Academic Service (MAS) and applications. In: Proceedings of the 24th international conference on world wide web, ACM, pp 243–246

Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 990–998

Tian Y, Hankins RA, Patel JM (2008) Efficient aggregation for graph summarization. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data, ACM, New York, NY, USA, SIGMOD '08, pp 567–580. https://doi.org/10.1145/1376616.1376675

Traud AL, Mucha PJ, Porter MA (2012) Social structure of facebook networks. Physica A Stat Mech Appl 391(16):4165–4180. https://doi.org/10.1016/j.physa.2011.12.021

Tsourakakis C, Bonchi F, Gionis A, Gullo F, Tsiarli M (2013) Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '13, pp 104–112. https://doi.org/10.1145/2487575.2487645

van Leeuwen M, De Bie T, Spyropoulou E, Mesnage C (2016) Subjective interestingness of subgraph patterns. Mach Learn 105(1):41–75. https://doi.org/10.1007/s10994-015-5539-3

Wang X, Jin D, Cao X, Yang L, Zhang W (2016) Semantic community identification in large attribute networks. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, AAAI Press, AAAI'16, pp 265–271. http://dl.acm.org/citation.cfm?id=3015812.3015851

Wei X, Xu L, Cao B, Yu PS (2017) Cross view link prediction by learning noise-resilient representation consensus. In: Proceedings of the 26th international conference on world wide web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17, pp 1611–1619. https://doi.org/10.1145/3038912.3052575

Wu Y, Zhong Z, Xiong W, Jing N (2014) Graph summarization for attributed graphs. In: 2014 International conference on information science, electronics and electrical engineering, vol 1, pp 503–507. https://doi.org/10.1109/InfoSEEE.2014.6948163

Xu Z, Ke Y, Wang Y, Cheng H, Cheng J (2012) A model-based approach to attributed graph clustering. In: Proceedings of the 2012 ACM SIGMOD international conference on management of data, ACM, New York, NY, USA, SIGMOD '12, pp 505–516. https://doi.org/10.1145/2213836.2213894

Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowl Inf Syst 42(1):181–213. https://doi.org/10.1007/s10115-013-0693-z

Yin Z, Gupta M, Weninger T, Han J (2010) A unified framework for link recommendation using random walks. In: 2010 international conference on advances in social networks analysis and mining, pp 152–159. https://doi.org/10.1109/ASONAM.2010.27

Zhang N, Tian Y, Patel JM (2010) Discovery-driven graph summarization. In: 2010 IEEE 26th international conference on data engineering (ICDE 2010), pp 880–891. https://doi.org/10.1109/ICDE.2010.5447830

Zhou Y, Cheng H, Yu JX (2009) Graph clustering based on structural/attribute similarities. Proc VLDB Endow 2(1):718–729. https://doi.org/10.14778/1687627.1687709

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.