Anti-Monotonic Overlap-Graph Support Measures

Toon Calders (t.calders@tue.nl)
Eindhoven University of Technology

Jan Ramon (jan.ramon@cs.kuleuven.be)
Katholieke Universiteit Leuven

Dries Van Dyck (dries.vandyck@uhasselt.be)
Hasselt University, Transnational University of Limburg

Abstract

In graph mining, a frequency measure is anti-monotonic if the frequency of a pattern never exceeds the frequency of a subpattern. The efficiency and correctness of most graph pattern miners relies critically on this property. We study the case where the dataset is a single graph. Vanetik, Gudes and Shimony already gave sufficient and necessary conditions for anti-monotonicity of measures depending only on the edge-overlaps between the intances of the pattern in a labeled graph. We extend these results to homomorphisms, isomorphisms and homeomorphisms on both labeled and unlabeled, directed and undirected graphs, for vertex and edge overlap. We show a set of reductions between the different morphisms that preserve overlap. We also prove that the popular maximum independent set measure assigns the minimal possible meaningful frequency, introduce a new measure based on the minimum clique partition that assigns the maximum possible meaningful frequency and introduce a new measure sandwiched between the former two based on the poly-time computable Lovász θ-function.

1 Introduction

Recently, graph mining has emerged as a new field within contemporary data mining that got a lot of attention over the last several years. The central task 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 [13], a good frequency measure must be such that the frequency of a super-pattern is always at most as high as that of a subpattern. This property is called *anti-monotonicity*. Also for reasons of efficiency, anti-monotonicity of the frequency measure is highly desirable, as it allows for pruning large parts of the search space. The efficiency and correctness of most graph pattern miners relies critically on the anti-monotonic property of the used frequency measure.

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 [13] 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 [1, 6]. 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 [11]. In this paper we extend the results of Vanetik, Gudes and Shimony to these settings as well. Our main contributions are:

1. We study systematically all 24 combinations of iso-, homo-, or homeomorphism, on labeled or unlabeled, directed or undirected graphs, with edge- or vertex-overlap and extend the anti-monotonicity results.

2. In our proofs, we use reductions which are also of interest in their own right, as they allow to transfer results for different types of morphisms and overlap from one setting to another. For an overview of the different reductions, see Figure 1.

3. An interesting consequence of the reductions is that any unlabeled, undirected graph is a potential vertex- and edge-overlap graph in all considered settings.

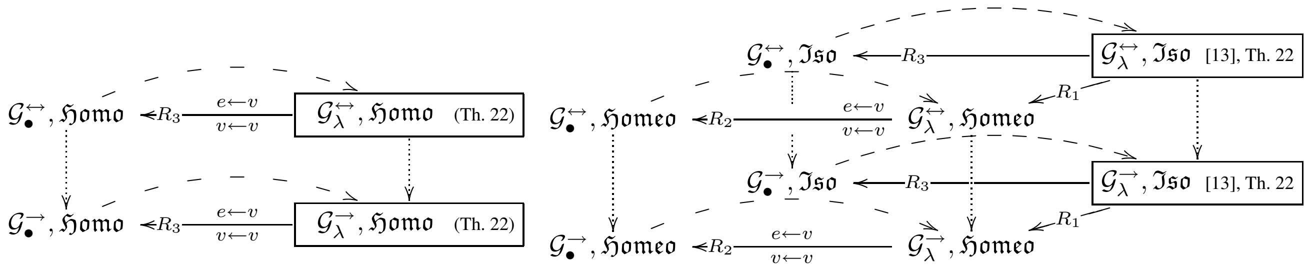

$\mathcal{G}^{\leftrightarrow}_{\bullet}, \mathfrak{Homo}$
R_3
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}^{\leftrightarrow}_{\lambda}, \mathfrak{Homo}$ (Th. 22)
$\mathcal{G}^{\rightarrow}_{\bullet}, \mathfrak{Homo}$
R_3
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}^{\rightarrow}_{\lambda}, \mathfrak{Homo}$ (Th. 22)
$\mathcal{G}^{\leftrightarrow}_{\bullet}, \mathfrak{Iso}$
R_3
$\mathcal{G}^{\leftrightarrow}_{\lambda}, \mathfrak{Iso}$ [13], Th. 22
R_1
$\mathcal{G}^{\leftrightarrow}_{\bullet}, \mathfrak{Homeo}$
R_2
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}^{\leftrightarrow}_{\lambda}, \mathfrak{Homeo}$
$\mathcal{G}^{\rightarrow}_{\bullet}, \mathfrak{Iso}$
R_3
$\mathcal{G}^{\rightarrow}_{\lambda}, \mathfrak{Iso}$ [13], Th. 22
R_1
$\mathcal{G}^{\rightarrow}_{\bullet}, \mathfrak{Homeo}$
R_2
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}^{\rightarrow}_{\lambda}, \mathfrak{Homeo}$

Figure 1. Overview of the reductions. R_1, R_2, and R_3 are proven respectively in Theorems 25, 29, and 33; the dashed and the dotted arrows in Prop. 12. Arrows representing reductions changing γ-overlap into γ'-overlap are labeled with $\gamma \rightarrow \gamma'$.

4. We show that (under reasonable assumptions) the *maximum independent set measure* (MIS) of Vanetik, Gudes and Shimony [13] 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.

5. In general, both the MIS measure and the MCP measure are NP-hard to compute in the size of the overlap graph. The Lovász measure is computable in polynomial time and is sandwiched between the former two measures. We show that the Lovász measure induces an anti-monotonic overlap-graph based frequency measure.

Due to space limitations, we omit the proofs for our theorems and only state the results and intuitions in this version. In Section 2 we review basic concepts from graph theory. In Section 3 we introduce overlap graphs, support measures and reductions between the different settings. Next, in Section 4 we present our results on minimal, maximal and poly-time computable meaningful overlap measures. In Section 5 we present our reductions and use them to extend the results to all 24 settings. Finally, we conclude in Section 6.

2 Preliminaries

We assume that the reader is familiar with most basic graph theoretic notions and computational complexity. Any textbook in these areas, such as [3] and [12] supply necessary background.

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* $E \subseteq \{\{v, w\} \mid v, w \in V, v \neq w\}$ or a set of *arcs* $E \subseteq \{(v, w) \mid v, w \in V, v \neq w\}$. 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 $V \rightarrow \Sigma$ assigning labels to the vertices. We 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., $\mathcal{G}^{\rightarrow}_{\bullet}$ for directed, unlabeled graphs.

