Data Mining and Knowledge Discovery, 2005

DOI: 10.1007/s10618-005-0003-9

Finding Frequent Patterns in a Large Sparse Graph*

MICHIHIRO KURAMOCHI kuram@cs.umn.edu
GEORGE KARYPIS karypis@cs.umn.edu
Department of Computer Science & Engineering,University of Minnesota, Minneapolis, MN 55455, USA

Abstract. Graph-based modeling has emerged as a powerful abstraction capable of capturing in a single and unified framework many of the relational, spatial, topological, and other characteristics that are present in a variety of datasets and application areas. Computationally efficient algorithms that find patterns corresponding to frequently occurring subgraphs play an important role in developing data mining-driven methodologies for analyzing the graphs resulting from such datasets. This paper presents two algorithms, based on the horizontal and vertical pattern discovery paradigms, that find the connected subgraphs that have a sufficient number of edge-disjoint embeddings in a single large undirected labeled sparse graph. These algorithms use three different methods for determining the number of edge-disjoint embeddings of a subgraph and employ novel algorithms for candidate generation and frequency counting, which allow them to operate on datasets with different characteristics and to quickly prune unpromising subgraphs. Experimental evaluation on real datasets from various domains show that both algorithms achieve good performance, scale well to sparse input graphs with more than 120,000 vertices or 110,000 edges, and significantly outperform previously developed algorithms.

Keywords: pattern discovery, frequent subgraph, graph mining

1. Introduction

In recent years, there has been an increased interest in developing data mining algorithms that operate on graphs. Such graphs arise naturally in a number of different application domains including network intrusion (Lee and Stolfo, 2000; Ko, 2000), semantic web (Berendt et al., 2002), behavioral modeling (Wasserman et al., 1994; Mooney et al., 2004), VLSI reverse engineering (Yoshida and Motoda, 1995), link analysis (Jensen and Goldberg, 1998; Kleinberg et al., 1999; Kleinberg, 1999; Palmer et al., 2002), and chemical compound classification (Dehaspe et al., 1998; Kramer et al., 2001; Gonzalez et al., 2001; Deshpande et al., 2003). Moreover, they can be used to effectively model the structural and relational characteristics of a variety of datasets arising in other areas such as physical sciences (e.g., chemistry, fluid dynamics, astronomy, structural mechanics, and ecosystem modeling), life sciences (e.g., genomics, proteomics, pharmacogenomics, and health informatics), and home-land defense (e.g., information assurance, network intrusion, infrastructure protection, and terrorist-threat prediction/identification).

*This work was supported in part by NSF CCR-9972519, EIA-9986042, ACI-9982274, ACI-0133464, and ACI-0312828; the Digital Technology Center at the University of Minnesota; and by the Army High Performance Computing Research Center (AHPCRC) under the auspices of the Department of the Army, Army Research Laboratory (ARL) under Cooperative Agreement number DAAD19-01-2-0014. The content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute.

The focus of this paper is on developing algorithms for a particular data mining task, which is that of finding frequently occurring patterns in graph datasets. Frequent patterns play a critical role in many data mining tasks as they can be used among other to derive association rules (Agrawal and Srikant, 1994), act as composite features for classification algorithms (Dehaspe et al., 1998; Muggleton, 1999; Srinivasan et al., 1997; Liu et al., 1998; Gonzalez et al., 2001; Li et al., 2001; Deshpande et al., 2002), cluster the (graph) transactions (Agrawal et al., 1998; Leibowitz et al., 1999; Jonyer et al., 2001a, 2001; Leibowitz et al., 2001; Guralnik and Karypis, 2001), and help in determining the similarity between graphs (Mitchell et al., 1989; Grindley et al., 1993; Koch et al., 1996; Pennec and Ayache, 1998; Chew et al., 1999; Leibowitz et al., 2001; De Raedt and Kramer, 2001; Raymond, 2002; Wang et al., 2002). Within the context of graphs, the most widely used definition of a pattern is that of a connected subgraph (Borgelt and Berthold, 2002; Yan and Han, 2002, 2003; Inokuchi et al., 2003; Hong et al., 2003; Huan et al., 2003; Kuramochi and Karypis, 2004a) and is the definition that we will use in this paper. However, different pattern definitions have been proposed as well (Inokuchi et al., 2003).

There are two distinct problem formulations for frequent pattern mining in graph datasets that are referred to as the *graph-transaction setting* and the *single-graph setting*. In the graph-transaction setting, the input to the pattern mining algorithm is a set of relatively small graphs (called transactions), whereas in the single-graph setting the input data is a single large graph. The difference affects the way the frequency of the various patterns is determined. For the graph-transaction setting, the frequency of a pattern is determined by the number of graph transactions that the pattern occurs in, irrespective of how many times a pattern occurs in a particular transaction, whereas in the single-graph setting, the frequency of a pattern is based on the number of its occurrences (i.e., embeddings) in the single graph. Due to the inherent differences of the characteristics of the underlying dataset and the problem formulation, algorithms developed for the graph-transaction setting cannot be used to solve the single-graph setting, whereas the latter algorithms can be easily adapted to solve the former problem.

In recent years, a number of efficient and scalable algorithms have been developed to find patterns in the graph-transaction setting (Borgelt and Berthold, 2002; Yan and Han, 2002; Inokuchi et al., 2003; Hong et al., 2003; Yan and Han, 2003; Huan et al., 2003; Kuramochi and Karypis, 2004a). These algorithms are complete in the sense that they are guaranteed to discover all frequent subgraphs and were shown to scale to very large graph datasets. However, algorithms that are capable of finding patterns in the single-graph setting has received much less attention, despite the fact that this problem setting is more generic and applicable to a wider range of datasets and application domains than the other. Moreover, existing algorithms that are guaranteed to find all frequent patterns (Ghazizadeh and Chawathe, 2002b; Vanetik et al., 2002) or algorithms that are heuristic, such as GBI (Yoshida et al., 1994) and SUBDUE (Holder et al., 1994), which tend to miss a large number of frequent patterns, are computationally expensive and do not scale to large datasets.

Developing algorithms that find the complete set of frequent patterns in the single-graph setting is the focus of this paper. We present two computationally efficient algorithms that can find subgraphs which are frequently embedded within a large sparse graph. The first algorithm, called HSIGRAM, follows a *horizontal approach* and finds the frequent

subgraphs in a breadth-first fashion, whereas the second algorithm, called VSIGRAM, follows a *vertical approach* and finds the frequent subgraphs in a depth-first fashion. These algorithms incorporate efficient algorithms for candidate generation and frequency counting that allow them to scale to graphs containing over 120,000 vertices and find patterns with relatively low occurrence frequency. Our experimental evaluation on eight real graphs shows that both HSIGRAM and VSIGRAM achieve reasonably good performance, scale to large graphs, and substantially outperform previously developed approaches for solving similar or simpler versions of the problem. A shorter version of this paper has previously appeared in Kuramochi and Karypis (2004b). This paper is enhanced with various modifications including a section on recent related work and new experimental evaluations.

The rest of this paper is organized as follows. Section 2 defines the graph model that we use, reviews some graph-related definitions, and introduces the notation that is used in the paper. Section 3 surveys related research in this area. Section 4 formally defines the problem of frequent subgraph discovery and discusses the challenges associated with finding them in a computationally efficient manner. Section 5 describes in detail the HSIGRAM and VSIGRAM algorithms that we developed for solving the problem of frequent subgraph discovery from a single large sparse graph. Section 6 provides a detailed experimental evaluation of the HSIGRAM and VSIGRAM algorithms on various real datasets and compares them against existing algorithms. Finally, Section 7 provides some concluding remarks.

2. Definitions and notation

A **graph** $G = (V, E)$ is made of two sets, the set of vertices V and the set of edges E. Each edge itself is a pair of vertices, and throughout this paper we assume that the graph is undirected, i.e., each edge is an unordered pair of vertices. Furthermore, we will assume that the graph is **labeled**. That is, each vertex and edge has a label associated with it that is drawn from a predefined set of vertex labels (L_V) and edge labels (L_E). Each vertex (or edge) of the graph is not required to have a unique label and the same label can be assigned to many vertices (or edges) in the same graph. If all the vertices and edges of the graph have the same vertex and edge label assigned to them, we will call this graph **unlabeled**.

Given a graph $G = (V, E)$, a graph $G_s = (V_s, E_s)$ is a *subgraph* of G if and only if $V_s \subseteq V$ and $E_s \subseteq E$. A graph is **connected** if there is a path between every pair of vertices in the graph. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are **isomorphic** if they are topologically identical to each other, that is, there is a mapping from V_1 to V_2 such that each edge in E_1 is mapped to a single edge in E_2 and vice versa. In the case of labeled graphs, this mapping must also preserve the labels on the vertices and edges. An **automorphism** is an isomorphism mapping where $G_1 = G_2$. Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, the problem of **subgraph isomorphism** is to find an isomorphism between G_2 and a subgraph of G_1, i.e., determine whether or not G_2 is included in G_1.

Given a subgraph G_s and a graph $\mathcal{G}$, two embeddings of G_s in $\mathcal{G}$ are called **identical** if they use the same set of edges of $\mathcal{G}$, and they are called **edge-disjoint** if they do not have any edges of $\mathcal{G}$ in common. Given a set of all embeddings of a particular subgraph G_s in a graph $\mathcal{G}$, the **overlap graph** of G_s is a graph obtained by creating a vertex for each

Table 1. Notation used throughout the paper.

Notation	Description
k-subgraph	A connected subgraph with k edges (also written as a size-k subgraph)
G^k,H^k	Graphs of size k
$E(G)$	Edges of a graph G
$V(G)$	Vertices of a graph G
cl(G)	Canonical label of a graph G
dia(G)	Diameter of a graph G
a, b, c, e, f	Edges
u, v	Vertices
d(v)	Degree of a vertex v
$l(v)$	Label of a vertex v
$l(e)$	Label of an edge e
$H = G - e$	A graph obtained by deleting edge $e \in E(G)$
$\mathcal{G}$	Input graph
$\mathcal{G}_i$	$\mathcal{G}$'s connected component
$\mathcal{S}(G^{k+1})$	Set of all connected size-k subgraphs of G^{k+1}
$\mathcal{M}(G) = \{m_i\}$	All embeddings of a subgraph G in $\mathcal{G}$
$\mathcal{A}(G) = \{e_i\}$	All anchor edges of a subgraph G in $\mathcal{G}$
C	Candidate subgraph
$\mathcal{C}^{\{k\}}$	Set of candidates with k edges
$\mathcal{C}$	Set of all candidates
F	Frequent subgraph
$\mathcal{F}^k$	Set of frequent k-subgraphs
$\mathcal{F}$	Set of all frequent subgraphs
$k*$	Size of the largest frequent subgraph in $\mathcal{G}$
L_E	Set of all edge labels in $\mathcal{G}$
L_V	Set of all vertex labels in $\mathcal{G}$

non-identical embedding and creating an edge for each pair of non-edge-disjoint embeddings. An example of a subgraph and its overlap graph are shown in figure 1.

The notation that we will be using throughout the paper is shown in Table 1.

