A New Polynomial-time Support Measure for Counting Frequent Patterns in Graphs

Jinghan Meng, Yi-Cheng Tu, and Napath Pitaksiranan

University of South Florida

Tampa, Florida

{jmeng,tuy,napath}@mail.usf.edu

ABSTRACT

Frequent subgraph mining (FSM) from graphs is an active subject in computer science research. One major challenge in FSM is the development of support measures, which are basically functions that map a pattern to its frequency count in a database. Current state-of-the-art in this topic features a hypergraph-based framework for modeling pattern occurrences which unifies the two main flavors of support measures: the overlap-graph based maximum independent set measure (MIS) and minimum image/instance based (MNI) measures. For the purpose of exploring the middle ground between these two groups, we introduce a new polynomial-time support measure, called maximum independent subedge set (MISS) measure to fill the gap between MIS and MI in terms of computation complexity and support count. Bounding theorems among all relevant support measures are also presented in this paper.

CCS CONCEPTS

• **Information systems** → *Graph-based database models*; • **Mathematics of computing** → *Hypergraphs*; *Graph algorithms*.

KEYWORDS

Graph mining, support measure, hypergraph

ACM Reference Format:

Jinghan Meng, Yi-Cheng Tu, and Napath Pitaksiranan. 2019. A New Polynomial-time Support Measure for Counting Frequent Patterns in Graphs. In *Proceedings of International Conference on Scientific and Statistical Database Management (SSDBM'19)*. ACM, New York, NY, USA, Article 4, 4 pages.

1 INTRODUCTION

The field of graph mining has attracted a lot of attention with applications in social networks, the web, and aviation maps. One of the important problems is frequent subgraph mining (FSM) [2], which aims to identify frequent patterns for the purpose of studying trend or locating hotspots. In such efforts, an essential component is the support measure of the given pattern, i.e., the frequency the pattern appears in the graph database. In this paper, we report our recent work related to support measures in the *single-graph* setting, in which the database is only one (often large) graph.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SSDBM'19, July 2019, Santa Cruz, CA, USA

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

DOI: 10.1145/3335783.3335810

Defining support measures is a non-trivial matter. First of all, a support measure has to satisfy the so-called *anti-monotonicity* property, which states that the frequency of a pattern should not exceed that of any of its subpatterns. This rule is essential to the majority of FSM algorithms because they rely on it to safely prune infrequent patterns from a tree structure to save time. Other facts to consider include computational cost and intuitive frequency counts.

The first anti-monotonic support measure is called the *maximum independent set based* (MIS) support [8]. The main idea of MIS is to count the maximum number of independent instances of a pattern from a data structure named *overlap graph*. While providing an intuitive count, MIS shows a major drawback in its lack of efficient algorithms – it is proved to be NP-hard that makes MIS infeasible to compute in even moderately sized graph datasets.

Based on the technique of vertex images, another type of support measure named the minimum-image-based (MNI) support [1] was designed. As another anti-monotonic support, MNI is linear-time computable. However, MNI falls short in *intuitiveness* – it could arbitrarily overestimate the frequency of a pattern. This is due to its lack of consideration of topological structure of the query pattern.

Our previous work [6] introduced a general framework based on the concept of *occurrence and instance hypergraph* to model many support measures. In this framework, MIS is shown to be equivalent to a *maximum independent edge set* (MIES) support, and MNI can be directly interpreted as a special case of the so-called *minimum-instance-based measure* (MI). Under this framework, MNI and MIS/MIES can be seen as the upper and lower bounds of the frequency spectrum among state-of-the-art measures. Moreover, another measure called the *minimum vertex cover measure* (MVC) is derived as an ultimate version of MNI. MVC is also connected with MIS through the concepts of dual hypergraph. Polynomial-time linear programming relaxations of MIES and MVC, denoted as RMIES and RMVC, stand in-between MIS and MVC in terms of count value. Bounding theorems that compare the frequency counts returned by aforementioned support measures are given as:

