

Discovering Representative Attribute-stars via Minimum Description Length

Jiahong Liu^{1†}, Min Zhou^{2§}, Philippe Fournier-Viger^{3§}, Menglin Yang⁴, Lujia Pan², Mourad Nouioua⁵

¹Harbin Institute of Technology, Shenzhen, China

²Huawei Noah’s Ark Lab, Shenzhen, China

³Shenzhen University, Shenzhen, China

⁴The Chinese University of Hong Kong, Hong Kong SAR, China

⁵University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria

{jiahong.liu21, mouradnouioua}@gmail.com, {zhoumin27, panlujia}@huawei.com, philfv@szu.edu.cn, mlyang@cse.cuhk.edu.hk

Abstract—Graphs are a popular data type found in many domains. Numerous techniques have been proposed to find interesting patterns in graphs to help understand the data and support decision-making. However, there are generally two limitations that hinder their practical use: (1) they have multiple parameters that are hard to set but greatly influence results, (2) and they generally focus on identifying complex subgraphs while ignoring relationships between attributes of nodes. Graphs are a popular data type found in many domains. Numerous techniques have been proposed to find interesting patterns in graphs to help understand the data and support decision-making. However, there are generally two limitations that hinder their practical use: (1) they have multiple parameters that are hard to set but greatly influence results, (2) and they generally focus on identifying complex subgraphs while ignoring relationships between attributes of nodes. To address these problems, we propose a parameter-free algorithm named CSPM (Compressing Star Pattern Miner) which identifies star-shaped patterns that indicate strong correlations among attributes via the concept of conditional entropy and the minimum description length principle. Experiments performed on several benchmark datasets show that CSPM reveals insightful and interpretable patterns and is efficient in runtime. Moreover, quantitative evaluations on two real-world applications show that CSPM has broad applications as it successfully boosts the accuracy of graph attribute completion models by up to 30.68% and uncovers important patterns in telecommunication alarm data.

I. INTRODUCTION

Data describing relationships between objects as a graph are abundant in many fields such as bioinformatics, chemistry, social network, and telecommunication [1], [2], [3], [4]. To help on the understanding of graph data and support decision-making, various algorithms have been developed for mining interesting patterns [1], [5], [6]. Nonetheless, these algorithms have two main limitations:

First, most algorithms focus on extracting patterns from the topology (e.g., frequent subgraphs) [1], [2]. However, they overlook the information carried by the attributes which is also important to understand the graph’s properties. For example, in a social network, the multiple attribute values associated with users are more informative than the topological structure. Thus, uncovering the correlation among attribute values can

further help to understand users’ characteristics and assist for several tasks such as user profile inference [7], [8]. A second example is telecommunication networks where devices (vertices) are connected by telecommunication links (edges), and a device can raise various alarms or errors (attribute values). Discovering relationships between these alarms is useful for network fault management to reduce the downtime and the maintenance cost [9].

Second, graph pattern mining algorithms typically require the setting of multiple parameters to obtain patterns [2], [10], [11]. However, finding suitable values for those parameters is time-consuming and often unintuitive. For instance, in frequent subgraph mining, a parameter called minimum support must be set. The setting of this parameter is hard as it depends on the dataset’s characteristics that are initially unknown to the user. Finding suitable values for those parameters is typically difficult and usually done by trial and error. In fact, if parameters are not set properly, too few or too many patterns may be found, which may lead to miss important information or to find many spurious patterns.

To address the above limitations, this paper proposes an algorithm, named CSPM (Compressing Star Pattern Miner), for identifying representative patterns in attributed graphs.

These patterns, named *attribute-stars* (*a-star*), are star-shaped and designed to reveal strong relationships between attribute values of connected nodes, in terms of conditional entropy.

An *a-star* indicates that if some attribute values occur in a node, some other attribute values will appear in some neighbors of that node. This type of attribute patterns is simple yet meaningful for many domains. For instance, in social networks, an *a-star* gives the information on characteristics of a user and his/her friends (e.g., friends of a smoker also tend to be smokers).

Such patterns can be used to perform *node attribute completion* (filling missing attribute values) [8], completing user profiles based on their friends’ information.

Note that a star shape is desirable as more complex structures are generally less meaningful. For instance, it is known that the influence of the friends of a friend tends to be much weaker than that of direct friends [7]. Also, patterns with

[†]Work mainly done during an internship at Huawei Noah’s Ark Lab.

[§]Corresponding Author

complex links are hard to interpret.

CSPM is easy to use as it is a parameter-free algorithm. Besides, CSPM applies a greedy search to quickly find an approximation of the best set of patterns that maximize the compression according to the MDL (Minimum Description Length) principle. Moreover, CSPM also relies on the concept of conditional entropy to assess how strong relationships between attributes are. Experiments have been performed on several benchmark datasets and show that the proposed algorithm is efficient and reveals insightful and interpretable patterns. Moreover, it has been quantitatively verified that CSPM successfully boosts the accuracy of graph attribute completion models (e.g., GCN [12], GAT [13]) and it can uncover important patterns in telecommunication alarm data.

The rest of this paper is organized as follows. Related work is reviewed in Section II. Preliminaries are introduced in Section III. CSPM is described in Section IV. Experiments are presented in Section VI. Finally, a conclusion is drawn in Section VII.

II. RELATED WORK

One of the most popular tasks to find patterns in a graph database or single graph is frequent subgraph mining (FSM) [2]. It consists of finding all connected subgraphs having an occurrence count (support) that is no less than a user-defined minimum support threshold.

However, finding large subgraphs is sometimes unnecessary for decision-making. Hence, special cases of the FSM problem that are easier to solve have been studied, such as mining frequent trees [14], [15] and paths [16], [17]. But many graph mining algorithms can only handle graphs with a single label per node. A few algorithms can find patterns in graphs with multiple labels per node (attributed graphs) [18]. Pasquier et al. proposed to mine frequent trees in a forest of attributed trees [14], while Atzmueller et al. designed the MinerLSD algorithm to find core subgraphs in an attributed graph [16]. Besides, algorithms were designed to find temporal patterns in dynamic attributed graphs [1], [19]. However, most graph pattern mining approaches have many parameters that users generally set by trial and error, which is time-consuming and prone to errors. Unsuitable parameter values can lead to long runtime or finding too many patterns, including spurious ones.

Mining compressing patterns. To select a small set of patterns that represent well a database, an emerging solution is to find *compressing patterns*. This idea was introduced in the Krimp algorithm [20] for discovering interesting itemsets (sets of values) in a transaction database (a binary table). Krimp applies the MDL principle [21] based on the idea that the best model will be insightful as it represents key patterns of the data. A database is compressed by a model M by encoding each occurrence of a pattern in the database by a code (stored in a code table). Though Krimp does not guarantee finding the best model, it typically finds a good one. However, the binary database format of Krimp is simple, which restricts its applications. To go beyond binary tables, variations of Krimp were proposed. SeqKrimp [22] was

designed to mine compressing sequential patterns in a set of sequences and the DITTO [23] algorithm was proposed to find compressing patterns in an event sequence. For graphs, GraphMDL was introduced to mine compressing subgraphs in a labeled graph [24]. But GraphMDL requires that the user first discovers frequent subgraphs using a traditional FSM algorithm [2] (similarly to Krimp). Hence, GraphMDL is not parameter-free, and results may vary widely depending on how the user sets the FSM algorithm’s parameters. Besides, GraphMDL handles a single or few attribute(s) and is not easily generalizable to multiple attributes. Hence, it fails to model richer data such as the multiple attribute values of social network users.

Differently from these approaches, this study presents a parameter-free algorithm to mine compressing patterns in graphs. To avoid relying on a traditional pattern mining algorithm, the proposed algorithm adopts an iterative approach to find a good set of compressing patterns, which is inspired by an improved version of Krimp, named SLIM [25]. During each iteration, candidates are found on-the-fly rather than using a predetermined set of patterns. Another difference distinguishing the algorithm presented in this paper from the previous work is that it handles an attributed graph as input and it finds a pattern type called *attribute-stars*. These patterns have a simple topological structure but they provide rich information about the relationships between attribute values of a node and its direct neighbors.

Summarizing and compressing a graph. A related research area is techniques for summarizing and compressing graphs. They focus on reducing storage space for large graphs to ease their processing or understanding. For example, Slash-Burn [26] removes high degree nodes of a graph to create large connected subgraphs. Then, the adjacency matrix is re-ordered to achieve high compression. Another method named VOG [27] decomposes a graph into subgraphs, and each subgraph is approximated using six pattern types (full/near clique, chain, full/near bipartite core, and star). The approximation having the MDL is selected. Another algorithm called HOSD [28] compresses a graph by searching for subgraph types of fixed size at multiple scales (from a local to a global perspective). But like VOG, HOSD focuses on compressing a graph’s topological structure and ignores node attributes.

To summarize a graph based on its topological and attribute structure, Greedy-Merge [29] adds virtual edges between nodes with same attribute values and then finds clusters of vertices that have a similar topological structure and attributes. The QB-ULSH [30] algorithm compresses an attributed graph by replacing nodes with similar edges by super-nodes using the MDL principle. However, the user needs to set multiple parameters, which directly influences results. Another recent approach is ANets [31]. It summarizes a directed weighted attributed graph by replacing nodes and edges by super-nodes and super-edges, through a process of matrix perturbations. The CSPM algorithm proposed in this paper is different. It is parameter-free and finds patterns that explain relationships between many attribute values of connected nodes. Though

TABLE I: Comparison between CSPM and related work

	CSPM	Krimp	SLIM	GraphMDL	VOG
Attributed graph?	✓	✗	✗	✗	✗
Attribute patterns?	✓	✗	✗	✗	✗
Compressing patterns?	✓	✓	✓	✓	✗
On-the-fly candidates?	✓	✗	✓	✗	✗

CSPM is compression-based, the compression is a mean to obtain representative patterns rather than the goal.