2.1. Canonical labeling

One of the key operations required by any frequent subgraph discovery algorithm is a mechanism by which to check whether two subgraphs are identical or not. One way of performing this check is to perform a graph isomorphism operation. However, in cases in which many such checks are required among the same set of subgraphs, a better way of

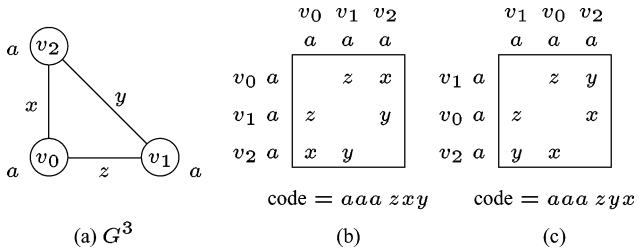


Figure 1 Simple examples of codes and canonical adjacency matrices.

performing this task is to assign to each graph a unique *code* (i.e., a sequence of bits, a string, or a sequence of numbers) that is invariant on the ordering of the vertices and edges in the graph. Such a code is referred to as the **canonical label** of a graph $G = (V, E)$ (Read and Corneil, 1977; Fortin, 1996), and we will denote it by cl(G). By using canonical labels, we can check whether or not two graphs are identical by checking to see whether they have identical canonical labels. Moreover, by comparing the canonical labels we can obtain a complete ordering of a set of graphs in a unique and deterministic way, regardless of the original vertex and edge ordering.

A simple way of defining the canonical label of a graph is as the string obtained by concatenating the upper-triangular entries of the graph's adjacency matrix when this matrix has been symmetrically permuted so that this string becomes the lexicographically largest (or smallest) over the strings that can be obtained from all such permutations. This is illustrated in figure 1 that shows a graph G^3 and the permutation of its adjacency matrix[1] that leads to its canonical label "*aaazyx*". In this code, "*aaa*" was obtained by concatenating the vertex-labels in the order that they appear in the adjacency matrix and "*zyx*" was obtained by concatenating the columns of the upper-triangular portion of the matrix. Note that any other permutation of G^3's adjacency matrix will lead to a code that is lexicographically smaller (or equal) to "*aaazyx*". If a graph has $|V|$ vertices, the complexity of determining its canonical label using this scheme is in $O(|V|!)$ making it impractical even for moderate size graphs. Note that the problem of determining the canonical label of a graph is equivalent to determining isomorphism between graphs, because if two graphs are isomorphic with each other, their canonical labels must be identical. Both canonical labeling and determining graph isomorphism are not known to be either in P or in NP-complete (Fortin, 1996).

In practice, the complexity of finding a canonical labeling of a graph can be reduced by using various heuristics to narrow down the search space or by using alternate canonical label definitions that take advantage of special properties that may exist in a particular set of graphs (McKay, 1981, n.d.; Fortin, 1996). As part of our earlier research we have developed such canonical labeling algorithm that fully makes use of edge- and vertex-labels for fast processing and various vertex invariants to reduce the complexity of determining the canonical label of a graph (Kuramochi and Karypis, 2001, 2002). Our algorithm can compute the canonical label of graphs containing up to 50 vertices extremely fast and will be the algorithm used to compute the canonical labels of the different subgraphs in this paper.

2.2. *Maximum independent set*

As discussed later in Section 4, our frequent subgraph discovery algorithm focuses on finding subgraphs whose embeddings are edge-disjoint. A critical step in obtaining this set of edge-disjoint embeddings for a particular subgraph is to find the maximum independent set of its overlap graph. Given a graph $G = (V, E)$, a subset of vertices $I \subset V$ is called **independent** if no two vertices in I are connected by an edge in E. An independent set I is called **maximal independent set** for every vertex v in I if there is an edge in E that connects v to a vertex in $V \setminus I$. A maximal independent set I is called **maximum independent set** (MIS) if I contains as many vertices of V as possible.

The problem of finding the MIS of a graph was among the first problems proved to be in NP-complete (Garey and Johnson, 1979), and remains so even for bounded degree graphs. Moreover, it has been shown that the size of MIS cannot be approximated even within a factor of $|V|^{1-o(1)}$ in polynomial time (Feige et al., 1991). However, the importance of the problem and its applicability to a wide-range of domains has attracted a considerable amount of research. This research has been focused on developing both faster exact algorithms as well as approximate algorithms. The faster exact algorithm to date is the algorithm by Robson (1986) that solves the MIS problem in time $O(1.211^{|V|})$, making it possible to solve in reasonable amount of time problem instances containing up to around 100 vertices. In this study, we used a fast implementation of the exact **maximum clique** (MC) problem solver *wclique* (Östergård, 2002) instead of those fast exact MIS algorithms. Because the MIS problem on a graph G is equivalent to the MC problem on a G's complement graph $\bar{G}$, we can use *wclique* as a fast exact MIS algorithm (EMIS). Heuristic algorithms focus on finding maximal independent sets whose size is bounded in terms of the size of the optimal solution, and a number of such methods have been developed (Hochbaum, 1983; Berman and Fujito, 1995; Khanna et al., 1994; Halldórsson and Radhakrishnan, 1997).

One of the most widely used heuristic is the **greedy algorithm** (GMIS) which selects a vertex of the minimum degree, deletes that vertex and all of its neighbors from the graph, and repeats this process until the graph becomes empty. A recent detailed analysis of the GMIS algorithm has shown that it produces reasonably good approximations of the MIS for bounded- and low-degree graphs (Halldórsson and Radhakrishnan, 1997). In particular, for a graph G with a maximum degree Δ and an average degree $\bar{d}$, the size $|I|$ of the MIS satisfies the following:

$$|I| \leq \min\left(\frac{\Delta+2}{3}|\text{GMIS}(G)|, \frac{\bar{d}+2}{2}|\text{GMIS}(G)|\right) \tag{1}$$

where $|\text{GMIS}(G)|$ is the size of the approximate MIS found by the GMIS algorithm. Note that Eq. (1) provides an upper-bound on the number of edge-disjoint embeddings of a particular subgraph, and we will use this bound to obtain a computationally tractable problem formulation that is guaranteed not to miss any subgraphs that can potentially be frequent.

3. Related work

The previous research on finding frequent subgraphs in graph datasets falls under two categories. The first category contains algorithms for finding subgraphs that occur multiple times in a single input graph (Yoshida et al., 1994, 2000; Holder et al., 1994; Ghazizadeh and Chawathe, 2002b; Vanetik et al., 2002) and are directly related to the algorithms presented in this paper, whereas the second category contains algorithms that find subgraphs that occur frequently across a database of small graphs (Dehaspe et al., 1998; Inokuchi et al., 2000, 2002, 2003; Kramer et al., 2001; Kuramochi and Karypis, 2001, 2004a; Borgelt and Berthold, 2002; Yan and Han, 2002; Hong et al., 2003; Huan et al., 2003; Cohen and Gudes, 2004). Between these two classes of algorithms, those developed for the latter problem are in general more mature as they have moderate computational requirements and scale to large datasets. In the rest of this section, we will describe the related research only on the single-graph setting as it is directly related to the topic of the paper. The reader should refer to Kuramochi et al. (2004) for a survey of the different algorithms for the graph-transaction setting.

The most well-known algorithm for finding recurring subgraphs in a single large graph is the SUBDUE system, originally developed in 1994, and improved over the years (Holder et al., 1994; Cook and Holder, 1994, 2000; Cook et al., 1995, 2000). SUBDUE is an approximate algorithm and finds patterns that can compress the original input graph by substituting those patterns with a single vertex. In evaluating the extent to which a particular pattern can compress the original graph it uses the minimum description length (MDL) principle, and employs a heuristic beam search to narrow the search-space. These approximations improve its computational efficiency but at the same time it prevents it from finding subgraphs that are indeed frequent. GBI (Yoshida et al., 1994) is another greedy heuristics based algorithm similar to SUBDUE. Ghazizadeh and Chawathe (2002b) developed an algorithm called SEuS that uses a data structure called *summary* to construct a lossy compressed representation of the input graph. This summary is obtained by collapsing together all the vertices of the input graph that have the same label and is used to quickly prune infrequent candidates. As the authors indicate, this summary data-structure is useful only when the input graph contains a relatively small number of frequent subgraphs with high frequency, and is not effective if there are a large number of frequent subgraphs with low frequency. Vanetik et al. (2002) presented an algorithm for finding all frequently occurring subgraphs from a single labeled undirected graph using the maximum number of edge-disjoint embeddings of a graph as a measure of its frequency. Each subgraph is represented by its minimum number of edge-disjoint paths (*path number*) and use a level-by-level approach to grow the patterns based on their path-number. Their emphasis is on efficient candidate generation and no special attention is paid for frequency counting.

4. Discovering frequent patterns in a single graph: Problem definition

A fundamental issue that needs to be considered by any frequent subgraph discovery problem formulation that is applicable to the single-graph setting is the counting method of the occurrence frequency. In general, there are two possible methods for determining the

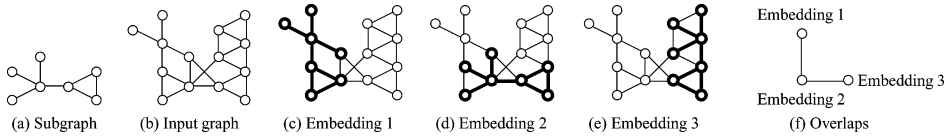


Figure 2 Overlapped embeddings.

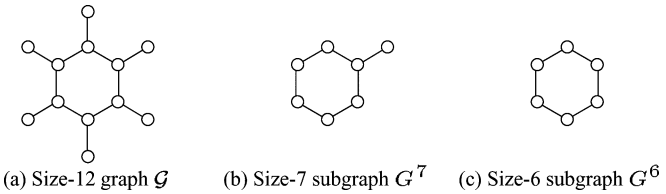


Figure 3 Patterns with the non-monotonic frequency.

frequency of a subgraph. In the first method, two embeddings of a subgraph are considered different, as long as they differ by at least one edge (i.e., non-identical). As a result, arbitrary overlaps of embeddings of the same subgraph are allowed. On the other hand, in the second method, two embeddings are considered different, only if they do not share edges (i.e., they are edge-disjoint). These two methods are illustrated in figure 2. In this example, there are three possible embeddings of the subgraph shown in figure 2(a) in the input graph of figure 2(b). Two of these embeddings (figures 2(c) and (e)) do not share any edges, whereas the third embedding (figure 2(d)) shares edges with the other two. Thus, if we allow overlaps, the frequency of the subgraph is 3, and if we do not it is 2.

These two ways of counting the frequency of a subgraph lead to problems with dramatically different characteristics. If we allow arbitrary overlaps between non-identical embeddings, then the resulting frequency is not any longer downward closed (i.e., the frequency of a subgraph does not monotonically decrease as a function of its length). This is illustrated in Figure 3. Both G^7 and G^6 are subgraphs of $\mathcal{G}$. Although the smaller subgraph G^6 has only one non-identical embedding, the larger G^7 has six non-identical embeddings. On the other hand, if we determine the frequency of each subgraph by counting the maximum number of its edge-disjoint embeddings, then the resulting frequency is downward closed (Vanetik et al., 2002).

