Data Min Knowl Disc (2011) 23:503–548
DOI 10.1007/s10618-011-0217-y

All normalized anti-monotonic overlap graph measures are bounded

Toon Calders · Jan Ramon · Dries Van Dyck

Received: 2 February 2010 / Accepted: 7 February 2011 / Published online: 27 February 2011

Abstract In graph mining, a frequency measure for graphs 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 frequent subgraphs have to be found in one graph. Vanetik et al. (Data Min Knowl Disc 13(2):243–260, 2006) already gave sufficient and necessary conditions for anti-monotonicity of graph measures depending only on the edge-overlaps between the instances 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. As a secondary contribution, we prove that the popular maximum independent set measure assigns the minimal possible normalized frequency and we introduce a new measure based on the minimum clique partition that assigns the maximum possible normalized frequency. In that way, we obtain that all normalized anti-monotonic overlap graph measures are bounded from above and below. We also introduce a new measure sandwiched between the former two based on the polynomial time computable Lovász θ-function.

Responsible editor: Chih-Jen Lin.

T. Calders
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: calders@tue.nl

J. Ramon (✉)
Katholieke Universiteit Leuven, Kortrijk, Belgium
e-mail: Jan.Ramon@cs.kuleuven.be

D. Van Dyck
Hasselt University, Transnational University of Limburg, Diepenbeek, Belgium
e-mail: Dries.VanDyck@UHasselt.be

Keywords Graph mining · Large network setting · Support measure · Anti-monotonicity · Morphism reductions

1 Introduction

Recently, graph mining has emerged as a new field within contemporary data mining that was a focus of interest over the last several years. The central task is to find subgraphs, called *patterns* that occur frequently in either a collection of graphs [e.g. databases of molecules De Raedt and Kramer (2001), game positions Ramon et al. (2000), scene descriptions, …], or in one large graph [e.g. the internet, citation networks Tong et al. (2007), social networks McGlohon et al. (2007), protein interaction networks He and Singh (2007), …]. Especially in the single-graph setting, the notion of frequency, however, is not at all straightforward. For example, the naive 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. Consider e.g. the unlabeled k-clique K_k. There are $\binom{k}{2}$ different embeddings under subgraph isomorphism of the unlabeled path of length 1 in K_k, whereas there are $3\binom{k}{3}$ embeddings of the path of length 2 in K_k. In fact, the number of different embeddings may increase exponentially in the size of the pattern. Hence, as pointed out by Vanetik et al. (2006), a good frequency measure must be such that the frequency of a superpattern is always at most as high as that of a subpattern. This property is called the *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 existing graph pattern miners relies critically on the anti-monotonicity of the frequency measure being used.

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 et al. (2006) 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 between the subgraphs. In Fig. 1, two examples of the overlap graph of a pattern in another graph have been given. Vanetik et al. (2006) described how the overlap graph can change when going from a superpattern to a subpattern by giving three graph operations. That is: one can transform one graph into another graph using the three operations if and only if there exists a database graph, superpattern and subpattern such that the two graphs represent respectively the overlap graph of a super- and of a subpattern. For example, in Fig. 1 the bottom overlap graph (of the superpattern) can be transformed into the top overlap graph (of the subpattern) via the three operations. A direct consequence of this result is that a support measure that is based only on the overlap graph must correspond to a graph property that is increasing under the three operations.

The results of Vanetik et al. (2006) are valid for subgraph isomorphism and labeled graphs. In the context of graph mining, however, not only subgraph isomorphism and labeled graphs are important. On the one hand, the importance of

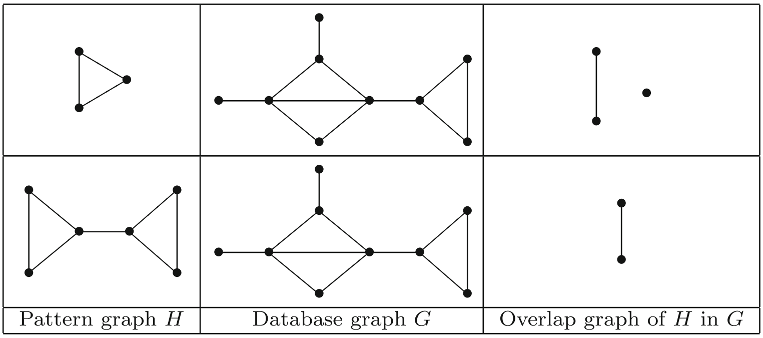


Fig. 1 Two examples of an overlap graph of H in G

homeomorphism-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 and De Raedt 1994). In this paper we extend the results of Vanetik et al. (2006) to these settings as well.