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

A *homomorphism* π from $H = (V_H, E_H, \Sigma, \lambda_H)$ to $G = (V, E, \Sigma, \lambda)$ is a mapping from $V_H \rightarrow V$, such that $\forall (v, w) \in E_H : (\pi(v), \pi(w)) \in E$. If such π exists, we say that H is homomorphic to G.We call π edge-surjective if $\forall (v', w') \in E\ \exists (v, w) \in E_H : \pi(v) = v' \wedge \pi(w) = w'$ and call it surjective if it is both vertex- and edge-surjective.

An *isomorphism* from H to G is a bijective homomorphism π from H to G. In that case, we say that H is isomorphic to G and write $H \cong G$. We use $H \subseteq G$ to denote that $H \cong g$, for some subgraph g of 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 $V(H) \rightarrow V(G)$ and from $E(H) \rightarrow P_G$, such that $\forall (v, w) \in E(H)$:

$\pi((v, w)) \in P_G(\pi(v), \pi(w)) \wedge$

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

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

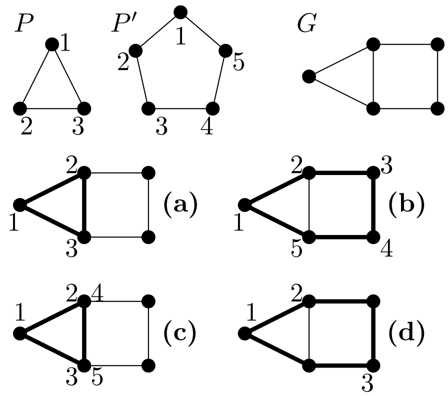


Figure 2. Examples of the different morphisms. An isomorphic image of P (a), P' (b), a homomorphic image of P' (c) and a homeomorphic image of P (d). The edges of the subgraph to which a pattern is mapped are in bold. The image of a vertex of the pattern is labeled with its identifier.

$(v, w) \neq (x, y) \Rightarrow \pi((v, w))$ and $\pi((x, y))$ disjoint [10].

We call π *surjective* if $\forall v' \in V(G)$ and $\forall e' \in E(G)$:
$[(\exists v \in V(H) : v' = \pi(v)) \vee (\exists e \in E(H) : v' \in \pi(e))]$
$\wedge\, [\exists e \in E(H) : e' \in \pi(e)]$.

By $\mathfrak{Homo}$, $\mathfrak{Iso}$ and $\mathfrak{Homeo}$, we denote the class of graph homomorphisms, isomorphisms and homeomorphisms, respectively.

If for $\pi : H \to G \in \{\mathfrak{Homo}, \mathfrak{Iso}, \mathfrak{Homeo}\}$ 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 $H, G \in \mathcal{G}_\lambda$.

Example 1. *Fig. 2 illustrates the introduced morphisms for unlabeled, undirected graphs. Note that in* **(c)** *the vertices 2,4 and 3,5 are mapped to the same vertex and in* **(d)** *the edges* $\{2, 3\}$ *and* $\{3, 1\}$ *are mapped to paths of length 2.*

3 Support measures and overlap graphs

Definition 2. *A* support measure *on* $\mathcal{G}_\beta^\alpha$ *is a function* $f : \mathcal{G}_\beta^\alpha \times \mathcal{G}_\beta^\alpha \to \mathbb{N}$ *that maps* (P, G) *to* $f(P, G)$ *where* P *is called the pattern,* G *is called the dataset graph and* $f(P, G)$ *is called the* support of P in G.

For efficiency reasons, most graph mining algorithms use a level-wise or depth-first approach to generate frequent patterns, expanding smaller patterns to larger ones, which requires an anti-monotonic support measure:

Definition 3. *A support measure* f *on* $\mathcal{G}_\beta^\alpha$ *is* anti-monotonic *iff* $\forall p, P, G \in \mathcal{G}_\beta^\alpha : p \subseteq P \Rightarrow f(P, G) \leq f(p, G)$.

Most support measures are based on the *matches* of a pattern in a graph:

Definition 4. *Let* $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}, \mathfrak{Homeo}\}$ *and* $P, G \in \mathcal{G}_\beta^\alpha$, $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$.

A $\mathfrak{K}$-match *of* P *in* G *is a subgraph* $g \subseteq G$ *for which there exists a surjective mapping* $\pi \in \mathfrak{K}$ *from* P *to* g. *An individual mapping* π *from* P *to* g *is called an* embedding *of* P *in* G.

We call an $\mathfrak{Iso}$*-match of* P *in* G *an* instance *of* P *in* G.

However, just counting the number of $\mathfrak{K}$-matches of a pattern in G does not result in an anti-monotonic support measure, as larger patterns may have more matches (e.g, in Figure 3, there are more instances of P in G than triangles).

3.1 Overlap graph

Most anti-monotonic measures in a single graph setting are based on the notion of an *overlap* graph G_P^γ [13, 9][1]:

Definition 5. *Let* $P, G \in \mathcal{G}_\beta^\alpha$, $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$. *Two subgraphs* g_1 *and* g_2 *of* G *have* vertex-overlap *if* $V(g_1) \cap V(g_2) \neq \emptyset$ *and* edge-overlap *if* $E(g_1) \cap E(g_2) \neq \emptyset$.

Let $\gamma \in \{$*vertex, edge*$\}$ *and* $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}, \mathfrak{Homeo}\}$. *The* $\mathfrak{K}$-γ-overlap graph G_P^γ *of a pattern* P *in the dataset* G *is an undirected, unlabeled graph in which each vertex corresponds to a* $\mathfrak{K}$*-match of the pattern* P *and two vertices are connected if the corresponding* $\mathfrak{K}$*-matches have* γ*-overlap.*

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

Let $p, P, G \in \mathcal{G}_\beta^\alpha$, $\gamma \in \{\text{vertex}, \text{edge}\}$, $\alpha \in \{\to, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, and $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}, \mathfrak{Homeo}\}$. Throughout the article, P denotes the (super)pattern, $p \subseteq P$ the subpattern and G the dataset, a single graph. G_P^γ (G_p^γ) is the $\mathfrak{K}$-γ-overlap graph of P (p) in G.

Vanetik, Gudes and Shimony [13] consider three operations on the overlap graph G_P^γ: clique contraction, edge removal and vertex addition, as defined below.

Definition 6. *Let* $G = (V, E) \in \mathcal{G}_\bullet^{\leftrightarrow}$. *Let* $K \subseteq G$ *be a clique in* G. *The* clique contraction $\mathsf{CC}(G, K)$ *yields a new graph* $G' = (V', E')$ *in which the subgraph* $K \subseteq G$ *is replaced by a new vertex* $k \notin V$ *adjacent to* $\{w \mid \forall v \in V(K) : \{v, w\} \in E\}$*:*

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

[1] [13] uses the term *instance* graph instead of overlap graph. The term *instance* suggests the use of isomorphisms, and we consider support measures based on any kind of morphism, we follow the terminology of [9] to avoid confusion.