Remark 1: The input data, output patterns, and the focus of CSPM are quite different from GraphMDL and VOG (see Table I). VOG focuses on summarizing the topology of patterns without attributes. GraphMDL is designed to find patterns in many small graphs instead of a large one. SLIM has a goal and procedure similar to CSPM since candidates are found on-the-fly rather than using a pre-determined set of candidates, but it can't handle attributed graphs.

III. PRELIMINARIES

In this section, important definitions related to graphs and the discovery of compressing patterns are presented.

Graphs. A graph $G = (V, E)$ is composed of two finite sets: a vertex set V and an edge set E . The vertex set V is a set of one or more elements called *vertices*. The edge set has zero or more elements called *edges*, where $E \subseteq V \times V$. A vertex u can reach and is adjacent to a vertex v if $\{u, v\} \in E$. An edge e_1 is adjacent to an edge e_2 if they contain the same vertex. A graph G is *connected* if any vertex u can be reached from any other vertex v by traversing a sequence of adjacent edges (a *path*). An *attributed graph* $G = (A, \lambda, V, E)$ is a set V of vertices, a set E of edges, a set of nominal attributes A , and a relation $\lambda : V \mapsto A$ that maps vertices to attribute values. Attributed graphs can model data from many domains such as social networks where nodes are persons, edges are relationships between persons, and attribute values describe a person's profile in terms of properties such as age, gender, and city. In this paper, the input is an attributed graph with nominal attributes, that is connected and does not contain self-loops (and edge from a vertex to itself). For example, Fig. 1(a) shows an attributed graph that will be the running example. It has five vertices $V = \{v_1, v_2, \dots, v_5\}$, three attribute values (a, b and c). For instance, v_2 has attribute values a and c .

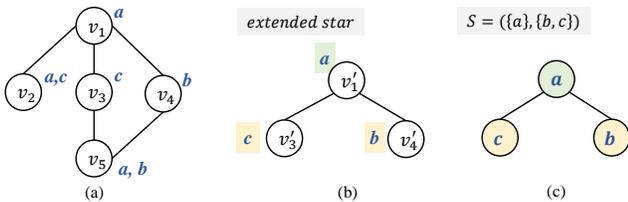


Fig. 1: An example of attributed graph

A graph can be represented as a *vertex adjacency list*, i.e. a list of tuples, where each tuple contains a vertex followed by the set of its adjacent vertices. For example, the adjacency list of the graph of Fig. 1 (a) is $\{(v_1, \{v_2, v_3, v_4\}), (v_2, \{v_1, v_5\}), (v_3, \{v_1, v_5\}), (v_4, \{v_1, v_5\}), (v_5, \{v_3, v_4\})\}$.

Many types of structural patterns were found in graphs in previous work such as trees, paths, cliques, bi-partite cores,

stars, and other complex subgraphs [1], [5], [27]. A *star* is a graph where a vertex w called the *core* is adjacent to all other vertices (called *leaves*) and no leaves are adjacent to each other. Formally, a *star* X is an undirected graph $X = (V, E, L, c)$ where V is a set of vertices, $c \in V$ is a vertex called the *core*, $L = V - \{c\}$ is a non empty set of vertices called *leaves*, and there are edges between the core and each leaf, i.e. $E = \{\{c, w\} | w \in V \wedge w \neq c\}$.

The definition of star does not consider attributes. To add attributes, the concept of star is extended as follows: An *extended star* $X = (V, E, L, c, A, \lambda)$ is a star with a set of attribute values A and a function $\lambda : V \mapsto A$ that maps vertices to categorical attribute values. To be able to locate occurrences of an extended star in an attributed graph, the concept of appearance is defined. An extended star $X = (V_x, E_x, L, c, A_x, \lambda_x)$ is said to *appear in* an attributed graph $G = (A_y, \lambda_y, V_y, E_y)$ if there exists a bijective mapping $f : V_x \rightarrow V_y$ (where $V_z \subseteq V_y$) that meets two conditions: First, for each pair $(v, a) \in \lambda_x$, there exists a corresponding pair $(f(v), a) \in \lambda_y$. Second, for each edge $(u, v) \in E_x$, there is a corresponding edge $(f(u), f(v)) \in E_y$.

Compressing Patterns. The basic framework for mining compressing patterns was defined in Krimp [20]. The aim is to discover the model M (a set of patterns) that best describes the database. The quality of a model is evaluated using the Minimum Description Length (MDL) principle on the basis that a model that is small and compresses well the database will capture the key information that the database contains.

In Krimp, a model M is composed of a set of patterns (sets of items, i.e. itemsets) and their corresponding code lengths. To compress a database, each itemset X of a model is replaced by a unique code. Patterns that appear frequently are given smaller codes to maximize compression. In Krimp, the length of a code for an itemset X is represented as a number of bits and is calculated using the Shannon Entropy as $L(X) = -\log_2 P(X)$, where $P(X)$ is the relative occurrence frequency of X . Before building a model, Krimp creates a two-column table called *Standard Code Table (ST)* where each line stores an item from the dataset in the left column and its corresponding code length in the right column. Then, the length of the original database $L(D|ST)$ without compression can be calculated by replacing all items in the database by their codes in ST and summing their code lengths with that of the ST . To build a better model M , Krimp initializes $M = \emptyset$ and then iteratively adds a pattern to the model if it improves compression. Each pattern is an itemset consisting of one or more items. A model is represented as a code table CT assigning a code length to each pattern. The description length (DL) of the database compressed with CT is calculated as $L(CT, D) = L(CT|D) + L(D|CT)$, where $L(CT|D)$ is the length of the code table, and $L(D|CT)$ is the length of the database encoded with that code table. To find a good model (a code table), Krimp employs a greedy search. Note that the ultimate goal of compressing pattern mining algorithms is to find a good set of patterns that describe the data well, instead of actually compressing the data. Hence, only the code length

of each pattern is necessary, which can be directly obtained by well derived formulas (e.g. Shannon Entropy).

IV. THE CSPM ALGORITHM

This section describes the proposed CSPM algorithm to discover star-shaped attribute patterns revealing interesting relationships between attributes in an attributed graph.

A. Pattern Format and Problem Statement

The goal of this study is to mine **attribute** patterns that can reveal strong inner relationships between attribute values with implicit connection information, instead of finding sub-graphs or substructures meeting the given requirements (e.g. frequency). With this premise, a star-shaped pattern format is selected. This format has some advantages. First of all, a star-shaped attribute pattern is simple, easy to understand, and can indicate correlation between attribute values of directly connected nodes (illustrated by edges). What's more, for practical applications such as social network analysis, a star-shaped attribute pattern can capture the influence relationship between a person and its friends [32]. This pattern format also has industrial application for network alarm analysis as it will be shown in Section VI.

Although extended stars (defined in Section III) could reveal interesting patterns, a more general type of patterns is defined to summarize multiple extended stars. The goal is to focus less on the structure of stars and more on the relationships between attribute values of cores and leaves. The proposed pattern type $S = (S_c, S_L)$, namely *attribute-star* (*a-star*), where S_c is a set of attribute values of a core node and S_L is a set of attribute values that appear in any of its leaf nodes. The sets S_c and S_L of an attribute-star are called *coreset* and *leafset*, and their elements, *core values* and *leaf values*, respectively.

The relationship between an attribute-star and a star is defined as follows: An attribute-star $S = (S_c, S_L)$ is *matching* with a star $X = (V, E, L, c, A, \lambda)$, if (1) $\forall a \in S_c$, there is a pair $(c, a) \in \lambda$ and (2) $\forall y \in S_L$, there is a leaf $u \in L$ such that $(u, y) \in \lambda$. Thus, an attribute-star can match with multiple stars which gives the possibility to summarize them. For example, consider the a-star $S = (\{a\}, \{b, c\})$ in Fig. 1(c). It indicates that a core has the attribute value a and some leaves have attribute values b and c . This a-star matches with a star depicted in Fig. 1(b).

In the proposed method, the role of core values within an a-star is different from that of leaf values. More precisely, the core values are viewed as influencing leaf values. In other words, for each a-star that is discovered, if core values appear in the core vertex, there is a high chance that the leaf values appear in its neighbor(s).

Problem statement. This study aims to mine compressing patterns in an attributed graph. The *input* is an attributed graph $G = (A, \lambda, V, E)$ with categorical attribute values. The *goal* is to find the set of a-stars $\{S_1, S_2, \dots, S_n\}$ that best compresses the original information of the attributed graph G losslessly, i.e., with the minimum description length. The *output* is a set

of a-stars ordered by ascending code lengths. An a-star with a shorter code length indicates that it is more informative.

Note that we aim to find a subset of a-stars that meets the above *goal* instead of finding *all* possible a-stars. For this problem, an approximate algorithm is desired since the brute force approach is impracticable. For instance, if only 500 a-stars exist, there is $2^{500} - 1$ possible subsets.

B. The Inverted Database Representation

To support the efficient discovery of a-stars, the CSPM algorithm transforms the input database into an *inverted database* (I) representation. This latter allows to easily find a-stars by merging lines representing smaller a-star patterns together. Finding a-stars that best compress the database using the MDL then becomes a problem of selecting appropriate pairs of lines to be merged to make new a-stars.