Being able to take advantage of a frequency counting method that is downward closed is essential for the computational tractability of most frequent pattern discovery algorithms. For this reason, our problem formulations uses edge-disjoint embeddings. Given this, one way of formulating the frequent subgraph discovery problem for the single-graph setting as follows (Vanetik et al., 2002):

Definition 1 (Exact discovery). Given an input graph $\mathcal{G}$ which is undirected and labeled, and a parameter f, find *all* connected undirected labeled subgraphs that have at least f edge-disjoint embeddings in $\mathcal{G}$.

By this definition, in order to determine if a subgraph is frequent or not, we need to find whether the overlap graph of its non-identical embeddings contain an independent set whose size is at least f. When a subgraph is relatively frequent compared to the frequency threshold f, by using approximate MIS algorithms we can quickly tell that such a subgraph is actually frequent. However, in the cases in which the approximate MIS algorithm does not find a sufficiently large independent set, the exact MIS needs to be computed before a pattern can be kept or discarded. Also, if we need not only to find frequent subgraphs, but also to find their exact frequency, then the exact MIS needs to be computed on the overlap graph of every pattern. In both cases, because solving the exact MIS problem is in NP-complete (see Section 2.2), the above definition of the frequent subgraph discovery problem may not be tractable, even for a relatively simple input graph.

To make the problem more practical, we propose two alternative formulations that can find frequent subgraphs without solving the exact MIS problem.

Definition 2 (Approximate discovery). Given an input graph $\mathcal{G}$ which is undirected and labeled, and a parameter f, find *as many as possible* connected undirected labeled subgraphs that have at least f edge-disjoint embeddings in $\mathcal{G}$.

Definition 3 (Upper bound discovery). Given an input graph $\mathcal{G}$ which is undirected and labeled, and a parameter f, find *as few as possible* connected undirected labeled subgraphs such that an upper bound on the number of their edge-disjoint embeddings is above the threshold f.

Essentially the solutions for those two problems become a subset and a superset of the solution for Definition 1, respectively. The first formulation, Definition 2, which asks for a subset of the solution of Definition 1, requires that the embeddings of each subgraph form an overlap graph that has an *approximate MIS* whose size is greater than or equal to f. The second formulation, Definition 3, which asks for a superset of the solution of Definition 1, requires that an upper bound on the size of the exact MIS of this overlap graph is greater than or equal to f. As discussed in Section 2.2, such upper bounds can be easily obtained for both the GMIS algorithm as well as for other approximate algorithms. Note that inherent in these two last problem formulations is our desire to obtain a set of subgraphs whose size is as closed as possible to the actual number of frequent subgraphs that satisfy Definition 1. Thus, the effectiveness of these formulations and the associated algorithms will be evaluated with respect to both the computational efficiency that they achieve, and the number of patterns (or the size of the largest frequent patterns) they discover compared with the number of patterns discovered by 1.

5. Algorithms for finding frequent subgraphs in a large graph

We developed two algorithms, called HSIGRAM[2] and VSIGRAM, which find all frequent subgraphs according to Definitions 1–3 described in Section 4. In both algorithms, the frequent patterns are conceptually organized in a form of a lattice that is referred to as the *lattice of frequent subgraphs*. The kth level of this lattice contains all frequent subgraphs

Algorithm 1. HSIGRAM($\mathcal{G}$, *MIS_type*, f)

```
1:  ▷ f is the minimum frequency threshold.
2:  ▷ MIS_type is either approximate, exact or upper bound.
3:  F ← ∅
4:  F^1 ← all frequent size-1 subgraphs in G
5:  F^2 ← all frequent size-2 subgraphs in G
6:  k ← 2
7:  while F^k ≠ ∅ do
8:    C^{k+1} ← HSIGRAM-GEN(F^{k-1}, F^k, f)
9:    F^{k+1} ← ∅
10:   for each candidate C in C^{k+1} do
11:     C.freq ← HSIGRAM-COUNT(C, MIS_type)
12:     if C.freq ≥ f then
13:       add C to F^{k+1}
14:   F ← F ∪ F^{k+1}
15:   k ← k + 1
16: return F
```

with k edges (i.e., size-k subgraphs), and a node at level k, representing a subgraph G^k, is connected to at most k nodes at level $k-1$, each corresponding to a distinct (i.e., non-isomorphic) connected size-$(k-1)$ subgraph of G^k. The goal of both HSIGRAM and VSIGRAM is to identify the various nodes of this lattice and the frequency of the associated subgraphs.

The difference between the two algorithms is the method they use to discover (i.e., generate) the nodes of the lattice. HSIGRAM follows a horizontal approach and discovers the nodes in a breadth-first fashion, whereas VSIGRAM follows a vertical approach and discovers the nodes in a depth-first fashion. Both horizontal and vertical approaches have been previously used to find frequent subgraphs in the graph-transaction setting (Inokuchi et al., 2002; Kuramochi and Karypis, 2004a; Yan and Han, 2002; Borgelt and Berthold, 2002) and have their origins on algorithms developed for finding frequent itemsets and sequences (Agrawal and Srikant, 1994, 1995; Han et al., 2000; Zaki and Gouda, 2003). A detailed description of HSIGRAM and VSIGRAM is provided in the rest of this section.

5.1. *Horizontal algorithm:* HSIGRAM

The general structure of HSIGRAM is shown in Algorithm 1 (the notation used in the pseudo-code is shown in Table 1). HSIGRAM takes as input the graph $\mathcal{G}$, the minimum frequency threshold f, and the parameter *MIS_type* that specifies the particular problem definition (as discussed in Section 4). It starts by enumerating all frequent single- and double-edge subgraphs in $\mathcal{G}$, and then enters its main computational loop (Lines 7–15). During each iteration, HSIGRAM first generates all candidate subgraphs of size $k+1$ by joining pairs of size-k frequent subgraphs (Line 8) and then computes their frequency (HSIGRAM-COUNT in Line 11). The candidate subgraphs whose frequency is lower than the minimum threshold f are discarded and the remaining are kept for the next level of the algorithm. The computation terminates when no frequent subgraphs are generated during a particular iteration.

Algorithm 2. HSIGRAM-GEN$(\mathcal{F}^{k-1}, \mathcal{F}^k, f)$

```
1: C^{k+1} ← ∅
2: for each F in F^{k-1} do
3:    for each pair F_i, F_j in F.children do
4:       C ← join F_i and F_j based on F
5:       ▷ test if the downward closure property holds.
6:       S(C) ← all connected size-k subgraphs of C
7:       P(C) ← two primary subgraphs of size k
8:       skip ← false
9:       for each S in S(C) do
10:         if S.freq < f then
11:            skip ← true
12:            break
13:      if skip ≠ true then
14:         add C to C^{k+1}
15:         ▷ P(C) = {H_1, H_2}
16:         add C to H_1.children and to H_2.children
17: return C^{k+1}
```

The two key components of the HSIGRAM algorithm that significantly affect its overall computational complexity are the method used to perform candidate generation and the method used to compute the frequency of the candidate subgraphs. In the rest of this section we provide additional details on how these operations are performed and describe various optimizations that are designed to reduce their runtime.

5.1.1. Candidate generation. HSIGRAM generates candidate subgraphs of size $k+1$ by joining two frequent size-k subgraphs. In order for two such frequent size-k subgraphs to be eligible for joining each of the two must contain the same size-$(k-1)$ connected subgraph. The simplest way to generate the complete set of candidate subgraphs is to join all pairs of size-k frequent subgraphs that have a common size-$(k-1)$ subgraph. Unfortunately, the problem with this approach is that a particular size-k subgraph may have up to k different size-$(k-1)$ subgraphs and as a result, if we consider all such possible subgraphs and perform the resulting join operations, we will end up generating the same candidate pattern multiple times, and generating a large number of candidate patterns that are not downward closed. Such an algorithm would spend a significant amount of time identifying unique candidates and eliminating non-downward closed candidates (both of which operations are non-trivial as they require to determine the canonical label of the generated subgraphs).

HSIGRAM addresses both of these problems by only joining two frequent subgraphs if and only if they share a certain, properly selected, size-$(k-1)$ subgraph. Algorithm 2 shows the pseudo-code for the candidate generation, where the properly selected size-$(k-1)$ subgraph is denoted by F. For each frequent size-k subgraph F_i, let $P(F_i) = \{H_{i,1}, H_{i,2}\}$ be the two size-$(k-1)$ connected subgraphs of F_i such that $H_{i,1}$ has the smallest canonical label and $H_{i,2}$ has the second smallest canonical label among the various connected size-$(k-1)$ subgraphs of F_i. We will refer to these subgraphs as the *primary subgraphs* of F_i. Note that if every size-$(k-1)$ subgraph of F_i is isomorphic to each other, $H_{i,1} = H_{i,2}$ and $|P(F_i)| = 1$. HSIGRAM will only join two frequent subgraphs F_i and F_j, if and only if $P(F_i) \cap P(F_j) \neq \phi$,

Algorithm 3. hSiGraM-Count$(C^{k+1}, \mathit{MIS_type})$

1: $(\mathcal{M}(C^{k+1}), \mathcal{A}(C^{k+1})) \leftarrow$ hSiGraM-Embed$(C, \mathcal{G})$
2: $G \leftarrow$ build an overlap graph from $\mathcal{M}(C^{k+1})$
3: $\{G_1, G_2, \ldots, G_m\} \leftarrow$ decompose G
4: $f_{\text{MIS}} \leftarrow 0$
5: **for each** G_i **in** $\{G_1, G_2, \ldots, G_m\}$ **do**
6: **if** G_i is easy to handle **then**
7: $f_{\text{MIS}} \leftarrow f_{\text{MIS}} + |\mathsf{EMIS}(G_i)|$
8: **else if** *MIS_type* = approximate **then**
9: $f_{\text{MIS}} \leftarrow f_{\text{MIS}} + |\mathsf{GMIS}(G_i)|$
10: **else if** *MIS_type* = exact **then**
11: $f_{\text{MIS}} \leftarrow f_{\text{MIS}} + |\mathsf{EMIS}(G_i)|$
12: **else if** *MIS_type* = upper bound **then**
13: $f_{\text{MIS}} \leftarrow f_{\text{MIS}} + |\mathsf{GMIS}(G_i)| \min((\Delta + 2)/3, (\overline{d} + 2)/2)$
14: $\triangleright$ $\mathcal{S}(C^{k+1})$ is a set of all connected size-k subgraphs in C^{k+1}
15: $f_p \leftarrow$ the lowest frequency among $\mathcal{S}(C^{k+1})$
16: **return** $\min(f_{\text{MIS}}, f_p)$

and the join operation will be done with respect to the common size-$(k-1)$ subgraph(s). The proof that this approach will correctly generate all valid candidate subgraphs is presented in Kuramochi and Karypis (2004a). This candidate generation approach dramatically reduces the number of redundant and non-downward closed patterns that are generated and leads to significant performance improvements over the naive approach (Kuramochi and Karypis, 2001).

5.1.2. Frequency counting. hSiGraM-Count in Algorithm 3 computes the frequency of a candidate subgraph C by first identifying all of its embeddings, constructing the overlap graph of these embeddings, and then, based on the *MIS_type* parameter, finding an approximate or exact MIS of this overlap graph. The outline of this process is shown in Algorithms 3 and 4.