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

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

The rationale behind these operations is that the $\mathfrak{K}$-γ-overlap graph of P can be transformed into the $\mathfrak{K}$-γ-overlap graph of p by means of these operations. This can be seen based on the following two observations:

Observation 1 Any $\mathfrak{K}$-match of P contains a $\mathfrak{K}$-match of p.

Observation 2 Let g_1, g_2 be two $\mathfrak{K}$-matches of P and $g_1' \subseteq g_1$ and $g_2' \subseteq g_2$ be two $\mathfrak{K}$-matches of p. If g_1' and g_2' have γ-overlap, so do g_1 and g_2.

These conditions hold for all settings considered in this article. We quickly sketch the main ideas of the transformation process and refer to [13] for the full details. For reasons of simplicity we assume that p contains at least one edge.

Let $g' \subseteq G$ be a match of p, and let $super(g')$ be all matches of P in G containing g'. Because of Observation 1, every match g of P in G must be in at least one $super(g')$. Because of Observation 2, $super(g')$ forms a clique in G_P^γ, as they all overlap on g'. Furthermore, if there is an edge $\{g_1', g_2'\}$ in G_p^γ, there is an edge between any two $g_1 \in super(g_1')$ and $g_2 \in super(g_2')$ in G_P^γ. As such, an induced subgraph of G_p^γ can be formed by subsequently contracting the cliques $super(g')$ until for all $g' \in G_p^\gamma$, either $super(g')$ is empty, or a singleton. It is easy to see that one can go from an induced subgraph of G_p^γ to G_p^γ: first add all vertices not in the induced subgraph with node additions, and then remove spurious edges with edge removals.

3.2 Overlap support measure

Definition 7. *A* graph measure *is a function* $\hat{f} : \mathcal{G}_\bullet^\leftrightarrow \to \mathbb{N}$. *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* $\forall G \in \mathcal{G}_\bullet^\leftrightarrow : \hat{f}(G) \leq \hat{f}(o(G))$.

Definition 8. *Let* $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, $\gamma \in \{$*vertex, edge*$\}$ *and* $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}, \mathfrak{Homeo}\}$.

A support measure f on $\mathcal{G}_\beta^\alpha$ is a $\mathfrak{K}$-γ-overlap support measure *on $\mathcal{G}_\beta^\alpha$, if there exists a graph measure $\hat{f}$ such that* $\forall P, G \in \mathcal{G}_\beta^\alpha : f(P, G) = \hat{f}(G_P^\gamma)$.

Informally, an overlap support measure is a support measure that only depends on the overlap graph. Note that the associated graph measure $\hat{f}$ is always unique. An example of an anti-monotonic overlap support measure is the measure that assigns to every pattern P the size of the maximum independent set (MIS) [13] of G_P^γ; that is, the support is the maximal number of matches that fit in G without overlap. The main result of this article is the generalization of the following theorem of Vanetik, Gudes and Shimony [13]:

Theorem 9 (Vanetik, Gudes, Shimony)**.** *Let* $\alpha \in \{\rightarrow, \leftrightarrow\}$. *Any $\mathfrak{Iso}$-edge-overlap support measure f on $\mathcal{G}_\lambda^\alpha$ is anti-monotonic if and only if the associated graph measure $\hat{f}$ is non-decreasing under clique contraction, edge removal and vertex addition.*

We extend it to the complete space defined by the parameters α, β, $\mathfrak{K}$ and γ. More formally:

Theorem 10. *Let* $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, *and* $\gamma \in \{$*vertex, edge*$\}$.

Any $\mathfrak{K}$-γ-overlap support measure f on $\mathcal{G}_\beta^\alpha$ is anti-monotonic if and only if the associated graph measure $\hat{f}$ is increasing under clique contraction, edge removal and vertex addition.

The proof of sufficiency, i.e., that any $\mathfrak{K}$-γ-overlap support measure f is anti-monotonic if the associated graph measure is increasing under CC, VA and ER follows immediately from the fact that G_P^γ can be transformed into G_p^γ by these operations.

To prove necessity, Vanetik, Gudes and Shimony construct for every unlabeled graph H and every operation o, a triple (P, p, G) (where P is a super-pattern, p a subpattern and G a dataset) such that $G_P^\gamma \cong H$ and $G_p^\gamma \cong o(H)$. So if f would not be increasing under some $o \in \{\mathsf{CC}, \mathsf{ER}, \mathsf{VA}\}$, there would be a H such that $f(H) > f(o(H))$ and one could construct a G, P and p such that $f(G, P) > f(G, p)$, which would mean that f is not anti-monotonic. We follow the same approach.

3.3 Reductions

The necessity proofs for most settings are based on reductions from $\mathfrak{K}$-matches for $\mathcal{G}_\beta^\alpha$ to $\mathfrak{K}'$-matches for $\mathcal{G}_{\beta'}^{\alpha'}$.

Definition 11. *Let* $\mathfrak{K}, \mathfrak{K}' \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, $\alpha, \alpha' \in \{\rightarrow, \leftrightarrow\}$, $\beta, \beta' \in \{\bullet, \lambda\}$, *and* $\gamma, \gamma' \in \{$*edge*, *vertex*$\}$.

A $\mathfrak{K}, \gamma$-overlap on $\mathcal{G}_\beta^\alpha$ to $\mathfrak{K}', \gamma'$-overlap on $\mathcal{G}_{\beta'}^{\alpha'}$ reduction *is a function* $R : (\mathcal{G}_\beta^\alpha)^3 \to (\mathcal{G}_{\beta'}^{\alpha'})^3$ *that maps a triplet* (p, P, G) *to a triplet* (p', P', G') *such that:*

$$(1)\ p \subseteq P \textit{ iff } p' \subseteq P' \textit{ and } (2)\ G_p^\gamma \cong {G'}_{p'}^{\gamma'} \wedge G_P^\gamma \cong {G'}_{P'}^{\gamma'}.$$

Note that this definition does not automatically imply that the number of $\mathfrak{K}$-embeddings of P in G equals the number of embeddings of P' in G', as P' might have more/less automorphisms than P.

The following property gives reductions from unlabeled to labeled graphs, and from undirected to directed graphs.

Property 12. *For all* $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\gamma \in \{$*vertex*, *edge*$\}$, $\beta \in \{\lambda, \bullet\}$, $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, *there exist reductions:*

- *from $\mathfrak{K}$, γ-overlap on $\mathcal{G}_\bullet^\alpha$ to $\mathfrak{K}$, γ-overlap on $\mathcal{G}_\lambda^\alpha$; and*
- *from $\mathfrak{K}$, γ-overlap on $\mathcal{G}_\beta^\leftrightarrow$ to $\mathfrak{K}$, γ-overlap on $\mathcal{G}_\beta^\rightarrow$.*

An overview of the reductions which follow from our results in Section 5 are shown in Figure 1.