$$MIS = MIES \leq RMIES = RMVC \leq MVC \leq MI \leq MNI \quad (1)$$

Our Contributions: Based on our previous work [6], we sense the opportunity to abstract out even more fundamental rules in this topic. In this paper, we present a new support measure named **maximum independent subedge set (MISS) support** with desirable intuitiveness and efficiency. Note that computing MIS/MIES and MVC are NP-hard, and that for MI and MNI are linear. It is an interesting problem as to whether there exist support measures between MIES/MVC and MI/MNI in terms of computational complexity and counts. RMIES and RMVC are such measures however their computational complexity is still of high-order polynomial. MISS fills

the aforementioned gap: it is proved to be anti-monotonic, returns counts between MIS and MI, and can be computed in low-order polynomial time.

Paper Organization: In Section 2, we show a sketch of the necessary background for the problem; In Section 3, we introduce our new MISS support measure and study its properties; In Section 4 we summarize our findings and discuss opportunities for future work.

2 PRELIMINARIES

In this section, we introduce basic notations and concepts to describe the problem and the necessary background.

2.1 Labeled Graphs

The graph considered in this paper is vertex-labeled graph, which is simply referred to as *graph* hereafter. In all figures of this article, we represent the label of a vertex by the shade.

DEFINITION 1. A (undirected) **labeled graph**

$$G = (V_G, E_G, \lambda_G)$$

consists of a set of vertices V_G , a set of edges $E_G \subseteq V_G \times V_G := \{(u, v) \mid u, v \in V_G, u \neq v\}$ and a labeling function $\lambda_G : V_G \rightarrow \Sigma$ that maps each vertex of the graph to an element of the alphabet Σ . We use $\mathcal{G}$ to denote the class of all graphs.

DEFINITION 2. A **pattern** $p = (V_p, E_p, \lambda_p)$ is a labeled graph we use as a query against another graph.

DEFINITION 3. Let P be a graph pattern, and p a subgraph of P , denoted by $p \subseteq P$. We call p a **subpattern** of P , and likewise, we call P a **superpattern** of p .

After knowing that copies of pattern p exist in data graph G , we are focused on how many times it appears in G . For that, we first need to define the concepts of *occurrence* and *instance* of the pattern in the data graph.

DEFINITION 4. Given a pattern p and a data graph G , an **occurrence** f is an isomorphism (an edge-preserving bijection between the vertex sets of two graphs) between pattern p and a subgraph of G . That is to say f is also a subgraph isomorphism from p to G .

DEFINITION 5. Given a pattern p and a data graph G , a subgraph S of G is an **instance** of pattern p in G when there exists an isomorphism between p and S .

2.2 Overlap Concepts and Support Measure

The purpose of defining support measure is to count the appearances of a pattern p in a data graph G . The definition of support measure is given below:

DEFINITION 6. A **support measure** of pattern p in data graph G is a function $\sigma : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}^+$, which maps (p, G) to a non-negative number $\sigma(p, G)$.

Naturally, one would pick the number of occurrences or instances of a pattern as support measure. Unfortunately, such choices violate the *anti-monotonicity* rule [4, 8]. For example, in Figure 1, pattern p has four occurrences f_1, f_2, f_3, f_4 in data graph, but its superpattern P has five occurrences $f'_{1,1}, f'_{1,2}, f'_{1,3}, f'_2, f'_3$.

DEFINITION 7. A support measure σ of pattern p in G is **anti-monotonic** if for any pattern p and its superpattern P , we have $\sigma(p, G) \geq \sigma(P, G)$.