The inverted database representation is based on the observation that an attributed graph contains two parts: The topology (a vertex adjacency list) and the mapping function between vertices and attributes. Each tuple in the adjacency list can be viewed as a star associating a vertex (the core) to a list of adjacent vertices (the leaves). For example, the tuple $(v_1, \{v_2, v_3, v_4\})$ in the adjacency list of the graph of Fig. 1(a) is a star. Note that any vertex from a graph can be the core of a star. For that reason, a-stars can be directly found by substituting the vertices in the adjacency list with their corresponding attribute values. For instance, the tuple $(v_1, \{v_2, v_3, v_4\})$ could be used to form the a-star $(\{a\}, \{a, b, c\})$. Considering the fact that the core in an a-star has a different role than the leaves, each attribute value can be labeled as a core value or leaf value according to the vertices where it appears.

The inverted database is a three-column table. The first column S_L contains leafsets, the second column S_c indicates the coresets that are connected to the corresponding leafsets in the first column to form an a-star, and the third column contains the set of vertices where the core values appears, which are called the *positions*. A mapping table can be also built to map each core to its list of positions.

Taking the graph depicted in Fig. 1 as example, the mapping table of this graph is shown in Fig. 2(a), and the inverted database generated from this graph is presented in Fig. 2(b). From this mapping table, it can be seen that the coreset $S_c : \{c\}$ appears at vertices v_2 and v_3 . Moreover, from the graph, it can be observed that $\{a\}$ is adjacent to $\{c\}$ in both cases at v_2 and v_3 . Accordingly, there is a blue record $\{\{a\}, \{c\}, \{v_2, v_3\}\}$ in the inverted database of Fig. 2(b).

There are several advantages of using this inverted database representation over covering attribute patterns directly on the original graph:

First, it is easy to calculate the code length of each a-star. The reason is that the co-occurrence frequency of leaf values that are adjacent to the same core value can be directly obtained by intersecting their position sets. Second, there is no need to scan the whole database again after the discovery of each new a-star pattern. This is because each line of

S_c	Position
a	v_1, v_2, v_5
b	v_4, v_5
c	v_2, v_3

S_L	S_c	Position
a	a	v_1, v_2
	b	v_4
	c	v_2, v_3
b	a	v_1, v_5
	b	v_4, v_5
	c	v_3
c	a	v_1, v_5
	b	v_5

Fig. 2: Mapping table and inverted database

the database is initialized as a basic a-star with one edge (core value, leaf value). Hence, all a-stars can be generated by merging two or more lines together. Third, the inverted database allows storing a-stars in a non-redundant way where each line of the inverted database represents a distinct a-star. Thus, it is convenient to encode each a-star by appending a distinct code for each line, instead of building an extra translation table.

C. MDL with the Inverted Database.

The CSPM algorithm is inspired by the basic framework for mining compressing patterns of Krimp [20]. Krimp relies on a two-column standard code table ST and a code table CT to calculate the total description length of the compressed database based on the MDL principle. But the data and pattern types considered by Krimp are simple as it only finds itemsets (sets of symbols) in a transaction database (a binary table) [20]. In this paper, the goal is to discover a-stars in an attributed graph, which is more complex as patterns have two parts which are core values and leaf values.

Therefore, the Krimp model is adapted to define a model M that is suitable for a-star patterns. The proposed model type contains two code tables, named CT_c and CT_L . The former is a traditional two-column code table used to encode coresets, while CT_L is used to encode leafsets. This latter contains pointers to CT_c to facilitate appending the codes of coresets and leafsets to calculate the description length. CT_L has a similar structure to the inverted database. The difference is that the third column of CT_L contains leaf values' codes instead of sets of positions.

Before explaining how CSPM calculates the DL, it is important to notice that it is necessary to first build the *standard code table* ST based on the frequencies of all attributes in the attributed graph. The standard code table is the optimal encoding of all attributes without labels and structure information. It contributes to the cost (description length) of coresets and leafsets stored in code tables. In the case where patterns have a single core value, the format of the code table CT_c is the same as that of the standard code table.

For the sake of simplicity, we first discuss how to mine a-stars having a single attribute value as core values. Then, the more general case of multiple core values will be explained in subsection IV-F. In the case of single-core value patterns, a-stars are found by simply finding the pattern set of S_L adjacent values for different core values to form the different a-stars. As example, Fig. 3(a) shows the CT_c and CT_L code tables for the graph depicted in Fig. 1. In that figure, numbers under the label

Usage indicate the frequency of CT_c in the attributed graph derived from the mapping function table. Note that this usage information is not a part of the CT_c table. In CT_L , the first two columns can be obtained by combining the corresponding codes of $Code_c$ in CT_c . Moreover, the column $Code_L$ records all codes of leafsets S_L adjacent to different coresets. To obtain the code of an a-star, a pointer from the column S_c to the corresponding core values S_c in CT_c is used.

In Fig. 3(a), values are provided in a column under the label f_L/f_c . The f_L means the frequency of the corresponding a-star (line) in CT_L , while f_c is that of the corresponding S_c in the inverted database. Note that, the column f_L/f_c is not a part of CT_L . Consider the a-star $(\{a\}, \{b\})$ as an example. This a-star can be represented by Fig. 3(b).

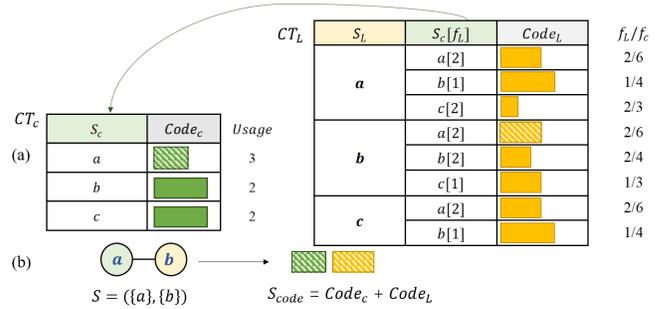


Fig. 3: Two code tables and an a-star $S = (\{a\}, \{b\})$ encoding

The description length $L(M, I)$ of a model and inverted database is then calculated by the following equations as Equation (1):

$$L(M, I) = L(M) + L(I|M). \quad (1)$$

$$L(M) = L(CT_c|I) + L(CT_L|I). \quad (2)$$

Note that, the calculation of $L(M)$ seen as Equation (2) is similar to the equation used by the Krimp algorithm. It is calculated by adding the cost of all columns together.

The length $L(I|M)$ is derived directly from the proposed inverted database representation by Equation (3):

$$L(I|M) = \sum_{S \in I_{new}} L(S_{code}) * S(f_L), \quad (3)$$

$$L(S_{code}) = L(Code_c) + L(Code_L). \quad (4)$$

where $S(f_L)$ is the frequency of the a-star (line) S in the new inverted database I_{new} , which is described using M . And $L(S_{code})$ is the code length of S by summing up the code lengths of its coreset $Code_c$ and leafset $Code_L$.

The CSPM algorithm is designed to find a-stars that indicate strong relationships between core values and leaf values according to the Minimum Description Length (MDL) principle. That is, CSPM is designed to find a-stars meet the following two conditions. First, an interesting a-star is supposed to be frequent to some extent. And the second condition is that the leaf values need to have high relevance for some core values while being less relevant to other core values. Relevance is assessed through the conditional entropy, described next.

D. Conditional Entropy

The code length of an a-star is required to calculate the DL of a model. The proposed method differs from traditional methods such as Krimp that are based on the Shannon entropy to obtain optimal lengths. There are two reasons. First, the proposed model considers two entities (based on core and leaf values), while Shannon entropy is only suitable for the case of one variable. Second, our goal is to evaluate the relevance between leaf values and core values instead of merely focusing on the frequency.

Calculating the code $Code_c$ of core values in M is only based on the frequency of attribute values in the mapping function between vertices and attribute values, as Equation (5). To optimally encode leaf values connected to each coreset

in each a-star, conditional entropy is used as a key measure.

$$L(cv) = -\log(P(cv)). \quad (5)$$

In information theory, the conditional entropy $H(Y|X)$ measures the cost of describing a random variable Y supposing that the value of another random variable X is given [33]. Similarly, in this paper, the set of coresets S_c^M can be regarded as X and the set of leafsets S_L^M as Y which can be calculated given that the coresets S_c have been obtained. Based on this fact, the relevance of each coresets to its leafset (a set of leaf values) can be revealed using the conditional entropy. According to the principle of conditional entropy, the code length of each a-star (each line in CT_L) with leafset S_L and coresets S_c is given by Equation (6). Hence, the $Code_L$ for each line is only determined by the corresponding f_L and f_c values.

$$L(Y = S_L|X = S_c) = -\log \frac{p(S_L, S_c)}{p(S_c)} = -\log \frac{f_L}{f_c}. \quad (6)$$

Then, $H(Y|X)$ is calculated by Equation (7), which represents the average encoding cost of each line in CT_L after using the conditional entropy encoding method.

$$\begin{aligned} H(Y|X) &= - \sum_{S_c \in X, S_L \in Y} p(S_c, S_L) \log \frac{p(S_c, S_L)}{p(S_c)} \\ &= - \sum_{j=1}^m \sum_{i=1}^n \frac{l_{ij}}{s} \log \frac{l_{ij}}{c_j} \end{aligned} \quad (7)$$

In that equation, s is the total frequency of a-stars (lines) in the inverted database, which is the sum of f_L , m is the total number of coresets, n is the total number of a-stars (lines) in the inverted database, l_{ij} is the frequency f_L of the i th leafset and j th coresets, and c_j is the frequency f_c of the j th coresets.

