What is Frequent in a Single Graph?

Björn Bringmann and Siegfried Nijssen

K.U. Leuven, Celestijnenlaan 200 A, B-3001 Leuven, Belgium
{Bjoern.Bringmann,Siegfried.Nijssen}@cs.kuleuven.be

Abstract. The standard, *transactional* setting of pattern mining assumes that data is subdivided in transactions; the aim is to find patterns that can be mapped onto at least a minimum number of transactions. However, this setting can be hard to apply when the aim is to find graph patterns in databases consisting of large graphs. For instance, the web, or any social network, is a single large graph that one may not wish to split into small parts. The focus in network analysis is on finding structural regularities or anomalies in one network, rather than finding structural regularities common to a *set* of them. This requires us to revise the definition of key concepts in pattern mining, such as support, in the *single-graph* setting. Our contribution is a support measure that we prove to be computationally less expensive and often closer to intuition than other measures proposed. Further we prove several properties between these measures and experimentally validate the efficiency of our measure.

Keywords: graph mining, pattern mining, network analysis.

1 Introduction

The traditional *transactional* setting of pattern mining is popular for many types of data with well-known examples being basket analysis [1], or molecular fragment mining [10]. As stated in the abstract, the single-graph setting introduces problems that do not appear in the transactional setting; the most prominent one being the definition of the support of a pattern. Naïve definitions of support have the problem that they are not anti-monotonic; thus they cannot be used effectively in pattern mining, as anti-monotonicity is required to prune the search space. To address this problem, Kuramochi and Karypis [7] as well as Fiedler and Borgelt [5] studied anti-monotonic support measures based on computing maximum independent sets (*MIS*) in overlap graphs.

Anti-monotonicity is however not the only requirement for efficient frequent pattern mining. It is also important that the frequency measure can be evaluated efficiently. We argue that the computation of overlap-based support measures is not feasible in many graph databases, and that more scalable support measures are needed to enable the use of frequent graph mining algorithms on network data. We propose a new support measure, and provide practical and theoretical evidence that this measure is more scalable, more general and more widely applicable than the support measures mentioned earlier. We show how this measure, and the overlap-graph based measures, relate to each other, thus providing deeper understanding in support measures for graph mining.

2 The Support of a Pattern

A labeled graph $g = (\mathbb{V}_g, \mathbb{E}_g, \lambda_g)$ consists of a set of nodes $\mathbb{V}_g$, a set of edges $\mathbb{E}_g \subseteq \mathbb{V}_g \times \mathbb{V}_g$ and a labeling function $\lambda_g : \mathbb{V}_g \cup \mathbb{E}_g \to \Sigma$ that maps each element of the graph to an element of the alphabet Σ. Let G_Σ be the set of all graphs over the alphabet Σ. We define support as a function $\sigma : G_\Sigma \times G_\Sigma \to \mathbb{N}$.

As stated earlier, minimum support needs to be anti-monotonic to allow efficient search. This means that for all graphs g, p and p', where p is a subgraph of p', it must hold that $\sigma(p, g) \geq \sigma(p', g)$. Anti-monotonicity is quite easily upheld in the transactional setting, but is more tricky for the *single-graph* setting. The cause of this problem is that it is not clear what exactly should be counted.

Occurrence of a Pattern Given a pattern $p = (\mathbb{V}_p, \mathbb{E}_p, \lambda_p)$ and a data graph $g = (\mathbb{V}_g, \mathbb{E}_g, \lambda_g)$, an occurrence is a function $\varphi : \mathbb{V}_p \to \mathbb{V}_g$ mapping the nodes of p to the nodes in g such that (I) $\forall v \in \mathbb{V}_p \Rightarrow \lambda_p(v) = \lambda_g(\varphi(v))$ and (II) $\forall (u, v) \in \mathbb{E}_p \Rightarrow (\varphi(u), \varphi(v)) \in \mathbb{E}_g$. The image of a set of nodes in an occurrence is denoted $\varphi(\mathbb{V}_p) = \{\varphi(v) | v \in \mathbb{V}_p\}$; similarly, we define the image of a set of edges.

The problem of the support measure on a single graph is explained in Figure 1. p_1 has one occurrence in g, and p_2 is a specialization of p_1. What is the support of p_2 in g? In the transactional setting every instance with at least one occurrence counts. This is undesirable in the single-graph setting, since every graph would have a support of either zero or one. A naïve measure that assigns a support of 2 in our example, would not be anti-monotonic.