Until now, to our best knowledge, only one overlap graph based support measure has been proposed in the literature (Vanetik et al. 2006). It is based on the computation of the size of a Maximal Independent Set (MIS) of the overlap graph. Unfortunately, computing a Maximal Independent Set in a graph is an NP-hard task. It is therefore natural to consider the question whether a good frequency measure exists which can be computed efficiently. In this paper, we present a first such measure, based on the Lovász function.

The contents of the paper can be summarized as follows:

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.
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.
4. We show that (under reasonable assumptions) the *maximum independent set measure* (MIS) of Vanetik et al. (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. Hence, as a consequence, all anti-monotonic overlap graph measures are bounded both from above and below.
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 frequency measure based on the overlap-graph.

The remainder of this paper is structured as follows. In Sect. 2, we briefly review some basic concepts and notations from graph theory. In Sect. 3 we formalize support measures and overlap graphs and give an overview of earlier results. In Sect. 4, we introduce the MCP measure, prove that it is anti-monotonic and show that, under reasonable assumptions, the MIS measure is minimal and the MCP measure maximal. We also introduce the polynomial time computable Lovász measure which is sandwiched between MIS and MCP and we prove that it is anti-monotonic. Section 5 is devoted to the extension of the results to all 24 considered settings. We first prove a base case, being homomorphic and isomorphic matches in labeled graphs, and then prove reductions which allow for transferring all results to the remaining cases. Section 6 concludes the paper with an overview of the contributions in this article and a discussion of possible extensions for future work.

2 Preliminaries

We assume that the reader is familiar with basic graph theoretic notions and with computational complexity. Textbooks in these areas, such as Diestel (2000) and Papadimitriou (1994) supply the necessary background.

2.1 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 $V \to \Sigma$ 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.

A graph $G = (V, E)$ is said to be a subgraph of graph $H = (V_H, E_H)$, denoted $G \subseteq H$ if $V \subseteq V_H$ and $E \subseteq E_H$.

By $\mathcal{G}$, we denote the class of all graphs; by $\mathcal{G}^{\to}$ ($\mathcal{G}^{\leftrightarrow}$), 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., $\mathcal{G}_{\bullet}^{\to}$ for directed, unlabeled graphs.

For $G \in \mathcal{G}_{\bullet}^{\leftrightarrow}$,

$$\overline{G} := (V(G), \{\{v, w\} \mid v, w \in V\} \setminus E(G))$$

denotes the complement graph of G. By $K_k \in \mathcal{G}_{\bullet}^{\leftrightarrow}$ we denote the *complete graph* on k vertices, i.e.,

$$K_k := (\{v_1, \ldots, v_k\}, \{\{v_i, v_j\} \mid 1 \leq i \neq j \leq k\}).$$

A subgraph $K \subseteq G$ on k vertices for which all vertices are adjacent to all other vertices is called a k-clique. A *cycle* of length k is a connected subgraph on k vertices each of which is incident with exactly two edges.

2.2 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 $V_H \rightarrow V$, such that $\forall (v, w) \in E_H : (\pi(v), \pi(w)) \in E$. 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. The latter is the same as saying that there exists a *subgraph isomorphism* from H to G, which in turn is equivalent with saying that there exists a *vertex-injective homomorphism* from H to G.

A path[1] of length k in G is a sequence of $k+1$ distinct vertices $(v_0, \ldots, v_k)$ with for all $i = 1 \ldots k$, $(v_{i-1}, v_i) \in E$. 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 in 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 in G is denoted P_G, and of all paths with end vertices v and w, $P_G(v, w)$. If $P_G(v, w)$ is not empty, we say that v and w are *connected.*

Following (LaPaugh and Rivest 1978), a *subgraph homeomorphism* π from H to G is the union of an injective mapping from $V(H) \rightarrow V(G)$ and an injective mapping from $E(H) \rightarrow P_G$, such that

(i) for all $(v, w) \in E(H)$ it holds that $\pi((v, w)) \in P_G(\pi(v), \pi(w))$,
(ii) for all $(v, w) \in E(H)$, for all $x \in \pi((v, w))$ and for all $y \in V(H) \setminus \{v, w\}$ it holds that $\pi(y) \neq x$, and
(iii) for all $(v, w), (x, y) \in E(H)$ such that $(v, w) \neq (x, y)$, the paths $\pi((v, w))$ and $\pi((x, y))$ are disjoint.