DEFINITION 8. **Vertex overlap:** A vertex overlap of occurrences f_1 and f_2 of pattern $P = (V_P, E_P)$ in data graph $G = (V, E)$ exists if vertex sets $f_1(V_P)$ and $f_2(V_P)$ intersect, that is, $f_1(V_P) \cap f_2(V_P) \neq \emptyset$ where $f_i(V_P) = \{f_i(v) : v \in V_P\}$, $i = 1, 2$. A vertex overlap of instances $S_1 = (V_{S_1}, E_{S_1})$ and $S_2 = (V_{S_2}, E_{S_2})$ of pattern P exists if vertex sets of S_1 and S_2 intersect, that is, $V_{S_1} \cap V_{S_2} \neq \emptyset$.

In this paper, we consider two occurrences **overlap** when they overlap at one or more than one vertex. For a pattern $p = (V_p, E_p)$, a node set $T \subseteq V_p$, and occurrences f, f' of p in data graph G , the term **image** refers to vertex set $\{f(v)\}_{v \in T}$. Existing support measures are designed to count the images of V_p in various ways.

A recent work [6] developed a framework that unifies all existing measures (including MIS and MNI), and provides a foundation for introducing and analyzing new support measures. This framework is built on the concept of *hypergraph*.

DEFINITION 9. A **hypergraph** $H = (V, E)$ consists of a set $V = \{v_1, v_2, \dots, v_n\}$ of n vertices and a set $E = \{e_1, e_2, \dots, e_m\}$ of m edges, where each edge is a non-empty subset of V .

DEFINITION 10. Assume that pattern $p = (V_p, E_p)$ has m occurrences $\{f_i\}_{1 \leq i \leq m}$, the **occurrence hypergraph** of p in G is defined as $H_p = (V, E)$ where $V = \cup_{1 \leq i \leq m} f_i(V_p)$, and $E = \{e_i\}_{1 \leq i \leq m}$, each $e_i = f_i(V_p)$. In other words, hypergraph vertex set V is the collection of all pattern node images, and each edge e_i is a collection of pattern node images mapped by occurrence f_i . For each e_i , a label f_i is given to distinguish it from others.

For example, in Figure 1, occurrence hypergraphs of p and P are shown as H_p and H_P . Each edge e_i in H_p corresponds to occurrence f_i , $i = 1, 2, 3, 4$. In H_P , edges $e'_{1,1}, e'_{1,2}, e'_{1,3}, e'_2, e'_3$ correspond to $f'_{1,1}, f'_{1,2}, f'_{1,3}, f'_2, f'_3$ respectively.

3 DEVELOPMENT OF NEW MEASURES

3.1 A New Polynomial Time Support

Within the hypergraph framework a series of support measures is defined/redefined including the MIS/MIES and MNI [6]. MIS and MNI represent the two main categories of support measures yet they stand on far ends of computing efficiency and intuitiveness. Here we want to emphasize that intuitiveness is a highly application-specific concept. However, in most (if not all) of the papers in this field, discussions are done in a general setup without considering application-specific requirements, and intuitiveness is defined as **avoidance of overestimation** of a pattern's frequency count [9]. As a special note, MIS is considered to be the most *intuitive* measure as it returns the number of independent occurrences. In a statistical sense, a measure should be related to the number of independent observations of a phenomenon (the pattern) [9]. Thus, MIS is often used as a lower bound to benchmark non-trivial support measures.¹ However, it is infeasible to compute the NP-hard MIS for even moderately sized inputs. On contrary to that,

¹A support measure that returns an arbitrarily small number (such as 0) for all inputs is trivial and not interesting.

Figure 1: An example of occurrence hypergraphs generated from two patterns over a data graph. Occurrences and hypergraph edges of the superpattern are marked red

the MNI has linear complexity but can return an arbitrarily large count for a pattern [1]. According to Eq. (1), the MVC and MI we developed fill in between MIS and MNI, providing more options for the efficiency/intuitiveness tradeoff. Linear programming relaxations of MIS and MVC further reduce the frequency counts and only require polynomial time. However, RMVC is still not scalable - the best solution is an interior-point-method algorithm which runs at $O(n^{3.5}L)$ complexity, where n is the number of vertices and L is the number of bits in the input [3]. Our goal is to design a more efficient support measure that returns frequency counts that are at least as close to MIS as RMVC is.