Single Graph Support Measures For the support measure introduced in [7], all possible occurrences φ_i of a pattern p in the graph g are calculated. An *overlap-graph* is constructed where each occurrence φ_i corresponds to a node and there is an edge (φ_j, φ_k) iff $\varphi_j(\mathbb{E}_p) \cap \varphi_k(\mathbb{E}_p) \neq \emptyset$ (i.e.: φ_j and φ_k *share* an edge). The support for the pattern p is defined as the size of the maximum independent set (*MIS*) of the overlap-graph. For example, in Figure 2 there would be three occurrences of the pattern p in the graph g. Even though [7] defined overlap in terms of edges, the concept can also be applied to vertices. For this case, we formalize the following binary relationship:

Definition 1. *A* simple overlap *of occurrences φ and φ' of pattern p exists if* $\varphi(\mathbb{V}_p) \cap \varphi'(\mathbb{V}_p) \neq \emptyset$.

We denote the support measure based on *simple overlap* as $\sigma_\bullet$. It can be shown that this support measure is anti-monotonic. However, solving a *MIS* problem is NP-complete.

A refinement of the *simple overlap* based support measure was introduced in [5] and named *harmful overlap*. We will denote this by $\sigma_\circ$. The basic idea of this measure is that some of the simple overlaps can be disregarded without harming the anti-monotonicity of the support measure. As before, an overlap graph is constructed and the support is defined as the size of the *MIS*. Different is the definition of overlap:

Definition 2. *A* harmful overlap *of occurrences φ and φ' of pattern p exists if* $\exists v \in \mathbb{V}_p : \varphi(v), \varphi'(v) \in \varphi(\mathbb{V}_p) \cap \varphi'(\mathbb{V}_p)$.

Note that both *simple overlap* $\sigma_\bullet$ and *harmful overlap* $\sigma_\circ$ are based on shared nodes here. However, *All* measures can be used either based on *edges* or on *nodes*.

Both measures rely on computing an overlap graph, and subsequently solving a *MIS* problem. We propose a measure of support, which avoids potentially expensive *MIS* computations[1]. It is based on the number of unique nodes in the graph $g = (\mathbb{V}_g, \mathbb{E}_g)$ to which a node of the pattern $p = (\mathbb{V}_p, \mathbb{E}_p)$ is mapped.

Definition 3. *The minimum image based support of p in g is defined as*

$$\sigma_\wedge(p, g) = \min_{v \in \mathbb{V}_p} |\{\varphi_i(v) : \varphi_i \text{ is an occurrence of } p \text{ in } g\}|$$

By taking the node in p which is mapped to the least number of unique nodes in g, we can ensure the anti-monotonicity of $\sigma_\wedge$. From our definition of support, we can deduce several computational benefits: **(i)** instead of $O(n^2)$ potential overlaps, where n is the possibly exponential number of occurrences, we only need to maintain a set of data vertices for every node in the pattern, which can be done in $O(n)$; **(ii)** we do not need to solve an NP complete *MIS* problem; **(iii)** it is not necessary to compute all occurrences: it is sufficient to determine for every pair of $v \in \mathbb{V}_p$ and $v' \in \mathbb{V}_g$ if there is *one* occurrence in which $\varphi(v) = v'$. The computational burden can be reduced further by taking into account the automorphisms of the pattern graph.

Relationships and Dependencies All measures introduced are based on the occurrence of patterns, but they can give different results on the same data. An example for how the three measures work and that they give different results can be found in Figure 2.

Nevertheless, several relationships between these measures hold. We can show that our measure $\sigma_\wedge$ is an upper bound for the harmful overlap measure $\sigma_\circ$, which is in turn an upper bound for the simple overlap based measure $\sigma_\bullet$.

Theorem 1. *$\sigma_\wedge$ is an upper bound for $\sigma_\circ$: $\forall p \in \mathcal{P} : \sigma_\wedge(p, \mathcal{T}) \geq \sigma_\circ(p, \mathcal{T})$*

Proof. Let $v^* = \arg\min_{v \in \mathbb{V}_p} |\{\varphi_i(v) : \varphi_i \text{ an occurrence of } p \text{ in } \mathcal{T}\}|$. Then we know that $\forall \varphi, \varphi' : \varphi(v^*) = \varphi'(v^*)$ there is a harmful overlap of φ and φ' and hence at most one of the occurrences φ and φ' can be a member of the *MIS*. From this the claim follows. □

Theorem 2. *$\sigma_\circ$ is an upper bound for $\sigma_\bullet$: $\forall p \in \mathcal{P} : \sigma_\circ(p, \mathcal{T}) \geq \sigma_\bullet(p, \mathcal{T})$*

Proof. We know that for all φ, φ' such that φ and φ' overlap harmfully, there is a simple overlap. Hence the overlap graphs for both measures have the same nodes, and the edges for the harmful overlap are a subset of the edges for the simple overlap. Thus, the harmful overlap contains less constraints for the *MIS*, and the set is at least as big as for the simple overlap. □

Finally it is easy to see that all described measures are bounded by the real number of pattern occurrences in the graph.