5.1.2.1. Embedding identification In order to identify all the embeddings of a candidate C, hSiGraM-Embed shown in Algorithm 4 needs to solve the subgraph isomorphism problem. Performing the subgraph isomorphism for every candidate from scratch is expensive, especially when the input graph is large. The overall computational requirements can be dramatically reduced if for each frequent subgraph we keep track of the exact location, within the input graph, of its various non-identical embeddings. This information can then be used to constraint the subgraph isomorphism problem by narrowing down the search space to only around these locations. However, this solution represents the classical memory-time trade-off. By maintaining in memory the complete embeddings of each pattern at a particular level of the lattice it reduces the computational complexity at the expense of dramatically increasing the memory requirements.

hSiGraM takes the middle road between these two approaches (recompute everything vs. save everything) and it only stores partial information about the embeddings of the frequent subgraphs. Specifically, for each embedding m_i of a frequent pattern F, hSiGraM-Embed arbitrarily selects one of the edges in the original graph that is also a part of m_i and uses it to

Algorithm 4. HSIGRAM-EMBED$(C, \mathcal{G})$

1: $\triangleright$ $\mathcal{A}$: a set of all anchor-edges of C.
2: $\mathcal{A} \leftarrow$ intersection of anchor-edges across $\mathcal{S}(C)$
3: $\triangleright$ collect all unique embeddings of C into $\mathcal{M}$.
4: $\mathcal{M} \leftarrow \emptyset$
5: **for each** anchor-edge e **in** $\mathcal{A}$ **do**
6: $\quad \mathcal{M}_e \leftarrow$ all embeddings of C that includes the edge e
7: $\quad \mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{M}_e$
8: $\triangleright$ collect all unique anchor-edges of C into $\mathcal{A}$.
9: $\mathcal{A} \leftarrow \emptyset$
10: **for each** embedding m **in** $\mathcal{M}$ **do**
11: $\quad e \leftarrow$ choose one edge from m arbitrarily
12: $\quad$ add e to $\mathcal{A}$
13: **return** $(\mathcal{M}, \mathcal{A})$

represent this partial information. We will refer to this edge as the *anchor-edge*, and the set of anchor-edges of a particular pattern F, as its *anchor-edge list* $\mathcal{A}(F)$. Note that the memory requirements of this approach is just one edge per embedding and does not change with the size of the frequent pattern. In addition, in the cases in which there is a high degree of overlap between the different embeddings, there can be a number of identical edges being selected as anchor-edges of different embeddings, out of which HSIGRAM stores only one, further reducing the memory requirements.

These anchor-edge lists are used to reduce the overall computational requirements as follows. Suppose a k-candidate C contains a frequent $(k-1)$-subgraph F_i. Because there are k edges in $E(C)$, C may have up to k distinct such frequent subgraphs of size $k-1$, and each F_i holds the anchor-edge list. Before starting the frequency counting of C, first HSIGRAM-EMBED selects one of F_i whose frequency is the lowest among $\{F_i\}$. For each $e_n \in \mathcal{A}(F_i)$, HSIGRAM-EMBED checks if there is an edge $e_m \in \mathcal{A}(F_j)$ for all $j \neq i$ such that the shortest path length between e_n and e_m, denoted by d, is within the diameter of C, denoted by dia(C). If there is such an edge e_m from every $\mathcal{A}(F_j)$ for $j \neq i$, e_n may be a part of an embedding of C, because if C is a frequent subgraph of size k, there must be a set of frequent subgraphs of size $k-1$ inside the same embedding of C. To compute the exact path length between edges e_n and e_m in $\mathcal{G}_i$ requires all pairs shortest paths, which may be computationally expensive when $|E(\mathcal{G}_i)|$ is large. HSIGRAM-EMBED bounds this length d by the difference between two lengths, $|d_n - d_m|$, where d_n and d_m are the shortest path lengths from an arbitrarily chosen vertex $v \in V(\mathcal{G}_i)$ to e_n and e_m respectively. Note that because this v is fixed through the entire process of the algorithm, it is necessary to compute the shortest paths such as d_n and d_m only once. If e_n and e_m are in the same embedding of C_i, always $d \leq \text{dia}(C)$ holds and $d_n \leq d_m + d$. Thus, if $|d_n - d_m| \leq \text{dia}(C)$ is true, then e_n and e_m may belong to the same embedding of C, otherwise e_n and e_m cannot be in the same embedding (see Figure 4). If e_n cannot find such e_m from every $\mathcal{A}(F_j)$ for $j \neq i$, e_m is removed from $\mathcal{A}(F_i)$ (Line 2). Because the subgraph isomorphism will be performed for each e_n, this pruning procedure is quite effective in reducing the overall runtime.

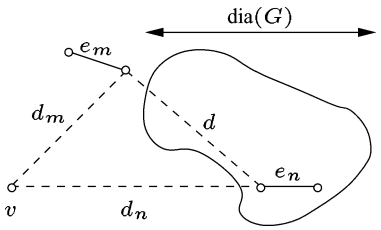


Figure 4 Distance estimation between two edges.

Finally, after removing unnecessary anchor-edges, for each of the remaining anchor-edges, all the subgraph isomorphisms of C are repeatedly identified and the set of embeddings $\mathcal{M}$ is built (Line 6).

5.1.2.2. Computing the frequency. The frequency of each subgraph C^{k+1} is computed by the HSIGRAM-COUNT function shown in Algorithm 3.

HSIGRAM-COUNT computes two different frequencies. The first, denoted by f_{MIS}, is computed based on the size of the MIS of the overlap graph created from the embeddings of C^{k+1}. The second, denoted by f_p, is the least frequency of all the connected size-k subgraphs of C^{k+1} (Line 15), which represents an upper bound on C^{k+1}'s frequency derived entirely from the lattice of frequent subgraphs. In the case in which f_{MIS} is computed using Definition 3, the frequency bound provided by f_p may actually be tighter, and thus may lead to more effective pruning. For this reason, the overall frequency of C^{k+1} is obtained by taking the minimum of f_{MIS} and f_p.

The frequency f_{MIS} is computed as follows (Lines 2–13). Given a pattern and all of its non-identical embeddings, HSIGRAM-COUNT generates its overlap graph G. Then, HSIGRAM-COUNT decomposes G into its connected components $G_1, G_2, \ldots, G_m$ ($m \geq 1$). Next, for each connected component G_i, it checks the maximum degree of its vertices and if it is less than or equal to two (a cycle or a path), it computes its maximum independent set directly by the EMIS algorithm because it is trivial to compute the exact MIS for this class of graphs (Line 7). If the maximum degree is greater than two, HSIGRAM-COUNT uses either the result of the GMIS algorithm (Line 9), the result of the EMIS algorithm (Line 11), or the upper bound on the size of the exact MIS (Eq. (1)). The summation of those MIS sizes for the components is the final value of f_{MIS}. Note that the decomposition of the overlap graph into its connected components allow us to take advantage of the properties of the special graphs and also obtain tighter bounds for each component as the maximum degree for some of them will be lower than the maximum degree of the entire overlap graph.

In addition, every edge is marked if it is included in any embedding of a frequent subgraph. Unmarked edges are removed before proceeding to the next iteration.

5.2. *Vertical algorithm:* VSIGRAM

The most computationally expensive step in the HSIGRAM algorithm is frequency counting as it needs to repeatedly perform subgraph isomorphism computations. The overall time can

be greatly reduced if instead of storing only the anchor-edges we store the complete set of embeddings across successive levels of the algorithm. Because of HSIGRAM's level-by-level structure, these complete embeddings need to be stored for the entire set of frequent and candidate patterns of each successive pair of levels. This substantially increases the memory requirements of this approach, making it impractical for most of interesting datasets. On the other hand, within the context of a vertical algorithm, storing the complete set of embeddings is feasible since we need to do that only for the subgraphs along the path from the current node to the root. Thus, a vertical algorithm has potentially a computational advantage over a horizontal algorithm, which motivated the development of VSIGRAM.

However, before developing efficient algorithms that generate the lattice of frequent subgraphs in a depth-first fashion two critical steps need to be addressed. The first step is the method used to ensure that the same node of the lattice and the depth-first subtree rooted at that node should not be discovered and explored multiple times. This is important because each node at level k will be connected to up to k different nodes at level $(k-1)$. As a result, if there are no mechanisms by which to prevent the repeated generation of the same node, a depth-first algorithm will end-up performing redundant computations (i.e., generating the same nodes multiple times), adversely impacting the overall performance of the algorithm. VSIGRAM eliminates these redundant computations by assigning each node at level k (corresponding to a subgraph F^k) to a unique parent node at level $k-1$ (corresponding to a subgraph F^{k-1}), such that only F^{k-1} is allowed to create F^k. The subgraph F^{k-1} is called *the generating parent* of F^k. Details on how this is achieved is provided in Section 5.2.1.

The second step is the method used to create successor nodes in the course of the traversal. In the case of HSIGRAM, this corresponds to the candidate generation phase, and is performed by joining the frequent subgraphs of the previous level. However, since the lattice is explored in a depth-first fashion, such joining-based approach will not work, as the algorithm may not have yet discovered the required frequent subgraphs. To address this problem, VSIGRAM creates the successor nodes (i.e., extended subgraphs) by analyzing all the embeddings of the current subgraph F^k, and identifying the distinct one-edge extensions to these embeddings that are sufficiently frequent. The frequent extensions for which F^k is the generating parent are then used as the successor nodes during the depth-first traversal.

The general structure of VSIGRAM is shown in Algorithm 5. VSIGRAM starts by determining all frequent size-1 patterns and then uses each one of them as the starting point of a recursive depth-first extension (VSIGRAM-EXTEND function). VSIGRAM-EXTEND takes as input a size-k frequent subgraph F^k and all of its embeddings $\mathcal{M}(F^k)$ in $\mathcal{G}$ and proceeds as follows. For each size-k embedding $m \in \mathcal{M}(F^k)$, it identifies and stores every possible size-$(k+1)$ subgraph in $\mathcal{G}$ that contains m. From this set of subgraphs, it extracts all size-$(k+1)$ subgraphs which are not isomorphic to each other and stores them in $\mathcal{C}^{k+1}$. Then, VSIGRAM-EXTEND eliminates from $\mathcal{C}^{k+1}$ all the subgraphs that do not have F^k as their generating parent (Lines 5–6) or are infrequent (Lines 7–8). The subgraphs remaining in $\mathcal{C}^{k+1}$ are the frequent subgraphs of size-$(k+1)$ obtained by an one-edge-extension of F^k and are used as input for the next recursive call. The recursion terminates when $\mathcal{C}^{k+1} = \emptyset$, and the depth-first search backtracks.