3.1.1 The MISS Measure. Consider the fact that MIES counts only independent occurrences and $MIES \leq MVC$, we shall choose MIES as the starting point. This is important, because it means that the subedge version of MIES will most likely be bounded by MI/MNI which are subedge versions of MVC.

Following the above ideas, we propose a maximum independent subedge set (MISS) measure. This new support measure is also supported by the following observation: if two hypergraph edges are independent to each other (i.e., no vertex intersections), any subedges (subsets) within these two edges should also be independent. We hereby develop a formal definition of the MISS measure and conduct rigorous analysis on its features.

DEFINITION 11. Given a pattern p and a data graph G , in hypergraph framework, the **maximum independent subedge set (MISS) support measure** is defined as

$$\sigma_{MISS}(p, G) = \min_{w \in W_p} d_{H_p}(w), \quad (2)$$

where $W_p = \{\{u, v\} : u, v \in V_p\}$, $d_{H_p}(w)$ is the size of maximum independent edge set of images of w under edges of H_p .

Figure 2: An example of FSM problem, in which MISS = MIS = 1, and MI ≥ 2

An example of MISS is given in Figure 2. The red edges are images of node pair $\{v_1, v_3\}$ (which is not a pattern edge), blue edges are images of node pairs (pattern edges) $\{v_1, v_2\}$ and $\{v_2, v_3\}$. $d(\{v_1, v_3\}) = 1$. $d(\{v_1, v_2\}) = d(\{v_2, v_3\}) = 2$, thus $MISS = \min\{1, 2, 2\} = 1$. On the other hand, we go through all combinations of node subsets, all their image counts ≥ 2 , thus MI is at least 2.

3.1.2 Key Properties of MISS. We shall show that MISS satisfies the anti-monotonicity requirement and is polynomial-time computable.

THEOREM 1. The MISS support measure is anti-monotonic.

PROOF. Given pattern $p = (V_p, E_p)$ and its superpattern $P = (V_P, E_P)$ in data graph G . Assume that there are m occurrences $\{f_i\}_{1 \leq i \leq m}$ of pattern p , and they can be extended to l occurrences $\{f'_i\}_{1 \leq i \leq l}$ of P . Since $V_p \subseteq V_P$, we have $W_p \subseteq W_P$ by definition. For any $w \in W_p \subseteq W_P$, because each f'_i is extension of some f_j , we have $\{f'_i(w)\}_{1 \leq i \leq l} \subseteq \{f_i(w)\}_{1 \leq i \leq m}$. Hence if we denote the cardinalities of maximum independent edge set of $\{f'_i(w)\}_{1 \leq i \leq l}$ and $\{f_i(w)\}_{1 \leq i \leq m}$ as $d_{f'}(w)$ and $d_f(w)$ respectively, we have $d_{f'}(w) \leq d_f(w)$.

$$\begin{aligned} \sigma_{MISS}(P, G) &= \min_{w \in W_P} d_{f'}(w) && \text{by definition} \\ &\leq \min_{w \in W_p} d_{f'}(w) && \text{because of } W_p \subseteq W_P \\ &\leq \min_{w \in W_p} d_f(w) = \sigma_{MISS}(p, G) && \text{by definition.} \end{aligned}$$

Hence we obtain the anti-monotonicity of MISS. $\square$

THEOREM 2. The MISS support measure is of polynomial time complexity in the number of pattern occurrences.