Moreover, we call a subgraph homeomorphism π *surjective* iff

(i) for all vertices $v' \in V(G)$ either there exists some vertex $v \in V(H)$ such that $v' = \pi(v)$ or there is some edge $e \in E(H)$ such that $v' \in \pi(e)$, and
(ii) for all edges $e' \in E(G)$ there is some edge $e \in E(H)$ such that $e' \in \pi(e)$.

Informally, a homeomorphism differs from an isomorphism by allowing that edges can be mapped to (pairwise disjoint) paths. Hence, every graph obtained from G by replacing edges by paths, an operation called *subdividing* edges, is an image of G under some homeomorphism. Also the other direction holds: every homeomorphic

[1] Remark that we use the definitions of Diestel (2000); that is, *path* coincides with what some authors call a *simple path* and we use *walk* for what is sometimes referred to as a *path* — a sequence of adjacent vertices in which vertices may be repeated.

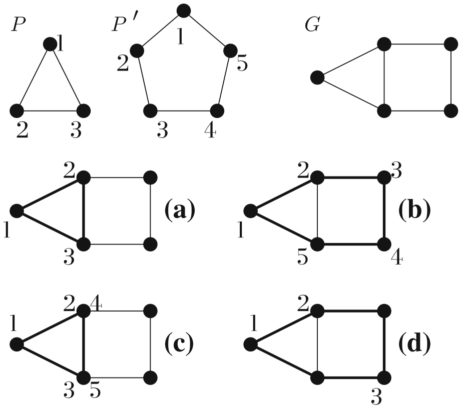


Fig. 2 Examples of the different morphisms. An isomorphic image of P (**a**), P' (**b**), a homomorphic image of P' (**c**) and a homeomorphic image of P (**d**). The edges of the subgraph to which a pattern is mapped are in *bold*. The image of a vertex of the pattern is labeled with its identifier

image of a graph can be obtained by subdivision. For details we refer the interested reader to Diestel (2000).

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

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

Example 1 Figure 2 illustrates the morphisms. The patterns P, P' and the graph G are unlabeled, undirected graphs. Each vertex of the patterns is labeled with an identifier, which is used to pinpoint its image in G. The edges of the subgraph $g \subseteq G$ to which a pattern is mapped are shown in bold. (a) and (b) show an isomorphic image of P resp. P', (c) a homomorphic image of P', and (d) a homeomorphic image of P. 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

As indicated in the introduction, in this paper we consider mining patterns from a single large database graph. In this section, we review and extend the concepts of support measure and overlap graph, which will be central in the rest of this paper.

Definition 2 A *support measure* on $\mathcal{G}^\alpha_\beta$ is a function $f : \mathcal{G}^\alpha_\beta \times \mathcal{G}^\alpha_\beta \to \mathbb{N}$ that maps (P, G) to $f(P, G)$ where P is called the pattern, G is called the database 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. Such an approach requires the support measure being anti-monotonic:

Definition 3 A support measure f on $\mathcal{G}^\alpha_\beta$ is *anti-monotonic* iff $\forall p, P, G \in \mathcal{G}^\alpha_\beta : p \subseteq P \Rightarrow f(P, G) \leq f(p, G)$. That is, the support in a graph G does not increase from a subpattern p to a superpattern P.

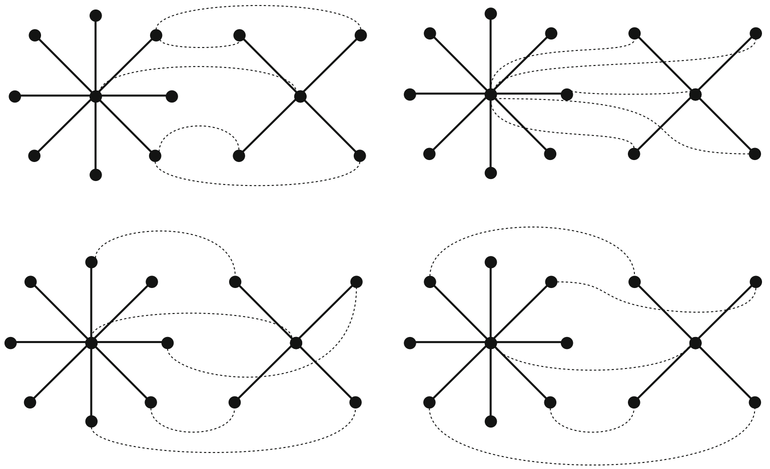


Fig. 3 The *dotted lines* indicate homomorphic embeddings from $Star_4$ (*right*) into $Star_8$ (*left*). Only the two embeddings depicted at the bottom are also $\mathfrak{Iso}$- and $\mathfrak{Homeo}$-embeddings, the two at the *top* are not