[1] This paper is an extended version of a paper presented at a workshop without publications [2].

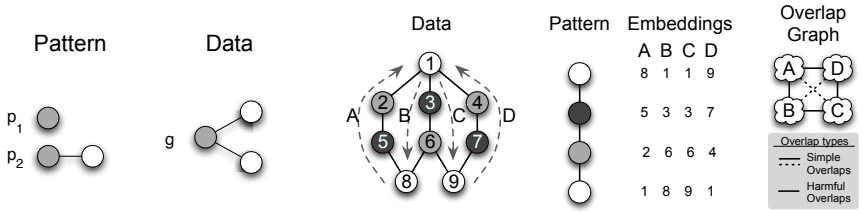

Fig. 1. The support problem in a single graph g: p_1 occurs once. How often occurs p_2?

Fig. 2. A graph with four different occurrences of a pattern. The three discussed measures evaluate to $\sigma_\bullet = 1 < \sigma_\circ = 2 < \sigma_\wedge = 3$.

Pattern	1	2	3	4	5	6	7	8	9	10
Nodes in Pattern	2	2	3	3	3	4	4	4	4	5
Image-based support	110	110	100	95	77	97	68	77	64	82
# Occurrences	432	418	1696	1606	815	5428	7380	2254	816	15878
Time for Occurrences	≈ 0s	≈ 0s	≈ 0s	≈ 0s	≈ 0s	≈ 0s	≈ 0s	≈ 0s	≈ 0s	≈ 0s
Edges in overlap-graph	3825	8714	328925	226886	167026	6662049	8362729	1401907	265249	66155623
Time for *MIS*	1s	1s	31s	57s	4s	958s	2456s	73s	2s	>45m
Size *MIS*	69	92	42	65	36	24	45	32	62	-

Pattern	11	12	13	14	15	16	17	18	19	20
Nodes in Pattern	5	6	7	8	9	10	11	12	13	14
Image-based support	68	80	69	69	66	66	63	63	62	62
# Occurrences	7988	44254	116580	287954	658540	1386328	2711828	5039624	9125850	16409046
Time for Occurrences	≈ 0s	2s	7s	22s	1m4s	2m56s	7m34s	18m57s	46m14s	110m49s
Edges in overlap-graph	9332671	-	47804219	-	-	-	-	-	-	-

Table 1. Details of the computations needed to determine the *MIS* support measures. *MIS* could not be computed for pattern 11 and above.

3 Experiments

To compare with the overlap-based support measures from [7, 5] we obtained the datasets *Aviation* and *Credit* described in [7] from the SUBDUE website[2]. We used the same thresholds as used in [7] and obtained for all three measures the same sets of frequent patterns as reported in [7]. A closer look revealed that both datasets are rather *transactional* than single graphs, consisting only of sets of trees of *depth one*. In the *Credit* dataset all trees have 21 nodes. A traditional transactional graph miner [10] yields identical results; no additional information on this data can be discovered using single graph mining.

The *WebKB* dataset[3] does not have this drawback and consists of four large graphs that correspond to the hyperlink structure of web pages from a computer science department. Nodes are labeled according to the seven types of web-page that they represent. Edges are unlabeled. Figure 4 summarizes the characteristics of the datasets. Table 1 lists details of the computation of the *MIS* for all patterns found on the Cornell dataset using our measure with a minimum support threshold of 61. *QUALEX-MS* [3][4], a state of the art (approximative) *MIS* solver [8] was used to calculate the *MIS*. The table shows that for larger

[2] http://cygnus.uta.edu/subdue/databases/index.html
[3] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
[4] http://www.stasbusygin.org/

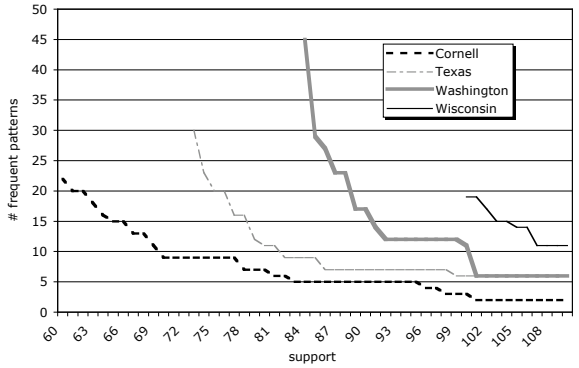

Fig. 3. Patternset sizes for different supports on the 4 WebKB graphs using $\sigma_{\wedge}$.