PROOF. By definition, $\sigma_{MISS}(p, G) = \min_{w \in W_p} d_{H_p}(w)$. Assume p has m occurrences $\{f_i\}_{1 \leq i \leq m}$. For $w \in W_p$, the images of w are $\{f_i(w)\}_{1 \leq i \leq m}$, because each $f_i(w)$ contains two vertices, they can be viewed as general graph edges. According to [7], there exists fast $O(\sqrt{nm})$ algorithm for finding maximum independent edge set (maximum matching) in general graphs, where n is the number of vertices and m is the number of edges in general graphs. For p that has k nodes, and H_p that has n vertices and m hyperedges, there are $\frac{k(k-1)}{2}$ node pairs in W_p . Thus the complexity of finding $d_{H_p}(w)$ is $O(\sqrt{nm})$ for each $w \in W_p$. The number of $w \in W_p$ is fixed (in terms of k), this concludes the proof. $\square$

Comparing Efficiency of MISS with RMVC. An important note is, $O(n^{3.5}L)$ time (L is the number of bits of the input to the algorithm) is needed for computing RMVC, as compared to the $O(\sqrt{nm})$ time for MISS. In the worse case ($m = n^2$), MISS runs at $O(n^{2.5})$ but in

practice we have $m \ll n^2$. The significant difference in computational performance between MISS and RMVC can be also seen from our experiments (in [5]).

3.2 New Bounding Theorems

We study the frequency counts returned by the newly-defined measures in comparison to existing ones. We first compare MISS with MIES, which is equivalent to MIS.

It is clear that if two hypergraph edges e and e' do not overlap then subedges (subsets) of e do not overlap with that of e' . Hence MIES is a lower bound of MISS, that is, $\sigma_{MIES}(p, G) \leq \sigma_{MISS}(p, G)$. The MISS measure is derived from maximum independent edge set, while MI and MNI are rooted at the minimum image based idea. This confirms the finding in the previous work [6] that MIES/MIS is the natural lower bound of support measures defined within the hypergraph framework. On the other hand, it is critical to study whether MISS is bounded by MI support measures - the value of MISS will be questionable if it finds larger counts than MI as the latter is of linear complexity.

THEOREM 3. *For pattern p in data graph G , we have*

$$\sigma_{MISS}(p, G) \leq \sigma_{MI}(p, G).$$

PROOF. Assume that p has m occurrences $\{f_i\}_{1 \leq i \leq m}$ in data graph G . We shall show that this inequality is true for any $\sigma_{MI}(p, G) = \min_{T \in \mathcal{T}} c_{H_p}(T)$ with $\mathcal{T} = \{T\}$, where T is a subset of V_p , $c_{H_p}(T)$ is the count of images of T under edges in H_p .

Case I: If T contains only one node u . Since in this paper we consider patterns containing more than one node, we can find another node v to form a node pair $w' = \{u, v\}$. Assume a maximum independent edge set of w' images under all occurrences is $I = \{f_i(w')\}_{1 \leq i \leq l} = \{\{f_i(u), f_i(v)\}\}_{1 \leq i \leq l}$, where $l \leq m$. It follows that $\{f_i(u)\}_{1 \leq i \leq l}$ are a set of distinct vertices. Hence $d_{H_p}(w') \leq c_{H_p}(T)$, and $\sigma_{MISS}(p, G) = \min_{w \in W_p} d_{H_p}(w) \leq d_{H_p}(w') \leq c(T)$.

Case II: If T contains two or more nodes, for any two nodes $u, v \in V_p$, let $w' = \{u, v\}$. Assume a maximum independent edge set of their images is $I = \{f_i(w')\}_{1 \leq i \leq l}$, where $l \leq m$. Because $f_i(w') \subseteq f_i(T)$, $1 \leq i \leq l$, and $f_i(w') \cap f_j(w') = \emptyset$, for $i \neq j$, we obtain that $f_i(T)$ and $f_j(T)$ are pairwise different for $1 \leq i, j \leq l$ and $i \neq j$. Hence, we have

$$\begin{aligned} \sigma_{MISS}(p, G) &= \min_{w \in W_p} d_{H_p}(w) \leq d_{H_p}(w') = l = |\{f_i(T)\}_{1 \leq i \leq l}| \\ &\leq |\{f_i(T)\}_{1 \leq i \leq m}| = c_{H_p}(T). \end{aligned}$$