4 Minimal, Maximal and PTIME overlap support measures

Let $\overline{G} = (V(G), \{\{v,w\} \mid v,w \in V\} \setminus E(G))$, denote the complement graph of $G \in \mathcal{G}_\bullet^\leftrightarrow$. E.g., for the *complete 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 support 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 support 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 [13]. A more compact proof can be found in [4].

4.1 MCP-measure

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

Definition 13. *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 $MCP(P,G) : (P,G) \rightarrow mcp(G_P^\gamma)$.*

Theorem 14. *Let $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, $\gamma \in \{vertex, edge\}$, and $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$. The MCP-measure is an anti-monotonic $\mathfrak{K}$-γ-overlap measure on $\mathcal{G}_\beta^\alpha$.*

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., [5], 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 a 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.

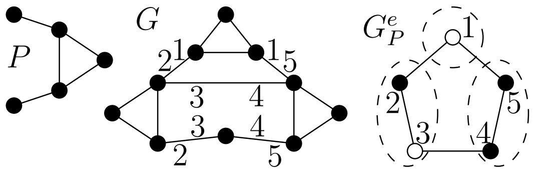


Figure 3. Left: A pattern P and a graph G. The 5 $\mathfrak{Iso}$-matches of P in G are indicated by the image in G of the edges outside the triangle. Right: The $\mathfrak{Iso}$-edge-overlap graph G_P^e with a MCP (dashed ellipses) and a MIS (white vertices).

Example 15. *Let us look at an example: consider pattern P and the graph G as shown in Figure 3. The 5 $\mathfrak{Iso}$-matches of P are indicated by an identifier on the image in G of the edges outside the triangle of P. The $\mathfrak{Iso}$-edge-overlap graph G_P^e of P in G is shown on the right in Figure 3 and is isomorphic to a pentagon. The white vertices mark the MIS $\{1,3\}$ and the dashed ellipses mark the MCP $\{\{1\},\{2,3\},\{4,5\}\}$ of G_P^e. Hence, if we count match 1 with MIS, we can only take match 3 or match 4 as second independent match, because 3 and 4 overlap, leading to a MIS-support of 2. This is a bit unnatural, because each of the 3 matches of the triangle can be extended to a match of P in a way that they do not overlap with each other, which would lead to a support of 3 of P.*

This more natural notion of counting independent matches is exactly what MCP-support allows us to do: we do not count individual matches, but groups of matches of P sharing a match of a subpattern p (a triangle) and allow to "switch" matches to decide whether a group is independent of an other. In this example, the group $\{1\}$ is independent of the groups $\{2,3\}$ and $\{4,5\}$, because it does not overlap with match 3 respectively match 4 and the group $\{2,3\}$ is independent of the group $\{4,5\}$ because, for instance, match 2 and match 5 do not overlap.

4.2 Bounding theorem and PTIME overlap support measure

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

Theorem 16. *Let $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, $\gamma \in \{vertex, edge\}$, $\alpha \in \{\rightarrow, \leftrightarrow\}$, and $\beta \in \{\lambda, \bullet\}$.*

For every meaningful $\mathfrak{K}$-γ-overlap measure f on $\mathcal{G}_\beta^\alpha$, and every $P, G \in \mathcal{G}_\beta^\alpha$, it holds that:

$$MIS(P, G) \leq f(P, G) \leq MCP(P, G) \ .$$

Proof. We use Theorem 10 to show both the minimality of MIS and the maximality of MCP.

Let $H = G_P$, let $mis(H) = k$, and let $I = \{v_1, \ldots, v_k\}$ be a MIS for H. Starting from the graph $(\{v_1, \ldots, v_k\}, \emptyset)$ we can add the vertices $V(H) \setminus I$ using VA and remove edges not in $E(H)$ by ER. Since f is meaningful, it is anti-monotonic and therefore $\hat{f}$ cannot decrease after each step, starting from $\hat{f}((\{v_1, \ldots, v_k\}, \emptyset)) = k$. As such, $\hat{f}(H)$ is larger than or equal to $k = mis(H)$.

On the other hand, let $mcp(H) = k$, and let $\{V_1, \ldots, V_k\}$ be an MCP for H and let $H_{cc} = \mathsf{CC}(\ldots \mathsf{CC}(\mathsf{CC}(H, V_1), V_2) \ldots, V_k)$. H_{cc} does not have edges: if it would, then joining the two cliques that were contracted to two connected vertices of H_{cc} would give us a smaller clique partition. Because f is anti-monotonic, $\hat{f}$ is increasing under CC and thus

$$\hat{f}(H) \leq \hat{f}(\mathsf{CC}(H, V_1)) \leq \cdots \leq \hat{f}(H_{cc}) = \hat{f}(\overline{K_k}) = k \ .$$

□

Unfortunately, both mis and mcp are known to be NP-hard to compute in the size of the overlap graph. This leads us to the following question: does there exist a meaningful overlap measure which is efficiently computable?

A well-known measure that is sandwiched between mis and mcp and that can be computed in polynomial time, is the theta function, also known as the Lovász function [8]. There are several equivalent characterizations of this function. The most concise definition is probably: $\theta(G) = \min_A \lambda_{\max}(A)$, 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 17. *θ is a meaningful overlap measure.*

5 Necessity for other morphisms and graph settings

In the previous sections we considered measures on overlap graphs that are anti-monotonic w.r.t. the operations VA, ER and CC. Let us now return to the connection between measures with this property and anti-monotonic graph support measures based on overlap graphs. [13] showed that when considering unlabeled, undirected graphs under subgraph isomorphism, a graph support measure based on the edge-overlap graph is anti-monotonic if an only if the corresponding graph measure on the edge-overlap graph is anti-monotonic w.r.t. VA, ER and CC. In this section, we generalize this result to all 24 settings. We do this by proving a base case and then applying reductions from all the other settings to this base case.

5.1 Necessity for labeled homomorphisms and isomorphisms

As base case, we prove the necessity of the non-decreasingness of $\hat{f}$ under the three graph operations for the anti-monotonicity of f for $\mathfrak{K}, \gamma$-overlap on $\mathcal{G}_\lambda^\alpha$, for all combinations of $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}\}$, $\gamma \in \{\text{vertex}, \text{edge}\}$, and $\alpha \in \{\rightarrow, \leftrightarrow\}$. The proof will not rely on reductions, but show the necessity directly for these cases. We show only the undirected case, as the proof for the directed case is very similar. Notice also that the directed case follows from the undirected-to-directed reductions shown in Property 12.