Dataset	Number of Nodes	Number of Edges
Cornell	627	1177
Texas	761	1507
Washington	1074	2158
Wisconsin	983	2311

Dataset	Node degree max	Node degree avg
Cornell	130	3.75
Texas	149	3.96
Washington	161	4.02
Wisconsin	162	4.70

Fig. 4. Characteristics of the real world datasets used for the experiments.

patterns, the number of occurrences is prohibitive; we were not able to compute the overlap graph, and consequently could not solve the *MIS* problem. As usually bigger patterns are of interest, this is a problem for the overlap-based measures.

The results of the experiments on the *WebKB* datasets are summarized in Figure 3. They suggest a relationship between the size of the database and the computational costs of the frequent pattern extraction. Expressing the support relative to the number of nodes in the data graphs, most datasets show the same behavior, except for *Washington*, where lower relative supports were feasible.

Moreover, we applied our algorithm to a life science database with up to 18.000 nodes and 24.000 edges [9]. However, these results require more research.

4 Conclusions and future work

We introduced a new support measure for mining frequent subgraphs in large single graphs and compared it experimentally and theoretically with existing measures. Existing measures are based on constructing overlap graphs, which soon grow impractically large; this makes solving the subsequent NP complete problem impossible. Since the proposed new measure does not suffer from this problem, it can be evaluated in cases where the old measures cannot be evaluated. Furthermore we showed that the new measure is an upper bound for the other measures, allowing us to guarantee a superset of patterns. We believe there are no clear advantages or disadvantages with respect to the interpretability of any of the measures.

We only compared with complete frequent (single)subgraph miners. Further applications may be found in heuristic single graph miners, of which *SubDue* [6] is an example, and graph miners dealing with additional constraints [4].

The proposed support measure is extendible in multiple ways. We present one example here. Our measure was introduced as a node-based support measure and is easily turned into an edge-based measure. More interestingly, it is possible to generalize our measure to more general substructures than nodes or edges.

Given a parameter k, we can define a support measure based on determining where each connected subgraph with k nodes of the pattern can be matched to.

Definition 4. *For a pattern p, a graph g, and a parameter k, the minimum k-image based support is defined as:*

$$\sigma_{\wedge}(p,g) = \min_{V \subseteq \mathbb{V}_p, |V|=k, V \text{ connected}} |\{\{\varphi_i(V)\} : \varphi_i \text{ is an occurrence of } p \text{ in } g\}|$$

Intuitively, we obtain a measure which achieves counts that are closer to the total number of occurrences of a pattern, while the counts are still anti-monotonic. Especially in larger patterns, it is sometimes more intuitive to allow for more overlap between occurrences than when only single nodes or edges are used.

Acknowledgments We thank Luc De Raedt for helpful comments and discussion, as well as Hannu Toivonen for providing the life sciences database. We thank Christian Borgelt for making the MoSS mining tool publicly available. The authors were supported by the EU FET IST project "Inductive Querying", contract number FP6-516169 and by a doctoral research grant by the KU Leuven research fund.

References

1. Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.
2. Björn Bringmann and Siegfried Nijssen. What is frequent in a single graph? In *International Workshop on Mining and Learning with Graphs (MLG)*, 2007.
3. S. Busygin. A new trust region technique for the maximum weight clique problem. *Discrete Applied Mathematics*, 154:2080–2096, 2006.
4. Chen Chen, Xifeng Yan, Feida Zhu, and Jiawei Han. gApprox: Mining frequent approximate patterns from a massive network. In *Proc. 2007 Int. Conf. on Data Mining (ICDM'07)*, 2007.
5. M. Fiedler and C. Borgelt. Support computation for mining frequent subgraphs in a single graph. In *International Workshop on Mining and Learning with Graphs (MLG)*, 2007.
6. Lawrence B. Holder, Diane J. Cook, and Surnjani Djoko. Substucture discovery in the SUBDUE system. In *KDD Workshop*, pages 169–180, 1994.
7. Michihiro Kuramochi and George Karypis. Finding frequent patterns in a large sparse graph. *Data Min. Knowl. Discov.*, 11(3):243–271, 2005.
8. T.S. Motzkin and E.G. Straus. Maxima for graphs and a new proof of a theorem of Turan. *Canadian Journal of Mathematics*, 17(4):533–540, 1965.
9. P. Sevon, L. Eronen, P. Hintsanen, K. Kulovesi, and H. Toivonen. Link discovery in graphs derived from biological databases. In *Data Integration in the Life Sciences*, volume 4075, 2006.
10. Xifeng Yan and Jiawei Han. gSpan: Graph-based substructure pattern mining. In *ICDM*, pages 721–724, 2002.