Min, Max and PTIME Anti-Monotonic Overlap Graph Measures

Toon Calders T.CALDERS@TUE.NL
Eindhoven University of Technology, Dept. of Math. and Comp. Science, MB 5600 Eindhoven, The Netherlands

Jan Ramon JANR@CS.KULEUVEN.BE
Katholieke Universiteit Leuven, Dept. of Computer Science, Celestijnenlaan 200A, 3001 Leuven, Belgium

Dries Van Dyck DRIES.VANDYCK@UHASSELT.BE
Hasselt University, Transnational University of Limburg, Agoralaan Building D, 3590 Diepenbeek, Belgium

1. Introduction

The central task in graph mining is to find subgraphs, called *patterns* that occur frequently in either a collection of graphs, or in one large graph. Especially in the single-graph setting, the notion of frequency, however, is not at all straightforward. For example, the naïve solution of taking the number of instances of the pattern as its frequency has the undesirable property that extending a pattern (i.e., making it more restrictive), may increase its frequency. Hence, as pointed out by Vanetik, Gudes and Shimony (2006), a good frequency measure must be *anti-monotonic*, i.e., the frequency of a super-pattern may not be higher as that of a subpattern. Not only the correctness, but also the efficiency of most existing graph pattern miners relies critically on this property, as it allows for pruning large parts of the search space.

An important class of anti-monotonic support measures in the single graph setting is based on the notion of an overlap graph — a graph in which each vertex corresponds to a match of the pattern and two vertices are connected by an edge if the corresponding matches overlap. Vanetik, Gudes and Shimony proved necessary and sufficient conditions for anti-monotonicity in the single, labeled graph setting, in which the vertices of the overlap graph represent subgraphs of the data set isomorphic to the pattern, and the edges represent edge overlap (2006) between the subgraphs.

In the context of graph mining, however, not only subgraph isomorphism and labeled graphs are important. On the one hand, the importance of homeomorphic based graph mining increased drastically with the study of biological networks (Bandyopadhyay et al., 2006; Grunewald et al., 2007). On the other hand, in applications where vertices can play several roles (e.g. social networks) homomorphism is more suitable. Homomorphism in the context of data mining has been thoroughly investigated in the field of inductive logic programming (Muggleton & Raedt, 1994).

The main contributions of this paper are: (1) We extend the anti-monotonicity results of Vanetik, Gudes and Shimony to all 24 combinations of iso-, homo-, or homeomorphism, on labeled or unlabeled, directed or undirected graphs, with edge- or vertex-overlap. (2) We show that (under reasonable assumptions) the *maximum independent set measure* (MIS) of Vanetik, Gudes and Shimony (2006) is the smallest anti-monotonic measure in the class of overlap-graph based frequency measures. We also introduce the new *minimum clique partition measure* (MCP) which represents the largest possible one. (3) In general, both the MIS and the MCP measure are NP-hard in the size of the overlap graph. We introduce the polynomial time computable Lovasz measure, which is is sandwiched between the former two, and show that is anti-monotonic.

2. Preliminaries

Graphs A graph $G = (V, E)$ is a pair in which V is a (non-empty) set of *vertices* or *nodes* and E is either a set of *edges* $E \subseteq \{\{v, w\} \mid v, w \in V\}$ or a set of *arcs* $E \subseteq \{(v, w) \mid v, w \in V\}$. In the latter case we call the graph *directed*. A *labeled* graph is a quadruple $G = (V, E, \Sigma, \lambda)$, with (V, E) a graph, Σ a non-empty finite, totally ordered set of labels, and λ a function $V \to \Sigma$ assigning labels to the vertices. We will use the notation $V(G)$, $E(G)$ and λ_G to refer to the set of vertices, the set of arcs (edges) and the labeling function of a graph G, respectively. By $\mathcal{G}$, we denote the class of all graphs; by $\mathcal{G}^{\rightarrow}$ ($\mathcal{G}^{\leftrightarrow}$), the restriction to directed (undirected) graphs; and by $\mathcal{G}_\lambda$ ($\mathcal{G}_\bullet$) the restriction to labeled (unlabeled) graphs. We often combine notation; e.g., $\mathcal{G}_\bullet^{\rightarrow}$ for directed, unlabeled graphs.

Morphisms The following concepts introduced in terms of $\mathcal{G}_\lambda^{\rightarrow}$ are also valid for undirected and/or unlabeled graphs by dropping the direction of the edges and/or the labels of the vertices.