Based on the above equation, the total encoding cost of the inverted database is obtained by Equation (8). Note that $\sum_{i=1}^n l_{ij}$ is equal to the frequency of c_j .

$$\begin{aligned} L(I|M) &= -s \times H(Y|X) = \sum_{j=1}^m \sum_{i=1}^n l_{ij} \log \frac{l_{ij}}{c_j} \\ &= \sum_{j=1}^m \sum_{i=1}^n l_{ij} \log c_j - \sum_{j=1}^m \sum_{i=1}^n l_{ij} \log l_{ij} \\ &= \sum_{j=1}^m c_j \log c_j - \sum_{j=1}^m \sum_{i=1}^n l_{ij} \log l_{ij}. \end{aligned} \quad (8)$$

E. Candidate Generation

The model is built iteratively. After determining all the coresets, the problem is simplified into selecting the leafsets with minimal conditional entropy. An intuitive way of doing this is to select the best set of a-stars among all possible combinations of leafsets with each coresets. But this is time-consuming for large databases. To generate patterns without assuming that a set of candidates has been previously mined by another algorithm as in Krimp, CSPM uses a greedy approach that merges the two leafsets that provide the best gain ΔL , i.e., in terms of the difference between the DLs before/after merging. The gain of a pair represents how much the merge step will reduce the DL. Thus, a larger gain is better.

Suppose that p is the pair of leafsets $\{(S_{Lx}, S_{Ly}) | S_{Lx} \in CT_L, S_{Ly} \in CT_L\}$ that CSPM is going to merge. A merge operation means that a new leafset pattern $\{S_{Lx} \cup S_{Ly}\}$ will cover all situations where S_{Lx} and S_{Ly} appear around the same coresets. This co-occurrence information is easily obtained using the inverted database by intersecting all sets of *positions* of the two merged lines that have the same S_c .

Two parts influence the DL to generate a new pattern $\{S_{Lx} \cup S_{Ly}\}$. The first one is the cost increase of the new pattern's leafset in the code table, which can be easily obtained through the standard code table ST . The other one is the gain ΔL from merging the pair of leafsets. From Equation (8), $L(I|M)$, Equation (9) is obtained for a simplified calculation of ΔL :

$$\begin{aligned} \Delta L &= \underbrace{\left(\sum_{i=1}^m c_j \log c_j - \sum_{i=1}^m c'_j \log c'_j \right)}_{P_1} - \\ &\quad \underbrace{\left(\sum_{j=1}^m \sum_{i=1}^n l_{ij} \log l_{ij} - \sum_{j=1}^m \sum_{i=1}^{n'} l'_{ij} \log l'_{ij} \right)}_{P_2} \end{aligned} \quad (9)$$

where variables with prime as superscript indicate the properties of the new code table CT'_L after the merge operation. Note that the number of coresets m does not change during the process because the set of coresets S_c^M is known.

To better explain the above equation, a detailed analysis of its two parts P_1 and P_2 is given. It can be observed that P_1 is related to coresets while P_2 is related to leafsets. In the following, for the sake of brevity, the lower case variable x is used to denote the f_L of the leafset S_{Lx} , i.e. $x = S_{Lx} \cdot f_L$. Similarly, the variable y is used to denote $S_{Ly} \cdot f_L$.

It can be observed that not all frequencies for each line f_L and each coresets f_c are changed by a merge operation. Suppose C is the coresets that S_{Lx} and S_{Ly} are both connected

with. For each coreset $e \in C$, f_e is the frequency of coreset e , and xy_e is the co-occurrence frequency of S_{Lx} and S_{Ly} . By the convenience of the inverted database structure, xy_e is obtained by intersecting the *positions* of a-stars (e, S_{Lx}) and (e, S_{Ly}) directly. Especially, if xy_e is equal to zero, the two lines can't be merged and the gain is equal to zero. Thus, P_1 in Equation (9) can be derived as in Equation (10):

$$P_1 = \sum_{e \in C} (f_e \log f_e - (f_e - xy_e) \log(f_e - xy_e)). \quad (10)$$

Due to the fact that not all values of l'_{ij} will become different from l_{ij} after merging, the P_2 part is calculated by only summing up all relative merged lines using Equation (11), where, P_e means the gain for generating a new a-star ($e, S_{Lx} \cup S_{Ly}$) by merging lines with core value e .

$$P_2 = \sum_{e \in C} P_e. \quad (11)$$

There are three cases to be considered according to the relationship between xy_e , x_e and y_e .

(1) **Partly merged.** In this case, the leafsets share some positions but each leafset has some unique positions, i.e., $xy_e \neq x_e, xy_e \neq y_e$. The formulation P_e can be derived as Equation (12).

$$P_e = (x_e \log x_e + y_e \log y_e) - [(x_e - xy_e) \log(x_e - xy_e) + (y_e - xy_e) \log(y_e - xy_e) + xy_e \log xy_e]. \quad (12)$$

(2) **Two lines totally merged.** In this case, $xy_e = x_e$ and $xy_e = y_e$. The two lines (a-stars) are merged as a new pattern ($\{e\}, \{S_{Lx} \cup S_{Ly}\}$) because they always appear at the same positions. After merging, there is no a-star ($\{e\}, \{S_{Lx}\}$) or ($\{e\}, \{S_{Ly}\}$) in the inverted database or code table anymore. Thus, P_e can be calculated as Equation (13).

$$P_2 = (x_e \log x_e + y_e \log y_e) - (xy_e \log xy_e) = xy_e \log xy_e. \quad (13)$$

(3) **One line totally merged.** Only one a-star will be removed by merging. In this case, $xy_e = x_e, xy_e \neq y_e$ or $xy_e \neq x_e, xy_e = y_e$, as shown as Equation (14) and (15), respectively.

$$P_e = (x_e \log x_e + y_e \log y_e) - (y_e \log y_e + xy_e \log xy_e) = y_e \log \frac{y_e}{y_e - xy_e} + xy_e \log(y_e - xy_e). \quad (14)$$

$$P_e = (x_e \log x_e + y_e \log y_e) - (y_e \log y_e + xy_e \log xy_e) = x_e \log \frac{x_e}{x_e - xy_e} + xy_e \log(x_e - xy_e). \quad (15)$$

Take Fig. 2 as example. Suppose that the leafsets $\{b\}$ and $\{c\}$ of column S_L are merged. There are two coresets $C = \{\{a\}, \{b\}\}$ that these leafsets are both connected with. For coreset $\{a\}$, the *Position* of a-stars ($\{a\}, \{b\}$) and ($\{a\}, \{c\}$) are the same as $\{v_1, v_5\}$, which means the two lines can be totally merged as a new pattern (Case 2). And for $\{b\} \in C$, the common position is $\{v_5\}$. Hence, a-star ($\{b\}, \{c\}$) will be totally merged while the line ($\{v_4, v_5\} - \{v_5\} = \{v_4\}$) remains for a-star ($\{b\}, \{b\}$) (Case 3). The new inverted database and code length $Code_L$ after merging are shown in Fig. 4. It is

observed in Fig. 4 that leafsets $\{b\}$ and $\{c\}$ are both connected to the same coreset $\{a\}$ of vertices $\{v_1, v_5\}$ and that they are connected to coreset $\{b\}$ of vertex $\{v_5\}$. Interestingly, the code length of leafset $\{b, c\}$ in CT_L corresponding to core value $\{a\}$ is much shorter than those of previous ($\{a\}, \{b\}$) and ($\{a\}, \{c\}$). Overall, the database is compressed and its DL is reduced by the merge operation.

S_L	S_c	Position	$Code_L$	f_i/f_c
a	a	v_1, v_2	■	2/4
	b	v_4	■	1/3
	c	v_2, v_3	■	2/3
{b, c}	a	v_1, v_5	■	2/4
	b	v_5	■	1/3
b	b	v_4	■	1/3
	c	v_3	■	1/3

Fig. 4: Inverted database and $Code_L$ after merging S_L pairs ($\{b\}, \{c\}$)

F. The Detailed Algorithm

The proposed CSPM algorithm takes as input an attributed graph G (topology and mapping function) and outputs a set of compressing a-star patterns. Algorithm 1 shows the detailed procedure. It is divided into two sub-procedures for (1) determining and encoding all coresets S_c^M , (2) and constructing the final a-stars by finding an approximation of the best leafsets for each coreset using a greedy search. The main steps of CSPM are:

Step 1: Encode attribute values and core values.

CSPM reads the mapping function without topology information, where each attribute value appearing at each vertex is seen as an item (symbol). If the user wishes to mine a-stars with coresets containing more than one core values, a traditional compressing pattern mining algorithm can be applied on a transaction database composed of the attribute values of vertices. Several algorithms can be used such as Krimp [20] and SLIM [25]. In the other case, attribute values are optimally encoded according to their global frequencies. The result is a standard code table ST of attribute values and a code table CT_c for coresets S_c^M . Note that CT_c is exactly the standard code table ST if all coresets have one core value.

Step 2: Construct the inverted database. The inverted database I is created, which contains leaf values S_L , core values S_c , and the topology information between core values and leaf values. Besides, each line in the initial inverted database stores the positions of each a-star with only one leaf value. The inverted database makes it easy to generate merge candidates and to cover the database with new patterns.

Step 3: Generate leafsets merge candidates. Merging candidates allows showing the quality of leafset pairs that can be merged as a new pattern. To do so, it is necessary to calculate the gains ΔL of all possible pairs of leafsets in the inverted database. Then, pairs with $\Delta L > 0$ are sorted in descending order to form the final candidates.