We will essentially use invariants under subgraph homomorphism to force an injective homomorphism; i.e., to ensure isomorphism.

Let $G \in \mathcal{G}^\leftrightarrow$. The *odd girth* $g_o(G)$ of G is the size of a smallest cycle of odd length in G. The *distance* $d_G(v, w)$ is equal to the length of a shortest path from v to w in G. If no such cycle respectively shortest path exist, we define $g_o(G)$ respectively $d_G(v, w)$ equal to ∞.

We will use the following well known invariants [7] to force each subgraph homomorphism into an isomorphism:

Proposition 18. *If $H, G \in \mathcal{G}_\beta^\alpha$, $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$ for which there exists a homomorphism $\pi : H \rightarrow G$, then*

1. $g_o(H) \geq g_o(G)$,
2. $\forall v, w \in V(H) : d_H(v, w) \geq d_G(\pi(v), \pi(w))$.

Remark that the first invariant implies that a cycle of length k, with k odd, can only be mapped by a homomorphism to an odd cycle of length at most k.

We will use the following graphs as patterns:

Definition 19. *Let $k + 1$ be an odd integer. B_k denotes the graph in $\mathcal{G}_\lambda^\leftrightarrow$ defined by:*

$$\begin{aligned}
&V(B_k) = V_a \cup V_b \cup V_c \cup V_d, \\
&\quad V_a = \{a_1, \ldots, a_k, a'_1, \ldots, a'_k\}, \ \ V_b = \{b_1, \ldots, b_{k+1}\}, \\
&\quad V_c = \{c_1, \ldots, c_{k+1}\}, \ \ V_d = \{d_1, \ldots, d_k, d'_1, \ldots, d'_{k+1}\}, \\
&E(B_k) = E_a \cup E_b \cup E_c \cup E_d \cup \cup_{i=1}^k \{\{a_i, b_i\}, \{c_i, d_i\}\} \\
&\qquad\qquad \cup \{b_{k+1}, c_{k+1}\}, \\
&\quad E_a = \cup_{i=1}^k \{a_i, a'_i\}, \ \ E_b = \{b_{k+1}, b_1\} \cup \cup_{i=1}^k \{b_i, b_{i+1}\}, \\
&\quad E_c = \{c_{k+1}, c_1\} \cup \cup_{i=1}^k \{c_i, c_{i+1}\}, \ \ E_d = \cup_{i=1}^k \{d_i, d'_i\}, \\
&\lambda_{B_k}(u) = x, \forall u \in V_x, x = a, b, c, d.
\end{aligned}$$

We call the edges $\{a_i, a'_i\}$ arms*, the edges $\{d_i, d'_i\}$ legs, $1 \leq i \leq k$, the cycle induced by V_b the* torso *and the cycle induced by V_c the* hip.

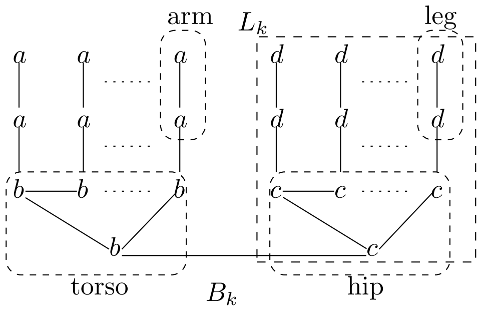


Figure 4. The graphs B_k and L_k.

$L_k \in \mathcal{G}_\lambda^{\leftrightarrow}$ denotes the subgraph of B_k induced by $V_c \cup V_d$ and is called the lower body *of B_k.*

An illustration of both graphs is shown in Figure 4.

Let $P[1], P[2]$ and $p[1], p[2]$ be two instances of $P = B_k$ respectively $p = L_k$ in a larger graph G and let $super(g)$ be equal to all matches of P in G containing $g \subseteq G$. We will use four types of overlap (see Figure 5):

lower body overlap: $P[1]$ and $P[2]$ share the complete lower body, which is a single instance of p, resulting in two adjacent vertices in G_P^γ and a single vertex in G_p^γ,

leg overlap: $P[1]$ and $P[2]$ share a leg, resulting in two adjacent vertices in G_P^γ and two adjacent vertices in G_p^γ,

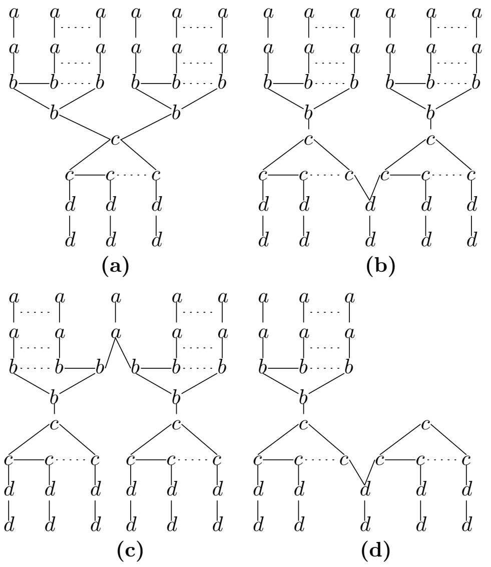
arm overlap: $P[1]$ and $P[2]$ share an arm, resulting in two adjacent vertices in G_P^γ and two independent vertices in G_p^γ,

partial leg overlap: $p[1] \subset P[1]$ shares a leg with $p[2]$, with $super(p[2]) = \emptyset$, resulting in a single vertex in G_P^γ and two adjacent vertices in G_p^γ.

Note that in each type of overlap, there is always vertex overlap if and only if there is edge overlap. When three or more instances of P or p overlap, we will always make sure that an arm or leg is shared by at most two instances of P or p, which is always possible by taking k sufficiently large. When two instances overlap, they always overlap once, i.e., if $P[1]$ and $P[2]$ have an overlap of type x, they do not have an additional overlap of type $y \neq x$. We will call these restrictions *the overlap condition* and assume implicitly that they are obeyed at all times when constructing graphs by overlapping instances of P and p.