A *homomorphism* π from $H = (V_H, E_H, \Sigma, \lambda_H)$ to $G = (V, E, \Sigma, \lambda)$ is a mapping from $V_H \to V$, such that $\forall (v, w) \in E_H : (\pi(v), \pi(w)) \in E$. We say that H is homomorphic to G and write $H \to G$.

An *isomorphism* from H to G is a bijective homomorphism π from H to G such that $(v, w) \in E(H)$ if and only if $(\pi(v), \pi(w)) \in E(G)$. In that case, we say that H is isomorphic to G and write $H \cong G$.

A path of length k in G is a sequence of vertices $(v_0, \ldots, v_k)$ with $(v_{i-1}, v_i) \in E$. The vertices $v_1, \ldots, v_{k-1}$ are called the *inner* vertices and v_0, v_k the *end* vertices of the path. Two paths P_1 and P_2 of G are called *disjoint* or *independent* if no inner node of P_1 is in P_2 and vice versa. The set of all paths of G is denoted P_G, and of all paths with end vertices v and w, $P_G(v, w)$. A *subgraph homeomorphism* π from H to G is a pair of injective mappings from $V(H) \to V(G)$ and from $E(H) \to P_G$, such that $\forall (v, w) \in E(H)$:

$\pi((v, w)) \in P_G(\pi(v), \pi(w)) \wedge$
$\forall x \in \pi((v, w)) : \forall y \in V(G) \setminus \{v, w\} : \pi(y) \neq x$,

and $\forall (v, w), (x, y) \in E(H)$:

$(v, w) \neq (x, y) \Rightarrow \pi((v, w))$ and $\pi((x, y))$ disjoint (LaPaugh & Rivest, 1978).

By $\mathcal{H}$, $\mathcal{I}$ and $\mathcal{O}$, we denote the class of graph homomorphisms, isomorphisms and homeomorphisms, respectively.

We call π *surjective* if $\forall v' \in V(G)$ and $\forall e' \in E(G)$:

$[(\exists v \in V(H) : v' = \pi(v)) \vee (\exists e \in E(H) : v' \in \pi(e))]$
$\wedge [\exists e \in E(H) : e' \in \pi(e)]$.

If for $\pi : H \to G \in \{\mathcal{H}, \mathcal{I}, \mathcal{O}\}$ it holds that $\lambda_H(v) = \lambda_G(\pi(v))$, we call π *label-preserving*. We will always implicitly assume that π is label-preserving when $H, G \in \mathcal{G}_\lambda$.

3. Support measures and overlap graphs

Definition 1. *Consider a pattern $P \in \mathcal{G}^\alpha_\beta$ and a single graph $G \in \mathcal{G}^\alpha_\beta$, $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$. A* support measure *on $\mathcal{G}^\alpha_\beta$ is a function f that maps (P, G) to $f(P, G) \in \mathbb{N}$. $f(P, G)$ is called the* support of P in G.

Definition 2. *A support measure f on $\mathcal{G}^\alpha_\beta$ is* anti-monotonic *if $\forall P, G \in \mathcal{G}^\alpha_\beta \quad \forall p \subseteq P : f(P, G) \leq f(p, G)$.*

Definition 3. *Let $\mathcal{K} \in \{\mathcal{H}, \mathcal{I}, \mathcal{O}\}$ and $P, G \in \mathcal{G}^\alpha_\beta$, $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$.*
*A $\mathcal{K}$-*match of P in G *is a minimal subgraph $g \subseteq G$, for which there exists a surjective mapping $\pi \in \mathcal{K}$ from P to g.*

Most anti-monotonic measures are based on the notion of an *overlap* graph G^γ_P (Vanetik et al., 2006; Kuramochi & Karypis, 2005)

Definition 4. *Let $P, G \in \mathcal{G}^\alpha_\beta$, $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, $\gamma \in \{$vertex, edge$\}$ and $\mathcal{K} \in \{\mathcal{H}, \mathcal{I}, \mathcal{O}\}$. Two subgraphs g_1 and g_2 of G have a* vertex-overlap *if $V(g_1) \cap V(g_2) \neq \emptyset$ and an* edge-overlap *if $E(g_1) \cap E(g_2) \neq \emptyset$.*
*The $\mathcal{K}$-γ-*overlap graph G^γ_P *of a pattern P in the dataset G is an undirected, unlabeled graph in which each vertex corresponds to a $\mathcal{K}$-match of the pattern P and two vertices are connected if the corresponding $\mathcal{K}$-matches have an γ-overlap.*

Note that G^γ_P is always undirected and that the edges depend on the used notion of overlap. For example, G^γ_P will be denser for vertex-overlap than for edge-overlap because the latter implies the former.