Algorithm 5. vSIGRAM

vSIGRAM($\mathcal{G}$, *MIS_type*, f)
1: $\mathcal{F} \leftarrow \emptyset$
2: $\mathcal{F}^1 \leftarrow$ all frequent size-1 subgraphs in $\mathcal{G}$
3: **for each** F^1 **in** $\mathcal{F}^1$ **do**
4: $\quad \mathcal{M}(F^1) \leftarrow$ all embeddings of F^1
5: **for each** F^1 **in** $\mathcal{F}^1$ **do**
6: $\quad \mathcal{F} \leftarrow \mathcal{F} \cup$ vSIGRAM-EXTEND($F^1, \mathcal{G}, f$)
7: **return** $\mathcal{F}$

vSIGRAM-EXTEND($F^k, \mathcal{G}$, *MIS_type*, f)
1: $\mathcal{F} \leftarrow \emptyset$
2: **for each** embedding m **in** $\mathcal{M}(F^k)$ **do**
3: $\quad \mathcal{C}^{k+1} \leftarrow \mathcal{C}^{k+1} \cup \{$all $(k+1)$-subgraphs of $\mathcal{G}$ containing $m\}$
4: **for each** C^{k+1} **in** $\mathcal{C}^{k+1}$ **do**
5: $\quad$ **if** F^k is not *the generating parent* of C^{k+1} **then**
6: $\quad\quad$ **continue**
7: $\quad$ compute C^{k+1}.freq from $\mathcal{M}(C^{k+1})$
8: $\quad$ **if** C^{k+1}.freq $< f$ **then**
9: $\quad\quad$ **continue**
10: $\quad$ add C^{k+1} to $\mathcal{F}$
11: **return** $\mathcal{F}$

In the rest of this section we provide additional details on how the various operations are performed and describe various optimizations that are designed to reduce vSIGRAM's runtime.

5.2.1. Generating parent identification. The scheme that vSIGRAM uses to determine the generating parent of a particular subgraph is as follows (see figure 5 for example). Suppose a size-$(k+1)$ frequent subgraph F^{k+1} is just created by extension from a size-k frequent subgraph F^k. By the canonical labeling, the order of edges and vertices in F^{k+1} is uniquely determined. vSIGRAM removes the last edge that does not disconnect F^{k+1} and obtains another size-k subgraph F. If F is isomorphic to F^k then F^k becomes the generating parent of F^{k+1}, and vSIGRAM keeps the further exploration from F^{k+1}. It is easy to see that for each subgraph F^{k+1} exactly one of its distinct size-k connected subgraphs will satisfy this property; thus, this approach will explore each subgraph exactly one time. Let us take an illustrative example. In figure 5, a size-6 pattern F^6 in (e) has four size-5 subgraphs (i.e., equivalence classes in terms of isomorphism) shown in (a), (b), (c) and (d). Suppose, by canonical labeling, every edge in F^6 is assigned a unique ID (from e_0 to e_5 as shown in figure 5(e)). Then, we remove the last edge e_5 from F^6 to obtain a size-5 subgraph. Because this size-5 subgraph is isomorphic to ${F_b}^5$ shown in figure 5(b), ${F_b}^5$ is the generating parent of F^6. In other words, when we reach a node in the depth-first search space corresponding to ${F_a}^5$, ${F_c}^5$ or ${F_d}^5$, we will stop the exploration that leads to F^6 because ${F_a}^5$, ${F_c}^5$ and ${F_d}^5$ are not the generating parent of F^6. Thus, only the extension from ${F_b}^5$ to F^6 is kept for further exploration.

Note that similar type of approaches have been used in the past in the context of vertical algorithms for the graph-transaction setting (Vanetik et al., 2002; Yan and Han, 2002).

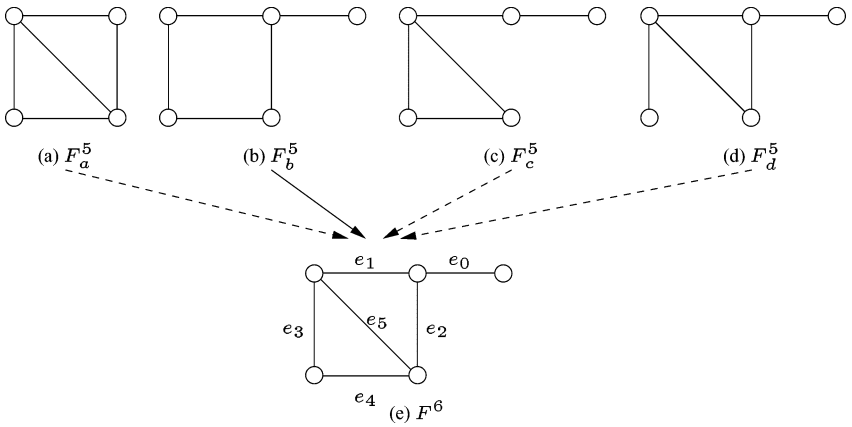


Figure 5 A simple example of parent identification.

All of these share the same idea, which avoids redundant frequent pattern generation and traverses the lattice of patterns as if it was a tree.

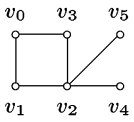
5.2.2. Efficient subgraph extension. Starting from a frequent size-k subgraph, VSIGRAM obtains the extended subgraphs of size $k+1$ by adding an additional edge (while preserving connectivity) to all of its possible embeddings. Specifically, for each embedding m of a frequent k-subgraph F, VSIGRAM enumerates all the edges that can be added to m to form a size-$(k+1)$ extended subgraph. Each of those edges is represented by a tuple of 5 elements $s = (x, y, u, v, e)$, called a *stem*, where x and y are the vertex IDs of the edge in $\mathcal{G}$, u and v, $u < v$, are the corresponding vertex IDs in F, and e is the label of the edge. For u and v, if there is no corresponding vertex in F, -1 is used to show that it is outside the subgraph F.

However, because of the automorphism of the subgraph F, we cannot use this stem representation directly. For a particular embedding m of a frequent subgraph F in $\mathcal{G}$, there may be more than one vertex mapping of the subgraph onto the embedding. If we simply used a pair of vertex IDs of the subgraph to represent a stem, depending on the mapping, the same edge addition might be considered a different stem, which would result in the wrong frequency of the subgraph. To avoid this problem, every time a stem is generated, its representation is normalized as follows. VSIGRAM enumerates all possible automorphisms of F, denoted by $\{\phi_i\}$. Note that each ϕ_i can be regarded as a mapping from a vertex ID v to another vertex ID w based on the automorphism (e.g., $w = \phi_i(v)$). Then, for each vertex ID v, the *canonical* vertex ID (cvid) is defined as

$$\text{cvid}(v) = \min_i \phi_i(v).$$

The automorphism with the least subscript that gives the canonical ID for v is called the *canonical automorphism*, denoted by ϕ_v^*. Formally ϕ_v^* is defined as follows:

$$\phi_v^* = \arg\min_{\phi_i} \phi_i(v), \quad i < j \quad \text{if } \phi_i(v) = \phi_j(v).$$



(a) Graph G

Input	v_0	v_1	v_2	v_3	v_4	v_5
ϕ_0	v_0	v_1	v_2	v_3	v_4	v_5
ϕ_1	v_0	v_1	v_2	v_3	v_5	v_4
ϕ_2	v_0	v_3	v_2	v_1	v_4	v_5
ϕ_3	v_0	v_3	v_2	v_1	v_5	v_4
cvid	v_0	v_1	v_2	v_1	v_4	v_4
ϕ^*	ϕ_0	ϕ_0	ϕ_0	ϕ_2	ϕ_0	ϕ_1

(b) Canonical vertex ID and automorphism

Figure 6 Size-6 graph G, canonical vertex IDs, and canonical automorphism.

As an example, let us take the graph G shown in figure (a) having 6 vertices whose vertex ID's are assigned from v_0 to v_5. There are 4 distinct automorphisms in total, from ϕ_0 to ϕ_3 as shown figure 6(b) (the mapping ϕ_0 is the trivial one). For instance, looking at the fourth row of the table in figure 6(b), denoted by "ϕ_2", we see the automorphism ϕ_2 maps v_0 to v_0, v_1 to v_3, v_2 to $v2$, v_3 to v_1, v_4 to v_4, and v_5 to v_5 . Next, looking at the fifth column, denoted by "v_3", we see that $\text{cvid}(v_3) = v_1$ and hence $\phi^*_{v_3} = \phi_2$, because ϕ_2 is the mapping with the smallest suffix, "2", that gives $v_1 = \phi_i(v_3)$. Finally, we can canonicalize every stem $s = (x, y, u, v, e)$ based on the canonical automorphism. In the canonical stem $s' = (x, y, u', v', e)$ for s, u' and v' are defined as:

$$\begin{array}{llll} u' \equiv \text{cvid}(u), & v' \equiv \phi^*_u(v) & \text{if } \text{cvid}(u) \leq \text{cvid}(v) \\ u' \equiv \phi^*_v(u), & v' \equiv \text{cvid}(v) & \text{otherwise.} \end{array}$$

Note that s' is an automorphism invariant representation of s and is used by vSIGRAM to properly determine the frequency of size-$(k+1)$ extended subgraphs. For example, suppose there exists a stem $s = (x, y, v_2, v_3, e)$. From the table in Figure 6(b), we can see that $\text{cvid}(v_2) = v_2$, $\text{cvid}(v_3) = v_1$ and $\text{cvid}(v_2) > \text{cvid}(v_3)$ by which we have ${v_2}' = \phi^*_{v_3}(v_2) = \phi_2(v_2) = v_2$ and ${v_3}' = \text{cvid}(v_3) = v_1$. Therefore, the canonical stem for s is eventually represented as $s' = (x, y, v_2, v_1, e)$.

5.2.2.1. Additional optimization: Keeping track of edge creation status. Each frequent subgraph maintains a three-dimensional table, called a *connection table*. Each element in the table is denoted by $\text{ct}(u', v', e)$ which shows if it is possible to form an edge between the vertices u' and v' whose edge label is e. Every time a stem (x, y, u', v', e) is discarded, the corresponding element in the connection table is updated to show that it is now impossible to create an edge with a label e between u' and v'. If $\text{ct}(u', v', e)$ is deactivated for a frequent subgraph of size k, then for any $l > k$, there can not be any frequent subgraph that has an edge between u' and v' with the edge label e. We can reduce the number of stems to be generated by looking up the connection table during the stem enumeration phase.

5.2.3. Frequency counting. vSIGRAM's frequency counting is similar to HSIGRAM-COUNT, except for the computation of f_p (see Line 15 in Algorithm 3). Recall from the discussion in Section 5.1.2.2, f_p is an estimate of the maximum frequency of a size-$(k+1)$ pattern p that is determined by looking at the minimum frequency of all of its connected size-k subgraphs. This estimate takes advantage of the downward closure property of the

Table 2. Datasets used in the experiments.

	Connected component							
		Size			Total number of		Total number of labels	
Dataset	Number	Min	Max	Avg	Vertices	Edges	Vertex	Edge
Aviation	2, 606	21	55	38	101, 088	98, 482	6, 171	51
Credit	700	20	20	20	14, 700	14, 000	59	20
Citation	613	1	19, 545	34	12, 628	21, 032	50	12
Contact Map	170	79	1, 995	660	33, 443	112, 244	21	2
DTP	2, 080	1	110	21	40, 879	43, 070	52	3
PPI	3, 225	1	1, 850	5	19, 228	17, 142	11, 103	7
VLSI	947	1	7, 768	12	11, 066	11, 542	23	1
Web	25, 374	1	21, 860	4	120, 166	98, 137	2, 299	1

Size (Min, Max, Avg) of connected component: the number of edges in the connected component of the minimum, maximum and average size, respectively.

minimum frequency constraint, and in the case of the upper-bound problem formulation (i.e., according to Definition 3), it can lead to a frequency upper bound that is smaller than that obtained from MIS.