Lemma 20. *Let $G \in \mathcal{G}_\lambda^{\leftrightarrow}$ be a graph constructed from n overlapping instances $P[1], \ldots, P[n]$ of $P = B_k$. Then, the $P[1], \ldots, P[n]$ are the only $\mathfrak{Homo}$-matches of P in G.*

Theorem 21. *Let $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\gamma \in \{$vertex, edge$\}$ and $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}\}$. Any undirected graph H is a $\mathfrak{K}$-γ-overlap graph, i.e., there always exist $P, G \in \mathcal{G}_\lambda^\alpha$ such that $H \cong G_P^\gamma$.*



(a) (b) (c) (d)

Figure 5. The four overlap types: (a) lower body overlap (b) leg overlap (c) arm overlap (d) partial leg overlap

Theorem 22. *Let $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\gamma \in \{$vertex, edge$\}$ and $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}\}$. Any $\mathfrak{K}$-γ-overlap support measure f on $\mathcal{G}_\lambda^\alpha$ is anti-monotonic only if the associated graph measure $\hat{f}$ is increasing under clique contraction, edge removal and vertex addition.*

5.2 Labeled isomorphisms to homeomorphisms

We will now show that we can reduce isomorphic mappings to homeomorphic mappings while preserving edge- and vertex-overlap. First we show how to reduce isomorphic mappings to homeomorphic mappings for the labeled case, and then we show how to reduce the labeled homeomorphisms to unlabeled homeomorphisms. We will only give the proofs for the undirected cases, as the directed cases can either be proven in a very similar way (replace all edges by arcs), or by composition of the reduction for the undirected case with the reduction from undirected to directed of Prop. 12.

We first prove some invariants under subgraph homeomorphism which will be important in the proof of correctness of both reductions in this section.

The *degree* of a vertex v in a graph $G \in \mathcal{G}^\alpha$ is defined as $\Delta_G(v) := \#\{w \mid \{v, w\} \in E(G)\}$. The *maximum degree* of G is then $\Delta(G) := max_{v \in V(G)} \Delta_G(v)$.

Definition 23. *The* connection strength between v and w *is defined as*

$$cs_G(v,w) := \max\{|P| \;:\; P \subseteq P_G(v,w) \text{ paths in } P \text{ pairwise disjoint}\}$$

and the maximal connection strength between two nodes of G *as:* $CS_G = \max_{v,w \in V_G} cs_G(v,w)$.

For $G \in \mathcal{G}^{\leftrightarrow}$, $cs_G(v,w)$ is often called the local connectivity of v and w. Notice that connection strength between nodes v and w is well known to be equivalent with vertex-connectivity (Menger's theorem [3]); i.e., the minimal number of vertices that need to be removed to make v and w disconnected.

Lemma 24. *Let π be a subgraph homeomorphism from H to G. Then,*

1. $|V(H)| \leq |V(G)|$ *and* $|E(H)| \leq |E(G)|$*;*
2. $\forall v \in V(H) : \Delta_H(v) \leq \Delta_G(\pi(v))$*;*
3. $\forall v, w \in V(H) : cs_H(v,w) \leq cs_G(\pi(v), \pi(w))$.

5.2.1 Labeled isomorphisms to labeled homeomorphisms

We first present the labeled case, i.e., a $\mathfrak{Iso}, \gamma$-overlap on $\mathcal{G}^{\leftrightarrow}_{\lambda}$ to $\mathfrak{Homeo}, \gamma$-overlap on $\mathcal{G}^{\leftrightarrow}_{\lambda}$ reduction. The reduction R_1 replaces each edge e by an induced subgraph (V_e, E_e) containing the original end vertices and some new vertices. We make sure that no new vertex can be the image of an original vertex by labeling them with a new label.

Formally, let $G \in \mathcal{G}^{\leftrightarrow}_{\lambda}$ with label alphabet Σ. Let $R_1 : G \rightarrow R_1(G) \in \mathcal{G}^{\leftrightarrow}_{\lambda}$ and $e = \{u, w\} \in E(G)$. We define (V_e, E_e) as follows:

$$\begin{aligned} V_e &= \{u, w\} \cup \{v_e^i \mid 1 \leq i \leq 5\}, \\ V_e \cap V(G) &= \{u, w\}, \\ E_e &= \{\{u, v_e^1\}, \{u, v_e^2\}, \{v_e^1, v_e^3\}, \{v_e^2, v_e^3\}, \\ &\quad \{v_e^3, v_e^4\}, \{v_e^3, v_e^5\}, \{v_e^4, w\}, \{v_e^5, w\}\} \end{aligned}$$

Let $e' = \{u', w'\} \in E(G)$. For $e' \neq e$, we make sure that $V_e \cap V_{e'} = \{u, w\} \cap \{u', w'\}$ and $E_e \cap E_{e'} = \emptyset$.

We are now ready to define $R_1(G)$:

$$\begin{aligned} V(R_1(G)) &= V(G) \cup \cup_{e \in E(G)} V_e, \\ E(R_1(G)) &= \cup_{e \in E(G)} E_e, \\ \lambda_{R_1(G)}(u) &= \lambda_G(u), \quad \forall u \in V(G), \\ \lambda_{R_1(G)}(v_e) &= l \notin \Sigma, \quad \forall v_e \in V(R_1(G)) \setminus V(G). \end{aligned}$$

An illustration of the subgraph replacing an edge from a vertex labeled a to a vertex labeled b is shown in Figure 6.

We can prove the following using connection strength:

Theorem 25. $(P, p, G) \rightarrow (R_1(P), R_1(p), R_1(G))$ *is a* $\mathfrak{Iso}, \gamma$*-overlap on* $\mathcal{G}^{\leftrightarrow}_{\lambda}$ *to* $\mathfrak{Homeo}, \gamma$*-overlap on* $\mathcal{G}^{\leftrightarrow}_{\lambda}$ *reduction,* $\gamma \in \{$*vertex, edge*$\}$.

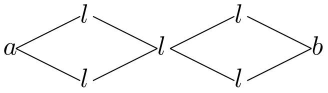


Figure 6. The subgraph (V_e, E_e) used in the reduction to replace an edge $e = \{u, w\}$, with $\lambda(u) = a$ and $\lambda(w) = b$.

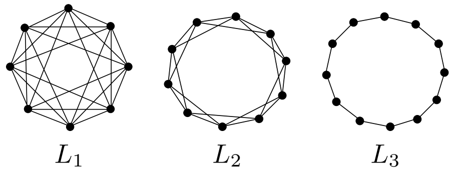


Figure 7. Label graphs for $n = 3$

5.2.2 Labeled to unlabeled homeomorphisms

We now show the reduction from the labeled case to the unlabeled one; i.e., from vertex-overlap of $\mathcal{G}^{\leftrightarrow}_{\lambda}$ to γ-overlap of $\mathcal{G}^{\leftrightarrow}_{\bullet}$ for all $\gamma \in \{\text{vertex}, \text{edge}\}$. We will use the following special label-graphs L_i^n to replace the labels $\Sigma = \{l_1, \ldots, l_n\}$.