Step 4: Select a-stars and update the database. A-stars are created by merging pairs in candidates that give the

maximal gain. Then, the code table CT_L and inverted database are updated. Then, Steps 3 and 4 are repeated until there is no candidate leafset pair that can compress the database more. A-stars left in model M are the set of patterns we want, where an a-star with shorter code length will have a higher ranking.

Algorithm 1: The CSPM algorithm

input : An attributed graph G with two parts: the adjacency list and the mapping function
output: Compressing a-star patterns

- 1 Encode all coresets S_c^M ; // **Step 1**
- 2 $CT_c \leftarrow (S_c, Code_c)$;
- 3 $CT_L \leftarrow (S_L, S_c, Code_L)$;
- 4 $I \leftarrow (S_L, S_c, Position)$; // **Step 2**
- 5 $candidates \leftarrow$
 generate_candidates(S_L^M); // Algorithm 2
- 6 **repeat**
- 7 // select pair with the max. gain ΔL
- 8 $p \leftarrow candidates.pop()$;
- 9 $I, CT_L \leftarrow merge(p)$; // **Step 4** (update I and CT_L)
- 10 $candidates \leftarrow generate_candidates(S_L^M)$; // **Step 3**
- 11 **until** $candidates = \emptyset$;
- 12 **return** all a-stars in M

Algorithm 2: Generate leafsets merge candidates

input : All leafsets S_L^M in CT_L , inverted database I
output: A candidate list of leafset pairs

- 1 $candidates = []$;
- 2 $P \leftarrow enumerate(S_L^M, 2)$;
- 3 **for** $p \in P$ **do**
- 4 $gain \leftarrow calculate_gain(p, I)$; // Equation (9)
- 5 **if** $gain > 0$ **then** $candidates.append(p)$;
- 6 **end**
- 7 Sort $candidates$ by descending order of gain;
- 8 **return** $candidates$

V. OPTIMIZATION

CSPM can mine a-star patterns efficiently. However, we found in empirical experiments that attributes are many while relationships between attribute values are sparse. Hence, few pairs can improve compression, among a huge number of possibilities. Besides, few gains are affected by a merge operation and need to be re-calculated and updated. Thus, it would be time-consuming to calculate all gains of all possible leafset pairs during each iteration. To improve the performance of the candidate generation process, an optimized CSPM version named CSPM-Partial is put forth. It updates only some of the gains and leafset pairs in candidates after each merge operation on account of the observation that two leafsets without a common coreset

will never be merged even if they are merged as a new pattern later. More precisely, leafset pairs that are not adjacent to the same coreset will never have positive gains. Thus, only the gains of pairs with *related* leafsets should be updated after each merge, instead of calculating the gain C_n^2 times, where

n is the leafset count in the inverted database. A leafset is called *related* to a merged pair if it contributes a positive gain when merged with a leafset in the merged pair p . For this optimization, a dictionary $rdict$ is built to store all *related* leafsets for each leafset of the inverted database.

Algorithm 3 shows the CSPM-Partial procedure. Line 6 creates the $rdict$ dictionary, which stores *key, value* pairs. Each key is a leafset and the value is the *related* leafsets that can be merged with it in $candidates$. Compared with CSPM-Basic, the generate_candidates function (line 9 in Algorithm 1) is substituted by an update operation (line 10 in Algorithm 3).

That way, after doing the merge operation in each iteration, only some gains are calculated and updated based on the previous $candidates$, instead of enumerating all possible pairs and re-calculating their gains again to generate a new $candidates$.

Algorithm 3: The CSPM-Partial algorithm

input : An attributed graph G with two parts: the adjacency list and the mapping function
output: Compressing a-star patterns

- 1 Encode all coresets S_c^M ;
- 2 $CT_c \leftarrow (S_c, Code_c)$;
- 3 $CT_L \leftarrow (S_L, S_c, Code_L)$;
- 4 $I \leftarrow (S_L, S_c, Position)$;
- 5 $candidates \leftarrow$
 generate_candidates(S_L^M); // Algorithm 2
- 6 $rdict \leftarrow related_dict(candidates)$
- 7 **repeat**
- 8 $p \leftarrow candidates.pop()$;
- 9 $I, CT_L \leftarrow merge(p)$;
- 10 $candidates, rdict \leftarrow update(p, I)$; // Algorithm 4
- 11 **until** $candidates = \emptyset$ or $len(rdict) = 0$;
- 12 **return** all a-stars in M

Details about the update operation on $rdict$ and $candidates$ (after the pair p is merged) are shown in Algorithm 4. l_{total} and l_{part} in line 1 contain the totally merged and partly merged leafsets of p , respectively. The overall update procedure consists of three operations (Remove, Add and Update). First, totally merged leafsets and related pairs are removed from $rdict$ and $candidates$ (line 4). Next, all possible *related* leafsets (rel) are obtained by intersecting $rdict[idx]$ and $rdict[idy]$ (line 6) based on the aforementioned observation. Each of them (rel) could be merged with the new leafset new_a . After calculating the gains, pairs with positive gains are added to $candidates$ and the *related* information is recorded in $rdict$ (line 10-12). Finally, the pairs whose gains are influenced by the merge operation are updated in $candidates$ (line 17-21). It is a fact that frequencies of a-stars having partly merged leafsets are always reduced by the merge operation. Thus, the influenced pairs may not contribute to compression anymore, i.e. the gains are no longer larger than zero. For this reason, the corresponding pairs and their *related* leafsets are removed from $candidates$ and $rdict$, respectively.

We here analyze the complexity of CSPM-Basic and the optimized CSPM-Partial. Given an attributed graph G with $|E|$

Algorithm 4: Update operation

input : The merged pair p , the inverted database I
output: Updated $candidates$ and $rdict$

```
1  $l_{total}, l_{part} \leftarrow merge\_state(p)$ ;  
2  $idx, idy, new_a \leftarrow p[0], p[1], \{p[0] \cup p[1]\}$ ;  
3 // (1) Remove totally merged leafsets  
4 if  $l_{total} \neq \emptyset$  then delete  $l_{total}$  in  $candidates, rdict$ ;  
5 // (2) Add pairs with new leafset  
6 for  $rel \in rdict[idx] \cap rdict[idy]$  do  
7    $pair \leftarrow (rel, new_a)$ ;  
8    $gain \leftarrow calculate\_gain(pair, I)$  // Equation (9)  
9   if  $gain > 0$  then  
10     $rdict[new_a].add(rel)$ ;  
11     $rdict[rel].add(new_a)$ ;  
12     $candidates.add(pair)$ ;  
13  end  
14 end  
15 // (3) Update influenced pairs  
16 if  $l_{part} \neq \emptyset$  then  
17   for  $rel \in rdict[l_{part}]$  do  
18     $pair \leftarrow (rel, l_{part})$ ;  
19     $gain \leftarrow calculate\_gain(pair, I)$  // Equation (9)  
20    if  $gain > 0$  then update  $pair.gain$  in  $candidates$   
21    else delete  $pair$  in  $candidates, rdict$ ;  
22  end  
23 return  $candidates, mdict$ 
```

edges, $|V|$ vertices, $|A|$ distinct attribute values. We denote $|\bar{A}|$ the average number of values per attribute, $|S_L^M|$ the number of leafsets after merging, and $|F|$ the number of generated a-stars. The time and space complexity of the two CSPM versions are analyzed in as below.

Time Complexity. First, CSPM encodes all coresets. It takes $O(1)$ time to insert elements in CT_c . Next, all the attributes are scanned in all vertices to construct the inverted database, which takes at most $O(|E| \times |\bar{A}|^2)$. Then, patterns are generated by doing merges, using at most $(|S_L^M| - |A|)$ iterations. For each iteration, CSPM-basic generates candidates (Algorithm 2) in at most $O(|S_L^M|^2)$ steps (the last iteration). The merge operation is $O(1)$ (add/remove line). CSPM-Partial optimizes this procedure by only considering some candidate pairs,

and the worst case is $O(|S_L^M|)$ in each iteration. But generally, few attributes are highly correlated with the merged leaves. Thus, the time complexity of recalculating gains in each iteration is $O(k)$, $k \ll |S_L^M|$. This is why CSPM-Partial outperforms CSPM-Basic in experiments.

Space Complexity. The dataset is stored as an inverted database, which is updated to generate a-stars. Thus, CSPM-Basic uses $O(|F|)$ space. The space complexity of CSPM-Partial is $O(|F| + |R|)$, where $|R|$ is the additional memory for keeping potentially related leaf-values, with the worst case of $O(|A| \times |A - 1|)$.

But for real applications, it is unlikely that all attributes are correlated with each other due to sparsity. Thus, the additional memory usage is generally very small.

VI. EXPERIMENTS

To evaluate the performance of the proposed CSPM algorithm Subsection VI-A first presents a runtime and gain update ratio analysis of the basic CSPM (SPM-Basic), and the optimized version (CSPM-Partial). Then, Subsection VI-B describes some simple a-stars found in real data to intuitively show that they are meaningful. The quality and usefulness of the a-stars as a whole are then evaluated in two application scenarios (Subsection VI-C and VI-D).

First, CSPM is quantitatively assessed for the popular graph attribute completion task. Then, CSPM is used for an industrial application to extract alarm correlation rules for fault management in telecommunication networks. Note that, CSPM-Partial is adopted for the two applications owing to its efficiency.

A. Evaluation of the CSPM-Partial Optimization

The first experiments were done to evaluate the benefits of using the proposed partial updating optimization.