However, vSiGraM cannot use this approach to obtain such bounds as it does not hold all size-k frequent subgraphs at the time a size-$(k+1)$ subgraph is created. As a result, vSiGraM uses the frequency of the size-k generating parent as the best estimate for f_p. As the experiments presented in Section 6 will show, the f_p estimates obtained by vSiGraM are not as tight as the corresponding estimates obtained by hSiGraM, which in turns increases the former's execution time as it will tend to find more subgraphs (note that these subgraphs will not be frequent according to Definition 1).

6. Experimental evaluation

In this section, we study the performance of the proposed algorithms with various parameters and real datasets. All experiments were done on an AMD Athlon MP 1800+ (1.53 GHz) machines with 2 GBytes main memory, running the Linux operating system. All the runtimes reported are in seconds.

6.1. Datasets

We used eight different datasets, each obtained from a different domain, to evaluate and compare the performance of hSiGraM and vSiGraM. The basic characteristics of these datasets are shown in Table 2. Note that even though these graphs consist of multiple connected components, the hSiGraM and vSiGraM algorithm treat them as one large graph and discover the frequent patterns according to Definitions 1–3 described in Section 4.

The *Aviation* and *Credit* datasets are obtained from the SUBDUE web site.[3] The Aviation dataset is originally from the Aviation Safety Reporting System Database and the Credit dataset is from the UCI machine learning repository (Blake and Merz, 1998). The directed edges in the original graph data were converted into undirected ones. For the Aviation dataset, we removed undirected edges to show "near_to" relation between two vertices because those edges form cliques which makes this graph difficult to mine.

The *Citation* dataset was created from the citation graph used in KDD Cup 2003.[4] Each vertex in this graph corresponds to a document and each edge corresponds to a citation. Because our algorithms are for undirected graphs, the direction of these citations was ignored. Since the original dataset does not have any meaningful label for vertices, we generated vertex labels as follows. We first used a clustering algorithm to form clusters of the document abstracts into 50 thematically coherent topics, and then assigned the cluster ID as the label to the corresponding vertices. For the edges, we used as labels the difference in the publication year of the two papers. For example, if two papers were published in 1997 and 2002, an edge is created between those two document vertices with the label "5". Finally, because some of the vertices in the resulting graph had a very high degree (i.e., authorities and hubs), we kept only the vertices whose degree was less or equal to 15.

The *Contact Map* dataset is made of 170 proteins from the Protein Data Bank (Berman et al., 2000) with pairwise sequence identity lower than 25%. The vertices in these graphs correspond to the different amino acids and the edges connect two amino acids if they are either at consecutive sequence positions or they are in contact in their 3D structure. Two amino acids are considered to be in contact if the distance between their C_α atoms is less than 8 Å. Furthermore, while creating the graphs we only considered non-local contacts that are defined as the contacts between amino acids whose sequence separation is at least six amino acids.

The *DTP* dataset is a collection of 2,319 chemical compounds randomly selected from the dataset of 223,644 chemical compounds provided by the Developmental Therapeutics Program (DTP) at National Cancer Institute.[5] Note that each chemical compound forms a connected component and there are 2,319 such components in this dataset. Each vertex corresponds to an atom and its label represents the atom type. An edge is formed between two vertices if the corresponding two atoms are connected by a bond. The type of a bond is used as an edge label, and there are three distinct edge labels.

The *PPI* dataset is created from Database of Interacting Proteins (DIP).[6] A vertex is either a particular protein sequence, or a property of a protein sequence. A property of a protein sequence is the information about the classification of a protein sequence. For example, if there is an entry about a particular protein sequence in PDB (RCSB Protein Data Bank), a vertex is created which corresponds to this entry in PDB. Likewise, if there is an entry about a particular protein sequence in SCOP (Structural Classification of Proteins), a vertex is created to represent this entry in SCOP. An edge is used to show either the possession of those properties, or the interaction between protein sequences. An edge used to show the possession of a property has an label to represent the type of the property. For example, if there is an entry about a protein sequence in PDB, there is an edge connecting the protein sequence vertex and the PDB entry vertex, with the edge label "PDB". All the other edges representing interactions between protein sequences have the same edge label

"link". Out of the total 19,228 vertices, 8,126 vertices correspond to protein sequences (i.e., the remaining 11102 vertices correspond to properties), and out of the total 17,142 edges, there are 3,988 edges having the "link" label.

The *VLSI* dataset was obtained from the International Symposium on Physical Design '98 (ISPD98) benchmark suite[7] and corresponds to the netlist of a real circuit. The netlist was converted into a graph by first removing any nets that are longer than four and then using a star-based approach to replace each net (i.e., hyperedge) by a set of edges. Note that for this dataset we limited the size of the largest discovered pattern to five edges. This is because for the values of the frequency threshold used in our experiments, the only frequent patterns that contained more than five edges were paths, and because of the highly connected nature of the underlying graph, there were a very large number of such paths, making it hard to find these longer path patterns in reasonable amount of time.

The *Web* dataset was obtained from the 2002 Google Programming Contest.[8] The original dataset contains various web pages and links from various "edu" domain. We converted the dataset into an undirected graph in which each vertex corresponds to a web page and an edge to a hyperlink between web pages. In creating this graph, we kept only links between "edu" domains that connected sites from different subdomains. Every edge has an identical label (i.e., unlabeled), whereas each vertex was assigned a label corresponding to the subdomain of a web server.

Note that for both the *PPI* and the *Web* datasets we removed vertices whose degree is greater than 10 and all edges incident to such vertices in order to make the entire graph sparse and to allow our algorithms finish the computation in a reasonable amount of time.

Compared to the other datasets, *DTP* and *Contact Map* datasets are both originally a collection of graphs (chemical compounds and protein sequences respectively). Hence, it is possible to use those two datasets in the graph-transaction setting where the task is to discover frequent patterns in terms of the number of supporting transactions. On the other hand, in this study, we used those two datasets in the single-graph setting in order to find frequent occurring patterns for which multiple embeddings in each connected component are counted.

6.2. Results

Table 3 shows the results obtained by the HSIGRAM and VSIGRAM algorithms for the different datasets, for a wide range of the minimum frequency threshold values f, and the three different MIS-based problem definitions. For each experiment, Table 3 shows the amount of time (in seconds) required by the particular algorithm, the total number of patterns that were discovered, and size of the largest pattern. Entries in the table marked with "—" represents experiments that were aborted because of high computational requirements.

From these results we can see that as expected, for all datasets and algorithms, as the value of f decreases, the runtime for finding the frequent patterns increases as well. The rate of increase in runtime follows the corresponding rate of increase in the number of patterns that are being discovered. Besides that, the results in this table help illustrate the relation between the two key variables in these experiments, which are the type of the particular

Table 3. Runtime in seconds and the number of found frequent patterns for the different datasets.

		Runtime[sec]						Number of Found Patterns						Largest Pattern Size					
Aviation		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	1750	779	342	787	342	789	341	2249	2249	2249	2249	2249	2249	9	9	9	9	9	9
	1500	1603	743	1674	745	1584	739	5207	5207	5207	5207	5207	5207	10	10	10	10	10	10
	1250	2726	1461	2720	1496	2781	1486	11087	11087	11087	11087	11087	11087	12	12	12	12	12	12
	1000	5256	3667	5158	3683	5596	3818	30331	30331	30331	30331	30331	30331	13	13	13	13	13	13
Citation		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	100	0.1	0.0	0.1	0.0	0.1	0.0	6	6	6	6	7	11	1	1	1	1	2	5
	50	0.1	0.1	0.1	0.1	0.6	–	39	39	39	39	113	–	2	2	2	2	7	–
	20	0.6	0.3	0.9	0.5	139	–	266	266	266	266	12203	–	3	3	3	3	16	–
	10	4.0	1.5	4.2	1.9	–	–	986	986	988	988	–	–	5	5	5	5	–	–
Contact Map		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	300	10	3	10	3	183	–	186	186	186	186	2358	–	2	2	2	2	10	–
	200	44	9	45	9	–	–	505	505	505	505	–	–	3	3	3	3	–	–
	100	362	63	356	71	–	–	3183	3183	3186	3186	–	–	5	5	5	5	–	–
	50	3505	607	3532	632	–	–	29237	29237	29298	29298	–	–	6	6	6	6	–	–
Credit		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	200	10	4	10	4	9	4	1325	1325	1325	1325	1325	1325	7	7	7	7	7	7
	100	49	20	45	21	45	20	11696	11696	11696	11696	11696	11696	9	9	9	9	9	9
	50	169	78	172	80	169	78	73992	73992	73992	73992	73992	73992	11	11	11	11	11	11
	20	2019	461	1855	468	1880	462	613884	613884	613884	613884	613884	613884	13	13	13	13	13	13
DTP		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	100	113	27	114	27	169	64	1244	1244	1244	1244	2484	3788	12	12	12	12	16	18
	50	145	34	134	35	247	103	4028	4028	4028	4028	8295	13622	14	14	14	14	18	21
	20	243	86	249	83	616	19998	21477	21477	21478	21478	52180	824702	16	16	16	16	20	81
	10	813	311	882	294	2018	–	112535	112535	112539	112539	232810	–	21	21	21	21	21	–
PPI		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	1000	3	1	–	–	–	–	2	2	–	–	–	–	2	2	–	–	–	–
	500	10	5	–	–	–	–	4	4	–	–	–	–	3	3	–	–	–	–
	200	160	155	–	–	–	–	12	12	–	–	–	–	6	6	–	–	–	–
	100	2374	2265	–	–	–	–	34	34	–	–	–	–	8	8	–	–	–	–
VLSI		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	100	42	7	–	–	54	10	379	379	–	–	519	609	5	5	–	–	5	5
	75	49	8	–	–	56	10	409	409	–	–	571	679	5	5	–	–	5	5
	50	236	15	–	–	282	17	683	683	–	–	946	1051	5	5	–	–	5	5
	25	428	18	–	–	469	20	1452	1452	–	–	1907	2131	5	5	–	–	5	5
Web		Apprx.		Exact		UB		Apprx.		Exact		UB		Apprx.		Exact		UB	
	f	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V	H	V
	200	1	1	1	1	1	–	11	11	11	11	15	–	1	1	1	1	4	–
	100	1	2	437	–	–	–	50	50	50	–	–	–	2	2	2	–	–	–
	50	2	2	–	–	–	–	197	197	–	–	–	–	3	3	–	–	–	–
	20	62	129	–	–	–	–	1160	1160	–	–	–	–	6	6	–	–	–	–

Note. Dashes indicate the computation was aborted because of the too long runtime or memory exhaustion.
f: the minimum frequency threshold, H: HSIGRAM, V:VSIGRAM
Approx.: with approximate MIS, Exact: with exact MIS, UB: with upper-bound MIS

algorithm (HSIGRAM vs VSIGRAM) and the type of frequency calculation (approximate MIS, exact MIS, or upper-bound MIS).