Definition 26. *Let $1 \leq s < k$ be integers. C_k^s denotes the graph in $\mathcal{G}^{\leftrightarrow}_{\bullet}$ with nodes $\{c_0, \ldots, c_{k-1}\}$ and edges*

$$E := \{\{c_i, c_j\} \mid j = (i+1) \; mod \; k \ldots (i+s) \; mod \; k\}.$$

Let $1 \leq i \leq n$ be integers. L_i^n denotes the graph $C^{n-i+1}_{2(n+i)}$.

Hence, C_k^s is a cycle of length k, with additional edges: every node is connected to its s successors (and hence also its s predecessors). The graphs L_i^n will be used in the proof to replace labels. In Figure 7, the label graphs for an alphabet of size 3 have been given. Intuitively, we will replace the labels of the vertices by "attaching" an appropriate L_i^n to the node. For a graph over the alphabet $\Sigma = \{l_1, \ldots, l_n\}$, L_i^n will be used to replace l_i. A first essential piece of the proof is that for a given n, no label-graph L_i^n can be mapped to another label-graph L_j^n, $j \neq i$ under homeomorphism.

Lemma 27. *Let $1 \leq i, j \leq n$ be integers. There exists a homeomorphism from L_i^n to L_j^n if and only if $i = j$.*

In all what follows, n will denote the size of the alphabet $\Sigma = \{l_1, \ldots, l_n\}$. We assume a function ι that maps the label l_i to its index i. We will slightly abuse the notation ι, and use $\iota(v)$ to denote $\iota(\lambda(v))$. Let G be a graph in $\mathcal{G}^{\leftrightarrow}_{\lambda}$ over the alphabet Σ, with vertices $V = \{v_1, \ldots, v_k\}$ and edges $E = \{e_1, \ldots, e_m\}$.

In the proofs we will need many copies of the same label graph L_i^n. Therefore, we will need to rename the vertices in these graphs to avoid confusion. We will use $L[v_i]$ to denote the following isomorphic copy of $L^n_{\iota(\lambda(v_i))}$:

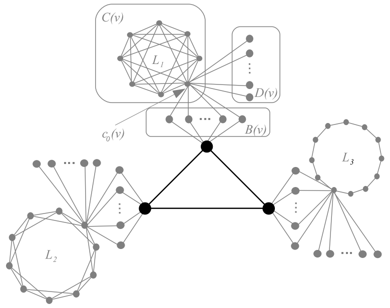


Figure 8. Reduction for removing labels in the case of homeomorphisms. The triangle in the middle is the original graph G. The labels of the top, left, right nodes were respectively l_1, l_2, and l_3.

$V(L[v_i]) = \{c^i_j \mid c_j \in V(L^n_{\iota(v_i)})\}$, and $E(L[v_i]) = \{\{c^i_j, c^i_k\} \mid \{c_j, c_k\} \in E(L^n_{\iota(v_i)})\}$. As such, any two $L[v]$ and $L[w]$ are disjoint whenever $v \neq w$, even if $\lambda(v) = \lambda(w)$. We now define the reduction, parameterized by c.

Definition 28. *For every $v_i \in V(G)$, let the following sets of vertices be given.*

$$\begin{array}{rclcrcl} B(v_i) & = & \{b^i_j \mid j = 1 \ldots c\} & \quad & C(v_i) & = & V(L[v_i]) \\ D(v_i) & = & \{d^i_j \mid j = 1 \ldots c\} & & & & \end{array}$$

We assume all these sets are disjoint. Furthermore, $c_0(v_i)$ denotes the node c^i_0; i.e., the first node in $L[v_i]$.

$R^c_2(G)$ is the following graph in $\mathcal{G}^{\leftrightarrow}_{\bullet}$:

$$\begin{array}{rcl} V(R^c_2(G)) & := & V(G) \cup \bigcup_{v \in V} \Big(C(v) \cup D(v) \cup B(v) \Big) \\ E(R^c_2(G)) & := & E(G) \cup \bigcup_{v \in V} E(L[v]) \\ & & \cup \bigcup_{v \in V} \{\{c_0(v), d\} \mid d \in D(v)\} \\ & & \cup \bigcup_{v \in V} \{\{v, b\}, \{b, c_0(v)\} \mid b \in B(v)\} \end{array}$$

An example of the reduction has been given in Figure 8. Intuitively, the rationale behind the reduction is as follows: the sub-graphs $L[v]$ replace the label of v. The nodes $D(v)$ are added in order to increase the degree of all nodes $c_0(v)$ to at least $2c + 2$. All other nodes have degree at most $2c$. This allows us to use degree-arguments to show that all c_0-nodes are mapped to c_0-nodes. The nodes $B(v)$ are added to connect any label graph $L[v]$ to the right node v (b of between). Their number c will be chosen so we can use connection strength arguments to show that in a match always label-graphs will be associated with the right node.

Using the arguments above, we can prove the following:

Theorem 29. *Let $p, P, G \in \mathcal{G}^{\leftrightarrow}_{\lambda}$, and let $c = \max\{\Delta(G), \Delta(P), \Delta(p), 2n\} + 1$. The function R_2 that maps (p, P, G) to $(R^c_2(p), R^c_2(P), R^c_2(G))$ is an $\mathfrak{Homeo}$, vertex-overlap on $\mathcal{G}^{\alpha}_{\lambda}$ to $\mathfrak{Homeo}$, γ-overlap on $\mathcal{G}^{\alpha}_{\bullet}$ reduction, for all $\alpha \in \{\rightarrow, \leftrightarrow\}$ and $\gamma \in \{$vertex, edge$\}$.*

5.3 From labeled to unlabeled homomorphisms and isomorphisms

Finally, we extend the results for homomorphism and isomorphism to unlabeled graphs. First, we will show that our constructions for labeled graphs can be extended to unlabeled graphs by using special subgraphs in the unlabeled case to encode the labels from the labeled case. We will focus on the most difficult case, homomorphism. For isomorphism, much simpler constructions are possible. Also, we will discuss only the undirected case here. The directed case is analogous.

The key idea for emulating labels with unlabeled subgraphs under homomorphism follows from the fact that cliques are always mapped on cliques of the same size.

Lemma 30. *Let $G \in \mathcal{G}^{\leftrightarrow}$. Let π be a homomorphism from K_k to G (where K_k is the complete graph with k vertices). Then, π is a subgraph isomorphism mapping, i.e. $\pi(K_k)$ is a k-clique of G.*

Apart from the notations introduced in Definition 32, we will also use

$$V^w_j = \{v_{w,j}, v_{w,j+1 \bmod \sigma(w)}, \ldots v_{w,j+K \bmod \sigma(w)}\}.$$

The subgraphs attached to the original vertices to represent the labels are isomorphic to the graphs $C^{\sigma(w)}_K$ as in Definition 26 and are illustrated in Figure 7.