Algorithms. Three algorithms were compared. The baseline algorithm is *CSPM-Basic* without optimization. The second algorithm is *CSPM-Partial* (CSPM with the partial updating optimization). Moreover, to provide a point of reference for experiments, the SLIM algorithm [25] is also adopted for comparison. In fact, although SLIM finds simple patterns (sets of values that co-occur) without considering inner relationship, it is close to this study as SLIM also is a compression-based algorithm and it can be easily applied to an attributed graph by treating coresets in each adjacency list tuple as items.

To our best knowledge, a-star patterns have not been considered before. Note that, graph pattern mining algorithms and summarization techniques mentioned in Section II such as VOG and GraphMDL, focus on mining different patterns with complicated structure (shown in Table I), which is quite different from our work and leads them to have a lower efficiency. Hence, these techniques are not included as baselines in this work.

Datasets. Four benchmark datasets having various characteristics were used. The statistics are summarized in Table II, where $|S_c^M|$ gives the number of coresets in the inverted database. *DBLP* [34] is a citation network indicating co-author relationships (edges) between researchers (vertices) during years 2006-2010. Attribute values are the conferences/journals a person has published in. The *DBLP-Trend* [35] is a variant of *DBLP* where the attribute indicates trends about publications. For example, (ICDE+, ICDE-, ICDE=) indicates that the number of publications in ICDE has increased, decreased, and stayed the same since the previous year, respectively. The *USFlight* [36] dataset contains data about flights (edges) between US airports (vertices) in 2005. The *Pokec* dataset¹ indicates friendship relationships (edges) between persons (vertices) on the Pokec social network. The attribute values store the musical tastes of users.

(1) **Runtime.** In the first experiment, the influence of the proposed optimization on the runtime was evaluated by com-

¹<https://stanford.io/3oZH9EI>

TABLE II: Statistics about datasets

Dataset	DBLP	DBLP-Trend	USFlight	Pokec
#Nodes	2,723	2,723	280	1,632,803
#Total edges	3,464	3,464	4,030	30,622,564
$ S_c^M $	127	271	70	914
Category	Citation	Citation	Airport	Music

paring CSPM-Partial with CSPM-Basic. Results are shown in TABLE III. Note that, the result of CSPM-Basic on the Pokec dataset is written as (–) because CSPM-Basic was not able to terminate after 48 hours. Three observations are made.

TABLE III: Runtime comparison

Dataset	SLIM	Runtime (s)	
		CSPM-Basic	CSPM-Partial
DBLP	4.69	43.13	0.98
DBLP-Trend	48.69	956.61	25.46
USFlight	1.25	10.16	1.43
Pokec	166,678.3	–	1,403.21

First, the CSPM-Basic algorithm is approximately 10 times slower than SLIM. This result is reasonable since CSPM-Basic solves a problem that has two variables (core and leaf values) and it considers the graph topology which is more complex. Moreover, CSPM-Basic merges and updates candidate pairs without using any optimization.

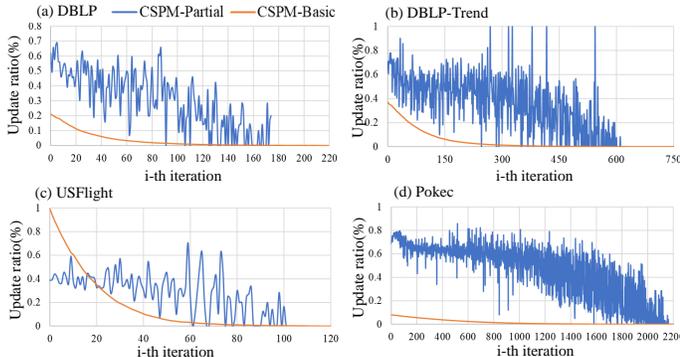


Fig. 5: Comparison of gain update ratio

Second, it was found that, the optimized algorithm, CSPM-Partial, is considerably faster than the CSPM-Basic. Especially for Pokec which is a large dataset. This result demonstrates that the proposed optimizations can make the algorithm more efficient.

A third observation is that, the runtime of CSPM-Partial increases as the number of coresets increases. In fact, CSPM-Partial optimizes the performance by generating several new leaf values during each iteration. However, the more leaf values that CT_L contains, the more the frequency of gain pairs calculations for each iteration. Therefore, the runtime is increased.

Overall, the optimized algorithm CSPM-Partial achieves high performance in terms of runtime especially for large datasets.

(2) **Gain update ratio.** The second experiment is performed to compare the relative amount of gain calculations achieved by the two CSPM versions. Fig. 5 shows the gain update ratio

of the two algorithms for each dataset from the first to the last iteration. Recall that, during an iteration and after merging two leafsets, merge candidates are updated. Moreover, the gain update ratio represents the ratio of gain values that are added or updated out of the total number of possible calculations in each iteration.

It can be observed that, CSPM-Partial often obtains a better gain update ratio compared with CSPM-Basic. This is why CSPM-Partial performs well on the used datasets such as DBLP and Pokec.

B. Pattern Analysis

To evaluate the interestingness of patterns discovered by the proposed CSPM algorithm, a manual inspection of the discovered patterns was performed. Besides, patterns with small code lengths are more frequent and have the strongest relationship between their core and leaf values.

(1) **DBLP.** Two patterns mined in DBLP and DBLP-Trend are depicted in Fig. 6(a). It was found that, there are many pattern occurrences where a researcher published a paper in *ICDM* and *EDBT* during a year, and his/her co-authors published in *PODS*, *ICDM* and *EDBT* during the same year. This is reasonable as *PODS*, *ICDM*, and *EDBT* belong to the same research area (data-mining), and generally, co-authors of a paper are interested in the same research areas. The pattern of Fig. 6(b) from DBLP-Trend can be interpreted in a similar way. But a difference is that patterns found in DBLP-Trend contain information about trends.

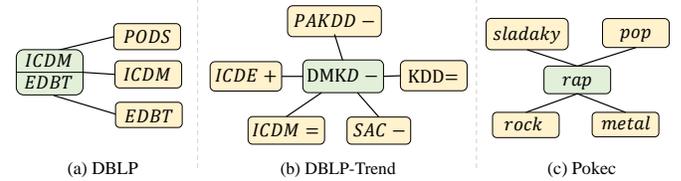


Fig. 6: Example a-stars from DBLP, DBLP-Trend and Pokec

(2) **USFlight.** An example pattern from the USFlight dataset is: ($\{NbDepart-\}$, $\{NbDepart+, DelayArriv-\}$), which means that, if the number of departure flights is reduced in an airport, there is a high chance that some connected airports will have an increase in their number of departure flights and a decrease in their number of delayed flights.

(3) **Pokec.** Pokec dataset contains the music preferences of people from different communities. Some interesting patterns were also discovered in this dataset using CSPM such as: ($\{rap\}$, $\{rock, metal, pop, sladaky\}$). This pattern has a short encoding length and a high frequency which means that it greatly contributes to the reduction of the conditional entropy. This pattern seems reasonable because some music types such as rock, metal, pop, sladaky, and rap are preferred by young people. Another pattern is found and it shows the music preferences of older people: ($\{disko\}$, $\{oldies, disko\}$).

C. Node Attribute Completion

In this section, we quantitatively assess the patterns to check whether CSPM can reveal underlying characteristics of

the data. More precisely, we apply CSPM to a graph-related classification task, namely node attribute completion [8]. Attribute-missing graph completion is useful for numerous real-world applications. Thus, various algorithms have been proposed [8], [37], [38], [39] to deal with this task. Intuitively, if attribute-stars extracted by CSPM can summarize well the data characteristics, incorporating mined patterns into the existing algorithms should improve or at least not degrade their performance.

We first present a scoring scheme which enables CSPM for the node completion task. It is based on the idea that a missing attribute value can be treated as a coreset with a core value and those of the neighbor nodes as leaf-values in an a-star. Then, a core value in an a-star with a smaller code length is more likely to be an attribute value of the targeted node. Therefore, we can score all possible attribute values using the code tables produced by CSPM.

The scoring module is summarized in Algorithm 5. In brief, given the a-star set M of CSPM and the attributes of the neighbors, the algorithm will return the scores of all possible attribute values for a targeted node v . In general, it is difficult to perfectly match the leaf-values in the leafset of an a-star with the attribute values of a neighbor. Thus, a weight is introduced to evaluate the similarity between leaf values in the leafset of an a-star and that of the neighbors of v . Intuitively, a leafset that is not similar to a set of neighbors will have a large weight w . As a result, it has a small score cl which indicates that the corresponding core-values are less likely to be an attribute value of v .

Algorithm 5: The scoring module of CSPM

input : An attributed graph G with n attribute values, the model M , a node v with missing attribute values
output: Scores for all possible attribute values

```

1  $scores \leftarrow [-\infty]_{n \times 1}$ ; // Set scores to  $-\infty$ 
2  $neighbors \leftarrow neighbor\_attributes(v)$ ;
3 foreach a-star  $S$  in  $M$  do
4    $S_{code} \leftarrow code\_length(S)$ ; // Equation (4)
5    $w \leftarrow similarity(S_L, neighbors)$ ; //  $S_L$ : leafset of  $S$ 
6    $cl \leftarrow -w \times S_{code}$ ;
7   if  $cl > scores[S_c]$  then
8      $scores[S_c] \leftarrow cl$ ; //  $S_c$ : coreset of  $S$ 
9   end
10 end
11 return  $scores$ 

```

Based on the fact that CSPM is designed to learn the underlying correlations between attribute values, it is suitable to assist other attribute completion models. This is done by integrating CSPM with attribute scoring into existing node attribute completion models. The prediction process is illustrated in Fig. 7. Besides, the output of a graph completion model for each vertex is a vector containing values indicating the possibilities that the attribute values appear while the vector obtained by the CSPM model records the scores of each attribute. In the proposed approach, the two vectors are normalized separately and then multiplied to get final scores.