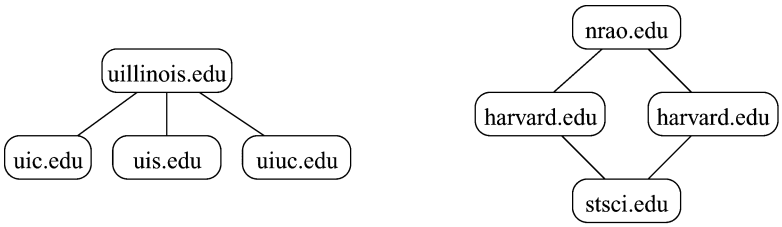
6.2.1. Comparison between HSIGRAM and VSIGRAM. In general, the amount of time required by VSIGRAM is smaller than that required by HSIGRAM. In fact, as the value of

the frequency threshold decreases, vSIGRAM is up to five times faster than hSIGRAM. This is true across all datasets for the approximate and exact MIS problem formulation, and for those datasets for which the upper-bound MIS formulation leads to the same number of frequent patterns for both algorithms. As discussed in Section 5.2, the reason for that performance advantage is the fact that by keeping track the embeddings of the frequent subgraphs along the depth-first path, vSIGRAM spends significantly less time in subgraph isomorphism related computations than hSIGRAM does.

However, for certain datasets, when the upper-bound MIS formulation is used, vSIGRAM ends up generating significantly more patterns than those generated by hSIGRAM. For example, in the case of the DTP dataset and $f = 20$, vSIGRAM generates almost 16 times more patterns than hSIGRAM. In such cases, the amount of time required by vSIGRAM is substantially greater than that required by hSIGRAM (32.4 times greater in the DTP example). The reason for that is the fact that because of its depth-first nature, vSIGRAM cannot take advantage of the frequent subgraph lattice to get a tight upper bound on the frequency of a subgraph based on the frequency of all of its subgraphs, and it bases its upper bound only on the frequency of the generating parent. On the other hand, because of its level-by-level nature, hSIGRAM can use the information from all its sub-patterns, and obtains better upper bounds (see discussion in Section 5.1.2.2).

6.2.2. Comparison between the different problem formulations. The results in Table 3 show that the approximate MIS-based formulation is able to effectively mine all datasets and tends to lead to the fastest execution times for both hSIGRAM and vSIGRAM. This result was expected, and was the key motivation behind developing this formulation. However, a somewhat surprising but positive outcome is that when comparing the number of patterns found by the approximate and the exact MIS-based formulations we can see that, in general, the approximate algorithm finds most of the patterns discovered by the exact algorithm and it only fails to discover a very small number of patterns.

On the other hand, both the exact and the upper-bound based MIS formulations are not as robust. Specifically, the exact formulation can mine only five out of the eight datasets, whereas the upper-bound formulation can mine only four. Even though these formulations tend to fail on approximately the same set of datasets (PPI and Web), the specific reasons for these failures are different. The exact formulation fails when the resulting overlap graph becomes highly interconnected, in which cases the time associated with finding the exact MIS is prohibitively expensive (i.e., the computations were aborted because they were not expected to finish within a reasonable amount of time). The upper-bound formulation fails when the bounds on the number of frequent patterns are not very tight, in which case the algorithm generates a very large number of patterns that exhaust the available memory. Note that in general, the extent to which we can obtain a tight upper-bound on the number of frequent patterns using the methodology described in Section 5.1.2.2, depends on the maximum and average degree of the overlap graph. Thus, when the overlap graph is not very sparse, then both the exact and the upper-bound formulations will tend to fail (this is the reason why both schemes fail for the PPI and the Web datasets). However, this correspondence is not perfect. For example, only the exact MIS-based formulation was able to finish for the Citation dataset, whereas only the upper-bound formulation was able to finish for the VLSI dataset.



(a) University of Illinois System (b) Astronomy related web pages

Figure 7 Examples of patterns discovered by SIGRAM.

6.2.3. Example subgraphs. To illustrate the types of subgraphs that SIGRAM can discover, we analyzed the subgraphs that were identified in the Web dataset. Recall from Section 6.1 that each vertex in this graph corresponds to an actual web-page, each edge to a hyperlink between two web-pages, and the vertex-labels to the subdomain of the server that hosts the web-page. Moreover, this graph was constructed by removing any hyperlinks between web-pages that have the same subdomain. As a result, a frequently occurring subgraph will represent a particular cross-linking structure among a specific set of institutions that occurs often, and it can identify common cross-university collaborations, interdisciplinary teams, or topic-specific communities.

Figure 7 shows two representative examples of the frequent subgraphs discovered by SIGRAM with the approximate MIS when the minimum frequency threshold is set to 20. The first subgraph (figure 7(a)) has a star topology and shows that there are links from a web page of the University of Illinois to web pages located at sites of all the three campuses, Chicago, Urbana-Champaign and Springfield, in the university system. SIGRAM identified 28 non-identical embeddings of this subgraph in the Web dataset, and 22 out of 28 embeddings are edge-disjoint.

The second subgraph (figure 7(b)) has a square-shaped topology consisted of web sites at Harvard, National Radio Astronomy Observatory (nrao.edu), and Space Telescope Science Institute (stsci.edu). An analysis of the complete uniform resource locators (URLs) of the embeddings of this subgraph showed that all the web pages had to do with astronomy and astrophysics. There are 64 non-identical embeddings of this substructure in total, and the number of edge-disjoint embeddings is 28.

6.3. Performance comparison with existing algorithms

6.3.1. Comparison with SUBDUE. We ran SUBDUE (Holder et al., 1994) version 5.0.6[9] on the same datasets described in Section 6.1 and measured the runtime, the number of discovered patterns, their size, and their frequency. These results are shown in Table 4. These results were obtained by using SUBDUE's default settings for all but the VLSI dataset. For the VLSI dataset, we run SUBDUE so that to find subgraphs that contain at most five edges, as was done in the case of HSIGRAM and VSIGRAM. Note that SUBDUE's default settings returns at most three subgraphs that were determined to be the most important. Also note

Table 4. SUBDUE Results.

Dataset	Runtime [sec]	Number of patterns	Pattern size			Frequency of found patterns		
Aviation	–	–	–	–	–	–	–	–
Citation	8812	3	27	26	27	1	1	1
Contact map	5043	3	224	223	223	1	1	1
Credit	517	3	6	5	5	341	395	387
DTP	1525	3	2	2	6	4957	4807	1950
PPI	2928	3	1	1	1	1	1	1
VLSI	16	3	1	1	1	773	773	244
Web	–	–	–	–	–	–	–	–

that for the Aviation and Web datasets we had to abort SUBDUE as it did not finish after three days.

Because of the inherent differences between SUBDUE and our algorithms, it is impossible to perform a direct comparison of the results that they generate. For this reason our comparisons will focus mostly on highlighting some key points. First, the amount of time required by SUBDUE is in general, considerably higher than that required by our algorithms. For example, SUBDUE did not finish the computation for the Aviation dataset after spending four entire days. Also for the Citation, Contact Map and PPI datasets, SUBDUE could not find any meaningful patterns at all, as the patterns that it found had a frequency of one. For the Credit dataset with the minimum frequency threshold of 50, both HSIGRAM and VSIGRAM with upper-bound MIS spent 169 and 78 seconds respectively to discover the same number of subgraphs, 73,992. The largest pattern has 11 edges and had a frequency of 58. In contrast, the largest pattern found by SUBDUE had six edges with a frequency of 341. This indicates that if there are small subgraphs that have relatively high frequency, SUBDUE will focus on them and will not discover the larger patterns. We can see the similar result for the DTP dataset. The size of the patterns SUBDUE found are very small, 2–6 edges, but their frequency is very high. On the other hand, the results in Table 3 show that with the minimum frequency threshold 20, both HSIGRAM and VSIGRAM under exact MIS spend 249 and 83 seconds respectively to find 21,478 frequent subgraphs, and the largest size is 16.

6.3.2. Comparison with SEuS. The SEuS (Ghazizadeh and Chawathe, 2002b) algorithm is designed to find all frequent subgraphs in a single-graph setting. However, when determining the frequency of a subgraph they consider all embeddings irrespective of whether they are disjoint or not. As a result, a subgraph may have high frequency even though it has small number of edge-disjoint embeddings because of overlapped embeddings. In Ghazizadeh and Chawathe (2002b), the runtime of SEuS on the PTE chemical dataset[10] is reported. SEuS (SEuS-S1) spent more than 20 seconds to find 34 frequent subgraphs, that is 1.4 frequent subgraphs per second. On the same dataset given the minimum frequency threshold of 500, VSIGRAM with upper-bound MIS requires 20 seconds to find 168 frequent subgraphs, which translates to 8.4 frequent subgraphs per second. Similarly, with the Credit

dataset (which is called "Credit-4" in Ghazizadeh and Chawathe (2002a)), SEuS-S1 spent 50 seconds to produce 48 frequent subgraphs (one frequent subgraphs per second), while VSIGRAM with upper-bound MIS finds 1,325 frequent subgraphs in four seconds for the minimum frequency threshold 200 (331 frequent subgraphs per second).

7. Conclusions

In this paper we addressed the problem of finding all the subgraphs that have many edge-disjoint embeddings in a large sparse graph, a step critical to discovering patterns in graph datasets. We studied three distinct formulations of the problem that were motivated by the complexity of identifying the maximum set of edge-disjoint embeddings of a subgraph, and developed two frequent subgraph mining algorithms for solving them. These algorithms are based on the horizontal and vertical paradigms, respectively. Our experimental evaluation on many real datasets showed that for most datasets and problem formulations both algorithms achieve good performance, with the vertical algorithm being two-to-five times faster.

Notes

1. The symbol v_i in the figure is a vertex ID, not a vertex label, and blank elements in the adjacency matrix means there is no edge between the corresponding pair of vertices.
2. SiGraM stands for Single Graph Miner.
3. http://cygnus.uta.edu/subdue/databases/index.html
4. http://www.cs.cornell.edu/projects/kddcup/datasets.html
5. DTP 2D and 3D Structural Information. http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html
6. http://dip.doe-mbi.ucla.edu/
7. http://vlsicad.cs.ucla.edu/~cheese/ispd98.html
8. http://www.google.com/programming-contest/
9. Although this version is not the latest one, it runs significantly faster than the current latest version, 5.0.8.
10. ftp://ftp.comlab.ox.ac.uk/pub/Packages/ILP/Datasets/carcinogenesis/progol/carcinogenesis.tar.Z

References

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In Proc. of 1998 ACM SIGMOD International Conference on Management of Data, pp. 94–105.

Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc. of the 20th International Conference on Very Large Data Bases (VLDB), J.B. Bocca, M. Jarke, and C. Zaniolo (Eds.), Morgan Kaufmann, pp. 487–499.

Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In Proc. of the 11th International Conference on Data Engineering (ICDE), P.S. Yu and A.L.P. Chen (Eds.), IEEE Press, pp. 3–14.

Berendt, B., Hotho, A., and Stumme, G. 2002. Towards semantic web mining. In International Semantic Web Conference (ISWC), pp. 264–278.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H. et al. 2000. The protein data bank. Nucleic Acids Research, 28:235–242.

Berman, P. and Fujito, T. 1995. On the approximation properties of independent set problem in degree 3 graphs. In Proc. of Workshop on Algorithms and Data Structures, pp. 449–460.

Blake, C.L. and Merz, C.J. 1998. UCI Repository of Machine Learning Databases.

Borgelt, C. and Berthold, M.R. 2002. Mining molecular fragments: Finding relevant substructures of molecules. In Proc. of 2002 IEEE International Conference on Data Mining (ICDM), pp. 51–58.

Chew, L.P., Huttenlocher, D., Kedem, K., and Kleinberg, J. 1999. Fast detection of common geometric substructure in proteins. In Proc. of the 3rd ACM RECOMB International Conference on Computational Molecular Biology, pp. 104–114.