Vanetik, Gudes and Shimony (2006) consider three operations on the overlap graph G^γ_P: clique contraction, edge removal and vertex addition, as defined below.

Definition 5. *Let $K \subseteq G$ be a clique in $G = (V, E)$. The* clique contraction $\mathsf{CC}(G, K)$ *yields a new graph $G' = (V', E')$ in which $K \subseteq G$ is replaced by a new vertex $k \notin V$ adjacent to $\{w \mid \forall v \in V(K) : \{v, w\} \in E\}$:*

$$\begin{aligned} V' &= V \setminus V(K) \cup \{k\} \\ E' &= E \setminus \{\{v, w\} \mid \{v, w\} \cap V(K) \neq \emptyset\} \\ &\quad \cup \{\{k, w\} \mid \forall v' \in V(K) : \{v', w\} \in E\}. \end{aligned}$$

The edge removal $\mathsf{ER}(G, e)$ *of the edge $e = \{v, w\}$ in the graph $G = (V, E)$ yields a new graph*

$G' = (V, E \setminus \{\{v, w\}\})$.

The vertex addition $\mathsf{VA}(G, v)$ *of the vertex $v \notin V$ in the graph $G = (V, E)$ yields a new graph*

$G' = (V \cup \{v\}, E \cup \{\{v, w\} \mid w \in V\})$.

The rationale behind these operations is that the $\mathcal{K}$-γ-overlap graph of P can be transformed into the $\mathcal{K}$-γ-overlap graph of p by means of these operations (Vanetik et al., 2006).

Definition 6. *A* graph measure *is a function $\hat{f} : \mathcal{G}^\leftrightarrow_\bullet \to \mathbb{N}$. Let o be a graph operation that transforms a graph G into a graph $o(G)$. A graph measure $\hat{f}$ is* increasing *under o if and only if $\forall G \in \mathcal{G} : \hat{f}(G) \leq \hat{f}(o(G))$.*

*Let $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, $\gamma \in \{$vertex, edge$\}$ and $\mathcal{K} \in \{\mathcal{H}, \mathcal{I}, \mathcal{O}\}$. A support measure f on $\mathcal{G}^\alpha_\beta$ is a $\mathcal{K}$-γ-*overlap measure *on $\mathcal{G}^\alpha_\beta$, if there exists a graph measure $\hat{f}$ such that $\forall P, G \in \mathcal{G}^\alpha_\beta : f(P, G) = \hat{f}(G^\gamma_P)$.*

The following theorem was originally proved by Vanetik, Gudes and Shimony (2006) for $\alpha \in \{\to, \leftrightarrow\}$, $\mathcal{K} = \mathcal{I}$ and $\gamma =$ edge and its generalization to the complete space defined by the parameters α, β, $\mathcal{K}$ and γ is our main result:

Theorem 7. *Let $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, $\mathcal{K} \in \{\mathcal{I}, \mathcal{H}, \mathcal{O}\}$, and $\gamma \in \{$vertex, edge$\}$.*

Any $\mathcal{K}$-γ-overlap measure f on $\mathcal{G}^\alpha_\beta$ is anti-monotonic if and only if the associated graph measure $\hat{f}$ is increasing under CC, ER *and* VA.

4. Minimal, maximal and PTIME overlap measures

Let $\overline{G} = (V(G), \{\{v, w\} \mid v, w \in V\} \setminus E(G))$, denote the complement graph of $G \in \mathcal{G}^\leftrightarrow_\bullet$. E.g., for the *complete*

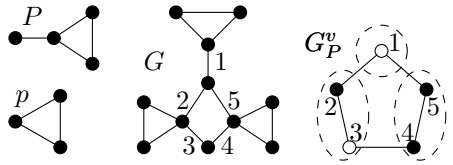

Figure 1. Left: Patterns P, p and a graph G. The 5 $\mathcal{I}$-matches of P in G are indicated by the image in G of the edge in P outside the triangle. Right: The $\mathcal{I}$-vertex-overlap graph G_P^v with a MCP (dashed ellipses) and a MIS (white vertices).

graph on k vertices, $K_k = (\{v_1, \ldots, v_k\}, \{\{v_i, v_j\} \mid 1 \leq i \neq j \leq k\})$, $\overline{K_k}$ is the graph with k isolated vertices. We call an overlap measure f *meaningful* if it is anti-monotonic and assigns the frequency k to k non-overlapping matches, i.e., $\hat{f}(\overline{K_k}) = k$.