Putting all together, we have $\sigma_{MISS}(p, G) \leq c_{H_p}(T)$ for any $T \in \mathcal{T}$. To conclude, $\sigma_{MISS}(p, G) \leq \min_{T \in \mathcal{T}} c_{H_p}(T) = \sigma_{MI}(p, G)$. $\square$

Theorem 3 tells us that no matter how the node subsets in a pattern are formed in a MI, the MISS count is always smaller. This also means that MISS provides a new level of efficiency-intuitiveness trade-off - MISS will not overestimate pattern frequency as much as MI does under the cost of longer computing time. With such results, the current spectrum of support counts with all measures mentioned in this paper is shown in Figure 3.

Note that there is no bound between MISS and RMVC (or MVC). However, MISS is found to be **as close to MIS as RMVC is**. This

$$\sigma_{MIS/MIES} \leq \sigma_{RMVC} \leq \sigma_{MVC} \leq \sigma_{MI}$$

Figure 3: New bounding theorems

can be seen from two aspects. First, in special cases where the pattern only contains one edge (i.e., two nodes), MISS values are found to be identical to MIS/MIES and thus better than RMVC. Second, MISS values are smaller than or equal to RMVC in most other scenarios, as shown by experiments running on real graph datasets (see [5] for more details).

In summary, MISS provides better efficiency and intuitiveness tradeoff than RMVC: it returns counts that are at least no larger than that of RMVC but requires asymptotically less time to compute.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we introduce the first non-relaxation polynomial-time frequent support measure named MISS which fills the gap between the NP-hard MIES/MIS and linear-time computable MI and MNI measures. For that, we believe this is a breakthrough in the topic of support measures in FSM. Analytical (and experimental [5]) results show that MISS significantly outperforms the best-known polynomial measure RMVC in running time and returns counts no larger than RMVC.

There are many opportunities for future research. Some ideas are: (1) more established graph combinatorics theorems can be incorporated into existing efforts in designing support measures; and (2) if certain support measures are absolutely needed while they have high computational complexity, parallel computing becomes a viable practical solution; (3) the increasing interest in web and social networks necessitates future research on support counting techniques that handle dynamic and streaming graph data sets.

REFERENCES

- [1] Björn Bringmann and Siegfried Nijssen. 2008. What is frequent in a single graph?. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 858–863.
- [2] Chuntao Jiang, Frans Coenen, and Michele Zito. 2013. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review* 28, 1 (2013), 75–105.
- [3] Narendra Karmarkar. 1984. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. ACM, 302–311.
- [4] Michihiro Kuramochi and George Karypis. 2005. Finding frequent patterns in a large sparse graph. *Data mining and knowledge discovery* 11, 3 (2005), 243–271.
- [5] Jinghan Meng, Napath Pitaksiranan, and Yicheng Tu. 2018. *Generalizing Design of Support Measures for Counting Frequent Patterns in Graphs*. Technical Report CSE/18-002, URL: www.csee.usf.edu/~tuy/pub/tech18-002.pdf. CSE Dept., University of South Florida, Tampa, FL, USA.
- [6] Jinghan Meng and Yi-cheng Tu. 2017. Flexible and Feasible Support Measures for Mining Frequent Patterns in Large Labeled Graphs. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17)*. ACM, New York, NY, USA, 391–402. <https://doi.org/10.1145/3035918.3035936>
- [7] Silvio Micali and Vijay V Vazirani. 1980. An $O(\sum |V| \sum |E|)$ algorithm for finding maximum matching in general graphs. In *Foundations of Computer Science, 1980., 21st Annual Symposium on*. IEEE, 17–27.
- [8] Natalia Vanetik, Ehud Gudes, and Solomon E. Shimony. 2002. Computing frequent graph patterns from semistructured data. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings*. 458–465.
- [9] Yuyi Wang and Jan Ramon. 2012. An efficiently computable support measure for frequent subgraph pattern mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 362–377.