Cohen, M. and Gudes, E. 2004. Diagonally subgraphs pattern mining. In Proc. of the 9th ACM SIGMOD workshop on research issues in Data Mining and Knowledge Discovery DMKD '04, pp. 51–58.

Cook, D.J. and Holder, L.B. 1994. Substructure discovery using minimum description length and background knowledge. Journal of Artificial Intelligence Research, 1:231–255.

Cook, D.J. and Holder, L.B. 2000. Graph-based data mining. IEEE Intelligent Systems, 15(2):32–41.

Cook, D.J., Holder, L.B., and Djoko, S. 1995. Knowledge discovery from structural data. Journal of Intelligent Information Systems, 5(3):229–245.

Dehaspe, L., Toivonen, H., and King, R.D. 1998. Finding frequent substructures in chemical compounds. In Proc. of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-98), R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro (Eds.), AAAI Press, pp. 30–36.

De Raedt, L. and Kramer, S. 2001. The level-wise version space algorithm and its application to molecular fragment finding. In Proc. of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01), pp. 853–862.

Deshpande, M., Kuramochi, M., and Karypis, G. 2002. Automated approaches for classifying structures. In Proc. of the 2nd Workshop on Data Mining in Bioinformatics (BIOKDD '02) pp. 11–18.

Deshpande, M., Kuramochi, M., and Karypis, G. 2003. Frequent sub-structure based approaches for classifying chemical compounds. In Proc. of 2003 IEEE International Conference on Data Mining (ICDM), pp. 35–42.

Feige, U., Goldwasser, S., Lovasz, L., Safra, S., and Szegedy, M. 1991. Approximating clique is almost NP-complete. In Proc. of the 32nd IEEE Symposium on Foundations of Computer Science (FOCS), pp. 2–12.

Fortin, S. 1996. The Graph Isomorphism Problem (Tech. Rep. No. TR96-20). Department of Computing Science, University of Alberta.

Garey, M.R. and Johnson, D.S. 1979. Computers and Intractability: A Guide to the Theory of np-completeness. New York: W. H. Freeman and Company.

Ghazizadeh, S. and Chawathe, S. 2002a. Discovering Freuqent Structures Using Summaries (Tech. Rep. No. CS-TR-4364). Department of Computer Science, University of Maryland.

Ghazizadeh, S. and Chawathe, S. 2002b. SEuS: Structure extraction using summaries. In Proc. of the 5th International Conference on Discovery Science, pp. 71–85.

Gonzalez, J., Holder, L.B. and Cook, D.J. 2001. Application of graph-based concept learning to the predictive toxicology domain. In Proc. of the Predictive Toxicology Challenge Workshop.

Grindley, H.M., Artymiuk, P.J., Rice, D.W., and Willett, P. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. Journal of Molecular Biology, 229:707–721.

Guralnik, V. and Karypis, G. 2001. A scalabale algorithm for clustering sequence datasets. In Proc. of 2001 IEEE International Conference on Data Mining (ICDM).

Halldórsson, M.M., and Radhakrishnan, J. 1997. Greed is good: Approximating independent sets in sparse and bounded-degree graphs. Algorithmica, 18(1):145–163.

Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In Proc. of ACM SIGMOD International Conference on Management of Data Dallas, TX, pp. 1–12.

Hochbaum, D.S. 1983. Efficient bounds for the stable set, vertex cover, and set packing problems. Discrete Applied Mathematics, 6:243–254.

Holder, L.B., Cook, D.J., and Djoko, S. 1994. Substructure discovery in the SUBDUE system. In Proc. of the AAAI Workshop on Knowledge Discovery in Databases, pp. 169–180.

Hong, M., Zhou, H., Wang, W., and Shi, B. 2003. An efficient algorithm of frequent connected subgraph extraction. In Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-03), Springer-Verlag, Vol. 2637, pp. 40–51.

Huan, J., Wang, W., and Prins, J. 2003. Efficient mining of frequent subgraph in the presence of isomophism. In Proc. of 2003 IEEE International Conference on Data Mining (ICDM'03), pp. 549–552.

Inokuchi, A., Washio, T., and Motoda, H. 2000. An apriori-based algorithm for mining frequent substructures from graph data. In Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'00), Lyon, France, pp. 13–23.

Inokuchi, A., Washio, T., and Motoda, H. 2003. Complete mining of frequent patterns from graphs: Mining graph data. Machine Learning, 50(3):321–354.

Inokuchi, A., Washio, T., Nishimura, K., and Motoda, H. 2002. A Fast Algorithm for Mining Frequent Connected Subgraphs (Tech. Rep. No. RT0448). IBM Research, Tokyo Research Laboratory.

Jensen, D. and Goldberg, H. (Eds.). 1998. Artificial Intelligence and Link Analysis Papers from the 1998 Fall Symposium. AAAI Press.

Jonyer, I., Cook, D.J., and Holder, L.B. 2001a. Discovery and evaluation of graph-based hierarchical conceptual clusters. Journal of Machine Learning Research, 2:19–43.

Jonyer, I., Holder, L.B., and Cook, D.J. 2001b. Hierarchical conceptual structural clustering. International Journal on Artificial Intelligence Tools, 10(1–2):107–136.

Khanna, S., Motwani, R., Sudan, M., and Vazirani, U.V. 1994. On syntactic versus computational views of approximability. In Proc. of IEEE Symposium on Foundations of Computer Science, pp. 819–830.

Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632.

Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A.S. 1999. The Web as a graph: Measurements, models and methods. Lecture Notes in Computer Science, 1627:1–17.

Ko, C. 2000. Logic induction of valid behavior specifications for intrusion detection. In IEEE Symposium on Security and Privacy (S&P), pp. 142–155.

Koch, I., Lengauer, T., and Wanke, E. 1996. An algorithm for finding maximal common subtopoloties in a set of protein structures. Journal of Computational Biology, 3(2):289–306.

Kramer, S., De Raedt, L., and Helma, C. 2001. Molecular feature mining in HIV data. In Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01), pp. 136–143.

Kuramochi, M., Deshpande, M., and Karypis, G. 2004. Data mining: Next generation challenges and future directions. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.), AAAI Press (in press), pp. 315–334.

Kuramochi, M. and Karypis, G. 2001. Frequent subgraph discovery. In Proc. of 2001 IEEE International Conference on Data Mining (ICDM), pp. 313–320.

Kuramochi, M. and Karypis, G. 2002. An Efficient Algorithm for Discovering Frequent Subgraphs (Tech. Rep. No. 02-026). University of Minnesota, Department of Computer Science.

Kuramochi, M. and Karypis, G. 2004a. An efficient algorithm for discovering frequent subgraphs. IEEE Transactions on Knowledge and Data Engineering, 16(9):1038–1051.

Kuramochi, M. and Karypis, G. 2004b. Finding frequent patterns in a large sparse graph. In Proc. of the 2004 SIAM International Conference on Data Mining (SDM04).

Lee, W. and Stolfo, S.J. 2000. A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security, 3(4):227–261.

Leibowitz, N., Fligelman, Z.Y., Nussinov, R., and Wolfson, H.J. 1999. Multiple structural alignment and core detection by geometric hashing. In Proc. of the 7th international conference on intelligent systems in molecular biology, Heidelberg: Germany, pp. 169–177.

Leibowitz, N., Nussinov, R., and Wolfson, H.J. 2001. MUSTA—a general, efficient, automated method for multiple structure alignment and detection of common motifs: Application to proteins. Journal of Computational Biology, 8(2):93–121.

Li, W., Han, J., and Pei, J. 2001. CMAR: Accurate and efficient classification based on multiple class-association rules. In Proc. of 2001 IEEE International Conference on Data Mining (ICDM), pp. 369–376.

Liu, B., Hsu, W., and Ma, Y. 1998. Integrating classification and association rule mining. In Proc. of the 4th Internation Conference on Knowledge Discovery and Data Mining (kdd-98), pp. 80–86.

McKay, B.D. (n.d.). Nauty users guide. http://cs.anu.edu.au/∼bdm/nauty/

McKay, B.D. 1981. Practical graph isomorphism. Congressus Numerantium, 30:45–87.

Mitchell, E.M., Artymiuk, P.J., Rice, D.W., and Willett, P. 1989. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. Journal of Molecular Biology, 212:151–166.

Mooney, R.J., Melville, P., Tang, L.R., Shavlik, J., Castro Dutra, I. de, Page, D. et al. 2004. Relational data mining with inductive logic programming for link discovery. In AAAI Press/The MIT Press, pp. 239–255.

Muggleton, S.H. 1999. Scientific knowledge discovery using Inductive Logic Programming. Communications of the ACM, 42(11):42–46.

Östergård, P.R.J. 2002. A fast algorithm for the maximum clique problem. Discrete Applied Mathematics, 120:195–205.

Palmer, C.R., Gibbons, P.B., and Faloutsos, C. 2002. ANF: A fast and scalable tool for data mining in massive graphs. In Proc. of the 8th ACM SIGKDD Internal Conference on Knowlege Discovery and Data Mining (KDD-2002) Edmonton, AB, Canada, pp. 81–90.

Pennec, X. and Ayache, N. 1998. A geometric algorithm to find small but highly simialar 3D substructures in proteins. Bioinformatics, 14(6):516–522.

Raymond, J.W. 2002. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. J. Chem. Inf. Comput. Sci., 42:305–316.

Read, R.C. and Corneil, D.G. 1977. The graph isomorph disease. Journal of Graph Theory, 1:339–363.

Robson, J.M. 1986. Algorithms for maximum independent sets. Journal of Algorithms, 7:425–440.

Srinivasan, A., King, R.D., Muggleton, S.H., and Sternberg, M.J.E. 1997. Carcinogenesis predictions using ILP. In Proc. of the 7th International Workshop on Inductive Logic Programming S. Džeroski and N. Lavrač (Eds.), Springer-Verlag, (Vol. 1297, pp. 273–287).

Vanetik, N., Gudes, E., and Shimony, S.E. 2002. Computing frequent graph patterns from semistructured data. In Proc. of 2002 IEEE International Conference on Data Mining (ICDM), pp. 458–465.

Wang, X., Wang, J.T.L., Shasha, D., Shapiro, B.A., Rigoutsos, I., and Zhang, K. 2002. Finding patterns in three dimensional graphs: Algorithms and applications to scientific data mining. IEEE Transactions on Knowledge and Data Engineering, 14(4):731–749.

Wasserman, S., Faust, K., and Iacobucci, D. 1994. Social network analysis : Methods and applications. Cambridge University Press.

Yan, X. and Han, J. 2002. gSpan: Graph-based substructure pattern mining. In Proc. of 2002 IEEE International Conference on Data Mining (ICDM), pp. 721–724.

Yan, X. and Han, J. 2003. CloseGraph: Mining closed frequent graph patterns. In Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), pp. 286–295.

Yoshida, K. and Motoda, H. 1995. CLIP: Concept learning from inference patterns. Artificial Intelligence, 75(1):63–92.

Yoshida, K., Motoda, H., and Indurkhya, N. 1994. Graph-based induction as a unified learning framework. Journal of Applied Intelligence, 4:297–328.

Zaki, M.J. and Gouda, K. 2003. Fast vertical mining using diffsets. In Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), pp. 326–335.