An *independent set* of G is a subset I of $V(G)$ such that $\forall v, w \in I : \{v, w\} \notin E(G)$. A *maximum independent set* (MIS) of G is an independent set of maximum cardinality and its size is notated as $mis(G)$. Up to now, all meaningful overlap measures f we are aware of are MIS-measures, i.e., the support of $f(P, G) = mis(G_P^\gamma)$. MIS was introduced and proven to be anti-monotonic in (Vanetik et al., 2006).

We introduce a new anti-monotonic overlap measure, inspired by the CC-operation:

Definition 8. *A* clique partition *of $G \in \mathcal{G}_\bullet^\leftrightarrow$ is a partitioning of $V(G)$ into $\{V_1, \ldots, V_k\}$ such that each V_i induces a clique in G. A* minimum clique partition *(MCP) is a clique partition of minimum size. Its size is denoted* $mcp(G)$. *The* MCP-measure *is defined by* $MCP(P, G) : (P, G) \to mcp(G_P^\gamma)$.

Theorem 9. *The MCP-measure is meaningful.*

It is interesting to compare MCP with MIS. Let $\chi(G)$ be the *chromatic number* of G, i.e., the minimal number of colors to color the vertices of G such that no two vertices with the same color are adjacent, and let $\omega(G)$ be the *clique number*; the size of the largest clique in G.

First, it is known that $mcp(G) = \chi(\overline{G})$ and $mis(G) = \omega(\overline{G})$ (see, e.g., (Gross & Yellen, 2004), section 5.5.1). Consequently, $mcp(G) \geq mis(G), \forall G \in \mathcal{G}_\bullet^\leftrightarrow$, since the size of a maximum clique is an lower bound for the chromatic number. Informally, it is easy to see why this is so: let $V_1, \ldots, V_k$ be an MCP and I a MIS for G. We know that I contains at most one vertex v_i of each V_i, $1 \leq i \leq k$. In other words, to decide whether we can include a match of V_i, MIS forces us to choose either no match or exactly one match v_i, which must be independent of all chosen $v_j \in V_j$. MCP, however, allows us to count a match in V_i as soon there is *a* match in V_i which does not overlap with *a* match in V_j. That is, we can make another choice for each (V_i, V_j) pair (see Figure 1 for a concrete example).

Interestingly, MIS and MCP turn out to be the minimal and the maximal possible meaningful overlap measures:

Theorem 10. *Let $\mathcal{K} \in \{\mathcal{I}, \mathcal{H}, \mathcal{O}\}$, $\gamma \in \{$vertex, edge$\}$, $\alpha \in \{\rightarrow, \leftrightarrow\}$, and $\beta \in \{\lambda, \bullet\}$. For every meaningful $\mathcal{K}$-γ-overlap measure f on $\mathcal{G}_\beta^\alpha$, and every $P, G \in \mathcal{G}_\beta^\alpha$, it holds that: $MIS(P, G) \leq f(P, G) \leq MCP(P, G)$.*

Unfortunately, both MIS and MCP are known to be NP-hard to compute in the size of the overlap graph. A well-known measure that is sandwiched between the MIS and the MCP and that can be computed in polynomial time, is the theta or Lovasz function (Knuth, 1994). There are several equivalent characterizations of this function. One definition is: $\theta(G) = \min_A \lambda_{\max}(A)$, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of matrix A and the minimum is taken over all feasible matrices A such that $A^\top = A$, $A_{ii} = 1$ and $A_{ij} = 1$ if $(i, j) \notin E(G)$.

Theorem 11. *θ is a meaningful overlap measure.*

References

Bandyopadhyay, S., Sharan, R., & Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison. *Genome Res., 16,* 428–435.

Gross, J. L., & Yellen, J. (2004). *Handbook of graph theory*. CRC Press.

Grunewald, S., Kristoffer, F., Dress, A., & Moulton, V. (2007). Qnet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. *Molecular Biology and Evolution, 24*, 532–538.

Knuth, D. E. (1994). The sandwich theorem. *Electron. J. Combin, 1*.

Kuramochi, M., & Karypis, G. (2005). Finding frequent patterns in a large sparse graph. *Data Min. Knowl. Discov., 11*, 243–271.

LaPaugh, A. S., & Rivest, R. L. (1978). The subgraph homeomorphism problem. *STOC '78* (pp. 40–50). New York, NY, USA: ACM Press.

Muggleton, S., & Raedt, L. D. (1994). Inductive logic programming : Theory and methods. *Journal of Logic Programming, 19,20*, 629–679.

Vanetik, N., Shimony, S. E., & Gudes, E. (2006). Support measures for graph data. *Data Min. Knowl. Discov., 13*, 243–260.