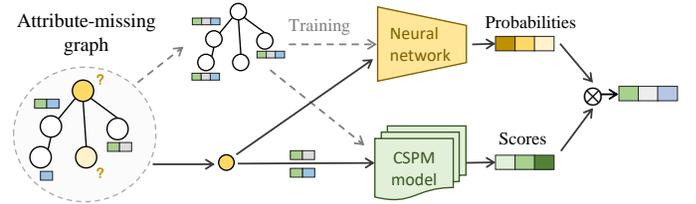


Fig. 7: The whole process of the attribute completion task using CSPM

Experiments are conducted on three benchmark citation networks in which attributes are categorical, namely Cora [40], Citeseer [41], and DBLP [34]. Similarly to [8], the performance is evaluated according to the Recall@K and NDCG@K [42], where larger values indicate better results. Specifically, the Recall reflects the correctness of the predicted attributes and the NDCG assesses ranking quality.

Note that DBLP is evaluated with a smaller K as it has less attribute values per node. The results are summarized in TABLE IV. It can be seen that, by integrating the a-stars-based scoring module into the attribute completion process, all the baseline algorithms are improved with different degrees. The most significant improvement is for NeighAggre or VAE which have relatively poor performance. These observations confirm that CSPM can successfully summarize the underlying characteristics of the given data.

D. Alarm correlation analysis

Telecommunication networks are cornerstones of the communication systems in modern society. Ensuring high-quality services in complex and large telecom networks is important and challenging since millions of faults and alarms are triggered across the devices everyday. Thus, discovering relationships between these alarms and filtering trivial alarms from important ones is critical to locate faults.

Recently, several studies have applied pattern mining techniques for alarm correlation analysis. Wang et al. [45] designed and deployed an alarm management system called AABD (Automatic Alarm Behavior Discovery) where the alarm correlation module mines frequent patterns in alarm logs via sequential pattern mining [46], [47]. The discovered patterns are then used to generate rules indicating that an alarm may be caused by another alarm. These rules are used to perform alarm compression (reduce the number of alarms presented to maintenance workers). For instance, (Low_signal, {Link_degrader, Microwave_stripping}) is an alarm rule, suggesting that the cause alarm Low_signal may be triggering derivative alarms (i.e., Link_degrader and Microwave_stripping). The alarm compression is achieved by only showing Low_signal to the maintenance workers when they appear simultaneously.

As AABD utilizes pattern mining algorithms which are not parameter-free, the quality of results is sensitive to the parameter values. Moreover, AABD ignores the connections between network devices and the alarm importance in a mined rule (i.e., cause alarm or not) have to be inferred by the experts knowledge. To solve the above problems, the ACOR

TABLE IV: Profiling evaluation for node attribute completion

	Metric	Recall@10	Recall@20	Recall@50	NDCG@10	NDCG@20	NDCG@50
	Method						
Cora	NeighAggre [39]	0.0895	0.1396	0.1944	0.1203	0.1534	0.1832
	CSPM+NeighAggre	0.1339	0.1842	0.3002	0.1856	0.2190	0.2796
	VAE [43]	0.0875	0.1215	0.2069	0.1219	0.1451	0.1901
	CSPM+VAE	0.1023	0.1512	0.2471	0.1429	0.1758	0.2266
	GCN [12]	0.1261	0.1782	0.2930	0.1727	0.2075	0.2680
	CSPM+GCN	0.1272	0.1843	0.3029	0.1749	0.2131	0.2754
	GAT [13]	0.1281	0.1811	0.2982	0.1747	0.2108	0.2722
	CSPM+GAT	0.1302	0.1862	0.3070	0.1780	0.2161	0.2794
	GraphSage [44]	0.1224	0.1711	0.2790	0.1689	0.2018	0.2586
	CSPM+GraphSage	0.1264	0.1769	0.2918	0.1738	0.2075	0.2680
	SAT [8]	0.1461	0.2132	0.3375	0.2069	0.2514	0.3177
	CSPM+SAT	0.1539	0.2184	0.3465	0.2131	0.2569	0.3244
Avg.improvement(%)	+12.94	+11.41	+14.57	+13.43	+12.56	+13.83	
Citeseer	NeighAggre [39]	0.0410	0.0737	0.1242	0.0658	0.0932	0.1274
	CSPM+NeighAggre	0.0574	0.0936	0.1742	0.0952	0.1255	0.1784
	VAE [43]	0.0369	0.0662	0.1292	0.0580	0.0825	0.1288
	CSPM+VAE	0.0438	0.0766	0.1480	0.0711	0.0985	0.1452
	GCN [12]	0.0613	0.1060	0.2000	0.1000	0.1370	0.1990
	CSPM+GCN	0.0628	0.1076	0.2039	0.1027	0.1402	0.2031
	GAT [13]	0.0535	0.0988	0.1958	0.0860	0.1238	0.1871
	CSPM+GAT	0.0585	0.1039	0.1979	0.0943	0.1322	0.1935
	GraphSage [44]	0.0561	0.1008	0.1939	0.0879	0.1250	0.1858
	CSPM+GraphSage	0.0579	0.1038	0.1971	0.0915	0.1297	0.1905
	SAT [8]	0.0657	0.1125	0.2103	0.1110	0.1501	0.2142
	CSPM+SAT	0.0684	0.1164	0.2141	0.1151	0.1552	0.2192
Avg.improvement(%)	+12.97	+9.30	+10.21	+14.57	+11.72	+10.52	
	Metric	Recall@3	Recall@5	Recall@10	NDCG@3	NDCG@5	NDCG@10
	Method						
DBLP	NeighAggre [39]	0.2022	0.2648	0.3755	0.2583	0.3000	0.3631
	CSPM+NeighAggre	0.3685	0.4646	0.6028	0.4302	0.4963	0.5709
	VAE [43]	0.2511	0.3119	0.4728	0.3071	0.3494	0.4370
	CSPM+VAE	0.3066	0.3801	0.5276	0.3755	0.4266	0.5050
	GCN [12]	0.3994	0.4832	0.6657	0.4943	0.5523	0.6510
	CSPM+GCN	0.5139	0.6400	0.7938	0.5916	0.6760	0.7582
	GAT [13]	0.3964	0.4910	0.6574	0.4902	0.5539	0.6419
	CSPM+GAT	0.5253	0.6634	0.8330	0.6018	0.6935	0.7814
	GraphSage [44]	0.3445	0.4258	0.6069	0.4227	0.4793	0.5760
	CSPM+GraphSage	0.4020	0.4840	0.6460	0.4790	0.5345	0.6195
	SAT [8]	0.4890	0.6418	0.8227	0.5708	0.6694	0.7629
	CSPM+SAT	0.4981	0.6445	0.8249	0.5780	0.6727	0.7664
Avg.improvement(%)	+30.68	+29.83	+20.80	+24.31	+24.52	+19.83	

algorithm [9] models alarm data as a dynamic attributed graph, and then extracts paired alarm having a high correlation by a tailored correlation measure which also give the importance of each alarm in an alarm pair. Though ACOR can generate alarm rules by combining alarm pairs with the same cause alarm, information loss is inevitable.

In this section, we evaluate the efficiency of CSPM for alarm correlation analysis by comparing with ACOR [9], checking whether the rules in the AABD rule library can be rediscovered with higher rankings. For this purpose, the coverage ratio [9] is used as an evaluation metric. It is defined as follows: $coverage = |A \cap B| / |A|$, where A is the set of valid rules and B is the rules found by CSPM or ACOR. A high coverage ratio indicates that most of rules discovered by the proposed framework are valid. Hereafter, the coverage ratio for top-K represents the ratio of valid rules in the k rules that are found by CSPM or ACOR and have the highest correlation scores. Note that, the alarm patterns extracted by CSPM are star-shaped where the core serves as the cause alarms, while the alarm patterns extracted by ACOR have a paired format. Hence, a-stars mined by CSPM are split into pairs. Note that the results are not influenced because the rankings and scores of all alarm rules are maintained. It is only done to compare with ACOR, which mines pairwise alarm rules.

The experiments were done on a real alarm dataset with 6 million triggered alarms collected from a metropolitan city of Southeast Asia from the 12th to 16th April, 2019. The collected alarms are categorized into 300 types, which fall into 11 rules stored in the AABD system [45]. These rules are then decomposed into 121 pair-rules [9]. The coverage ratios of the two algorithms are given in Fig. 8. It can be found that, the coverage ratio increases as more rules are selected, and finally all the valid rules are found. Moreover, as it is expected, the valid rules found by CSPM are ranked higher comparing with the valid rules found by ACOR. This is due to the fact that ACOR evaluates each pair-rule separately. However, the proposed algorithm simultaneously and systematically takes all rules into account via the MDL.

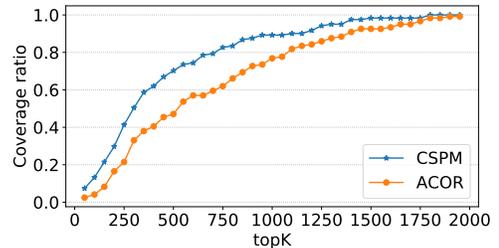


Fig. 8: The coverage ratio of ACOR and CSPM for alarm correlation analysis

VII. CONCLUSION

To find patterns that capture strong relationships between attributes in attributed graphs, this paper proposed a parameter-free and compression-based algorithm, named CSPM. It finds star-shaped attribute patterns that best compress the data according to the conditional entropy and MDL principle. Experiments on large graphs have shown that CSPM can identify interesting patterns and that its optimization improves its performance. Besides, CSPM was shown to boost the performance of state-of-the-art graph attribute completion models and find interesting patterns for alarm correlation analysis.