We now formalize the encoding of labels with undirected subgraphs:

Definition 31. *Let $G \in \mathcal{G}^{\leftrightarrow}_{\lambda}$. Let k be (an upper bound on) the clique number of G. A Schema for Labeling with Unlabeled Subgraphs (SLUS) for G is a pair (K, σ) such that*

- *$K \geq \max(k + 2, 2|\Sigma|)$;*
- *$\sigma : \Sigma \rightarrow \mathbb{N}$ is an injective function mapping every element from the alphabet Σ of labels on a distinct odd integer such that*

$$\forall l \in \Sigma : 4(K + 1) < \sigma(l) < 5(K + 1).$$

When it is clear that w is a vertex, we will use $\sigma(w)$ as a shorthand for $\sigma(\lambda_G(w))$. We now define a transformation from labeled to unlabeled graphs:

Definition 32. *Let $G \in \mathcal{G}_\lambda^{\leftrightarrow}$ Let (K, σ) be a SLUS for G. Then, we define the transformed (unlabeled) graph $R_3^{K,\sigma}(G)$ by*

- *the vertices of $R_3^{K,\sigma}(G)$ are*

$$V(R_3^{K,\sigma}(G)) = \cup_{w \in V(G)} V^w$$

where $V^w = \{w_j \mid 0 \leq j < \sigma(w)\}$ where for all w, $w_0 = w$ and $w_j, j = 1 \ldots \sigma(w)$ are new vertices.

- *the edges of $R_3^{K,\sigma}(G)$ are*

$$E(R_3^{K,\sigma}(G)) \quad = \quad E(G) \cup E_{K,\sigma}^{lab}(G)$$

with $E_{K,\sigma}^{lab}(G) = \cup_{w \in V(G)} E^w$ where

$$\begin{aligned} E^w \quad = \quad & \{\{v_{w,j}, v_{w,j+i \textit{ mod } \sigma(w)}\} \mid \\ & 0 \leq j < \sigma(w) \wedge 1 \leq i \leq K\} \end{aligned}$$

We can prove the following

Theorem 33. *Let $p, P, G \in \mathcal{G}_\lambda^{\leftrightarrow}$ such that G has no two adjacent vertices with the same label, and let (K, σ) be a SLUS for p, P and G. Then, the function $R_3^{K,\sigma}$ that maps (p, P, G) to $(R_3^{K,\sigma}(p), R_3^{K,\sigma}(P), R_3^{K,\sigma}(G))$ is a $\mathfrak{Homo}$, vertex-overlap on $\mathcal{G}_\lambda^\alpha$ to $\mathfrak{Homeo}$, γ-overlap on $\mathcal{G}_\bullet^\alpha$ reduction, for all $\alpha \in \{\rightarrow, \leftrightarrow\}$ and $\gamma \in \{vertex, edge\}$.*

To generalize this result so that G is not required to have two adjacent vertices of the same label, we can first perform an additional transformation, splitting all edges in two new edges and labeling the new vertices in the middle of the original edges with a new mid-edge label.

6 Discussion and conclusion

We extended the results of [13] to a range of different settings. We proved the results for labeled homomorphism as a base case and provided reductions which are more generally applicable to prove the results for the other settings.

We showed that *MIS* and *MCP* are minimal and maximal anti-monotonic overlap support measures. We also made a first step towards making the overlap support measures scalable by proving the anti-monotonicity of the Lovász θ-function, a polynomial-time computable graph measure sandwiched between *MIS* and *MCP*.

Several extensions of our work are possible, some of those leading to smaller overlap graphs. An interesting one concerns alternative definitions for matches. We considered matches to be all vertices (edges) of the embedding of a pattern in the dataset. Alternatively, we can consider patterns where only a few distinguished vertices are taken into account for overlap. Making the set of vertices relevant for overlap smaller reduces the size of the overlap graph. The extension is straightforward in most of the cases considered in this paper. As a special case, suppose only one vertex of a pattern is considered relevant. The overlap graph is then reduced to a set of isolated vertices of size at most $|V(G)|$. [2] proposed a measure $f(P, G) = \min_{v \in P} |\{w \in V(G) : \exists \pi \in \mathfrak{Iso} : (\pi(P) \subseteq G) \wedge (w \in V(\pi(P)))\}|$. One can see this as the minimum over several measures, each considering one of the vertices of P relevant. The minimum of anti-monotonic functions is anti-monotonic itself.

There exist also different notions of overlap. E.g. [4] defines harmful overlap, which is based on embeddings. Two embeddings π_1 and π_2 of a pattern P overlap iff $\exists v \in V(P) : \pi_1(v), \pi_2(v) \in \pi_1(V(P)) \cap \pi_2(V(P))$. This notion then results in harmful overlap graphs. We expect our reductions can be easily adapted to generalize also the harmful overlap notion to the considered combinations of directedness, labeledness and morphism choice.

References

[1] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, 16(3):428–435, March 2006.

[2] B. Bringmann and S. Nijssen. What is frequent in a single graph? In *Proceedings of Mining and Learning with Graphs 2007*, Florence, Italy, 2007.

[3] R. Diestel. *Graph Theory, Third edition*. Springer-Verlag, 2005.

[4] M. Fiedler and C. Borgelt. Support computation for mining frequent subgraphs in a single graph. In *Proceedings of Mining and Learning with Graphs 2007*, Florence, Italy, 2007.

[5] J. L. Gross and J. Yellen. *Handbook of Graph Theory*. CRC Press, 2004.

[6] Grunewald, Stefan, Forslund, Kristoffer, Dress, Andreas, Moulton, and Vincent. Qnet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. *Molecular Biology and Evolution*, 24(2):532–538, February 2007.

[7] P. Hell and J. Nešetřil. *Graphs and homomorphisms*. Oxford University Press, 2004.

[8] D. E. Knuth. The sandwich theorem. *Electron. J. Combin.*, 1:48 pp., 1994.

[9] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph. *Data Min. Knowl. Discov.*, 11(3):243–271, 2005.

[10] A. S. LaPaugh and R. L. Rivest. The subgraph homeomorphism problem. In *STOC '78: Proceedings of the tenth annual ACM symposium on Theory of computing*, pages 40–50, New York, NY, USA, 1978. ACM Press.

[11] S. Muggleton and L. De Raedt. Inductive logic programming : Theory and methods. *Journal of Logic Programming*, 19,20:629–679, 1994.

[12] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[13] N. Vanetik, S. E. Shimony, and E. Gudes. Support measures for graph data. *Data Min. Knowl. Discov.*, 13(2):243–260, 2006.