Some interesting possibilities for future work are: (1) to utilize a-stars found by CSPM for other graph-related learning problems such as graph classification, (2) to extend CSPM to handle other graph types such as dynamic attributed graphs and graphs with multiple attributes per edge. (3) to develop a distributed version of CSPM to process very large databases.

REFERENCES

- [1] P. Fournier-Viger, G. He, C. Cheng, J. Li, M. Zhou, J. C.-W. Lin, U. Yun, A survey of pattern mining in dynamic graphs, *WIREs Data Mining and Knowledge Discovery* (2020).
- [2] C. Jiang, F. Coenen, M. Zito, A survey of frequent subgraph mining algorithms, *Knowledge Engineering Review* 28 (2013) 75–105.
- [3] M. Yang, Z. Meng, I. King, FeatureNorm: L2 feature normalization for dynamic graph embedding, in: *2020 IEEE Int. Conf. on Data Mining (ICDM)*, IEEE, 2020, pp. 731–740.
- [4] M. Yang, M. Zhou, M. Kalander, Z. Huang, I. King, Discrete-time temporal network embedding via implicit hierarchical learning in hyperbolic space, in: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1975–1985.
- [5] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. T. Chi, J. Zhang, B. Le, A survey of itemset mining, *WIREs Data Mining and Knowledge Discovery* (2017).
- [6] X. Yan, J. Han, gspan: Graph-based substructure pattern mining, in: *Proc. 2nd IEEE Int. Conf. on Data Mining*, 2002.
- [7] R. Y. Dougnon, P. Fournier-Viger, J. C.-W. Lin, R. Nkambou, Inferring social network user profiles using a partial social graph, *Journal of intelligent information systems* 47 (2) (2016) 313–344.
- [8] X. Chen, S. Chen, J. Yao, H. Zheng, Y. Zhang, I. W. Tsang, Learning on attribute-missing graphs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
- [9] P. Fournier-Viger, H. Ganghuan, M. Zhou, M. Nouioua, J. Liu, Discovering alarm correlation rules for network fault management, 2020.
- [10] P. Fournier-Viger, C. Cheng, J. C.-W. Lin, U. Yun, R. U. Kiran, Tkg: Efficient mining of top-k frequent subgraphs, in: *Proc. 7th Int. Conf. on Big Data Analytics*, Springer, 2019, pp. 209–226.
- [11] X. Yan, J. Han, gspan: graph-based substructure pattern mining, in: *Proc. 2nd IEEE Int. Conf. on Data Mining*, 2002, pp. 721–724.
- [12] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2017.
- [13] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: *Proc. 6th Int. Conf. on Learning Representations*, 2018.
- [14] C. Pasquier, J. Sanhes, F. Flouvat, N. Selmaoui-Folcher, Frequent pattern mining in attributed trees: algorithms and applications, *Knowledge and Information Systems* 46 (3) (2016) 491–514.
- [15] A. Jiménez, F. Berzal, J.-C. Cubero, Frequent tree pattern mining: A survey, *Intelligent Data Analysis* 14 (6) (2010) 603–622.
- [16] M. Atzmueller, H. Soldano, G. Santini, D. Bouthinon, MinerLSD: Efficient local pattern mining on attributed graphs, *Applied Network Science* (2018) 219–228.
- [17] W. Luo, H. Tan, L. Chen, L. M. Ni, Finding time period-based most frequent path in big trajectory data, in: *Proceedings of the 2013 ACM SIGMOD Int. Conf. on management of data*, 2013, pp. 713–724.
- [18] M. Kargar, M. Zihayat, J. Szlichta, Mining and exploration of attributed graphs: theory and applications, in: *Proceedings of the 29th Annual Int. Conf. on Computer Science and Software Engineering*, 2019, pp. 397–398.
- [19] P. Fournier-Viger, G. He, J. C.-W. Lin, H. M. Gomes, Mining attribute evolution rules in dynamic attributed graphs, in: *Proc. 22nd Int. Conf. on Data Warehousing and Knowledge Discovery*, Springer, 2020, pp. 167–182.
- [20] J. Vreeken, M. Van Leeuwen, A. Siebes, Krimp: mining itemsets that compress, *Data Mining and Knowledge Discovery* 23 (1) (2011) 169–214.
- [21] J. Rissanen, Modeling by shortest data description, *Automatica* 14 (5) (1978) 465–471.
- [22] H. T. Lam, F. Mörchen, D. Fradkin, T. Calders, Mining compressing sequential patterns, *Statistical Analysis and Data Mining: The ASA Data Science Journal* 7 (1) (2014) 34–52.
- [23] R. Bertens, J. Vreeken, A. Siebes, Keeping it short and simple: Summarising complex event sequences with multivariate patterns, in: *Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 735–744.
- [24] F. Bariatti, P. Cellier, S. Ferré, GraphMDL: Graph pattern selection based on minimum description length, in: *Proc. 18th Int. Symposium on Intelligent Data Analysis*, Springer, 2020, pp. 54–66.
- [25] K. Smets, J. Vreeken, Slim: Directly mining descriptive patterns., in: *Proc. 12th SIAM Int. Conf. on Data Mining*, 2012, pp. 236–247.
- [26] Y. Lim, U. Kang, C. Faloutsos, Slashburn: Graph compression and mining beyond caveman communities, *IEEE Transactions on Knowledge and Data Engineering* 26 (12) (2014) 3077–3089.
- [27] D. Koutra, U. Kang, J. Vreeken, C. Faloutsos, Vog: Summarizing and understanding large graphs (2014) 91–99.
- [28] A. Ahmed, Z. R. Hassan, M. Shabbir, Interpretable multi-scale graph descriptors via structural compression, *Information Sciences* (2020).
- [29] Y. Liu, T. Safavi, A. Dighe, D. Koutra, Graph summarization methods and applications: A survey, *ACM Computing Surveys* 51 (3) (2018) 62.
- [30] K. U. Khan, W. Nawaz, Y. Lee, Set-based unified approach for attributed graph summarization, in: *Proc. 4th Int. Conf. on Big Data and Cloud Computing*, 2014, pp. 378–385.
- [31] S. E. Amiri, L. Chen, B. A. Prakash, Efficiently summarizing attributed diffusion networks, *Data Mining and Knowledge Discovery* 32 (5) (2018) 1251–1274.
- [32] Y.-A. Shih, B. Chang, J. Y. Chin, Data-driven student homophily pattern analysis of online discussion in a social network learning environment, *Journal of Computers in Education* (2020) 1–22.
- [33] T. M. Cover, *Elements of information theory*, John Wiley & Sons, 1999.
- [34] E. Desmier, M. Plantevit, C. Robardet, J.-F. Boulicaut, Cohesive evolution patterns in dynamic attributed graphs, in: *Proc. 15th Int. Conf. of Discovery Science*, Springer, 2012, pp. 110–124.
- [35] P. Fournier-Viger, C. Cheng, Z. Cheng, J. C.-W. Lin, N. Selmaoui-Folcher, Mining significant trend sequences in dynamic attributed graphs, *Knowledge-Based Systems* 182 (2019) 104797.
- [36] M. Kaytoue, Y. Pitarch, M. Plantevit, C. Robardet, Triggering patterns of topology changes in dynamic graphs, in: *Proc. of 6th Int. Conf. on Advances in Social Networks Analysis and Mining*, 2014, pp. 158–165.
- [37] F. Monti, M. M. Bronstein, X. Bresson, Geometric matrix completion with recurrent multi-graph neural networks, in: *Proc. 30: Annual Conference on Neural Information Processing Systems*, 2017, pp. 3697–3707.
- [38] R. v. d. Berg, T. N. Kipf, M. Welling, Graph convolutional matrix completion, in: *Proc. KDD18 Deep Learning Day*, 2018.
- [39] Ö. Şimşek, D. Jensen, Navigating networks by using homophily and degree, *Proceedings of the National Academy of Sciences* 105 (35) (2008) 12758–12762.
- [40] A. K. McCallum, K. Nigam, J. Rennie, K. Seymore, Automating the construction of internet portals with machine learning, *Information Retrieval* 3 (2) (2000) 127–163.
- [41] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, T. Eliassi-Rad, Collective classification in network data, *AI magazine* 29 (3) (2008) 93–93.
- [42] C. Distinguishability, A theoretical analysis of normalized discounted cumulative gain (ndcg) ranking measures (2013).
- [43] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: *Proc. 2nd Int. Conf. on Learning Representations*, 2014.
- [44] W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in: *Proc. 31st Annual Conference on Neural Information Processing Systems*, 2017, pp. 1024–1034.

- [45] J. Wang, C. He, Y. Liu, G. Tian, I. Peng, J. Xing, X. Ruan, H. Xie, F. L. Wang, Efficient alarm behavior analytics for telecom networks, *Information Sciences* 402 (2017) 1–14.
- [46] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu, Mining sequential patterns by pattern-growth: The prefixspan approach, *IEEE Transactions on knowledge and data engineering* 16 (11) (2004) 1424–1440.
- [47] T. Truong, H. Duong, B. Le, P. Fournier-Viger, Fmaxclohusm: An efficient algorithm for mining frequent closed and maximal high utility sequences, *Engineering Applications of Artificial Intelligence* 85 (2019) 1–20.