Most support measures are based on how often and in what way a pattern P occurs in a graph G. The following definition of a image formalizes the notion of occurrence of a pattern for all types of graphs and morphisms we consider in this paper:

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

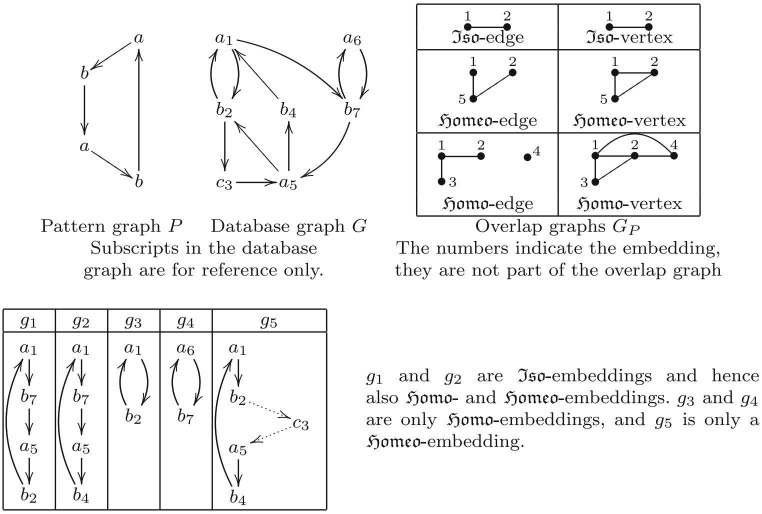
A $\mathfrak{K}$-*image of P in G* is a subgraph $g \subseteq G$ for which there exists a surjective mapping $\pi \in \mathfrak{K}$ from P to g [2]. We call *g the image through π*. An individual mapping π from P to g is called an $\mathfrak{K}$-*embedding* of P in G. $\mathfrak{K}$ is sometimes omitted from the notations if it is clear from the context.

By an *instance* of P in G we refer to a $\mathfrak{Iso}$-image of P in G

Hence, every embedding has exactly one image, but one image can have multiple embeddings.

However, just counting the number of $\mathfrak{K}$-images of a pattern in G does not result in an anti-monotonic support measure, as larger patterns may have more images, instead of less. Consider, e.g., the star graphs $Star_n = (\{v_0, \ldots, v_n\}, \{(v_0, v_i) \mid 1 \leq i \leq n\})$. In Fig. 3, different embeddings of $Star_4$ into $Star_8$ have been illustrated. In general, the number of $\mathfrak{Iso}$-images, $\mathfrak{Homo}$-images and $\mathfrak{Homeo}$-images of $Star_n$ in $Star_N$ (with $N \geq n > 2$) is $\binom{N}{n}$, $N^n + N$ and $\binom{N}{n}$ respectively. Clearly, $\binom{N}{n+1} > \binom{N}{n}$ if $N > 2n + 1$, and always $N^{n+1} + N > N^n + N$. Hence, counting the number of images does not result in an anti-monotonic support measure.

[2] Recall that surjective was defined as *both* vertex- *and* edge-surjective



Pattern graph P Database graph G
Subscripts in the database graph are for reference only.

Overlap graphs G_P
The numbers indicate the embedding, they are not part of the overlap graph

g_1 and g_2 are $\mathfrak{Iso}$-embeddings and hence also $\mathfrak{Homo}$- and $\mathfrak{Homeo}$-embeddings. g_3 and g_4 are only $\mathfrak{Homo}$-embeddings, and g_5 is only a $\mathfrak{Homeo}$-embedding.

Fig. 4 Illustration of the different types of overlap graphs, parameterized by the morphism and overlap type

3.1 Overlap graph

An important class of anti-monotonic measures are the ones that are based on the notion of an *overlap* graph $G_P^{\gamma,\mathfrak{K}}$ (Vanetik et al. 2006; Kuramochi and Karypis 2005)[3]. An overlap graph summarizes not only the images of the pattern in the database graph, but also how they overlap:

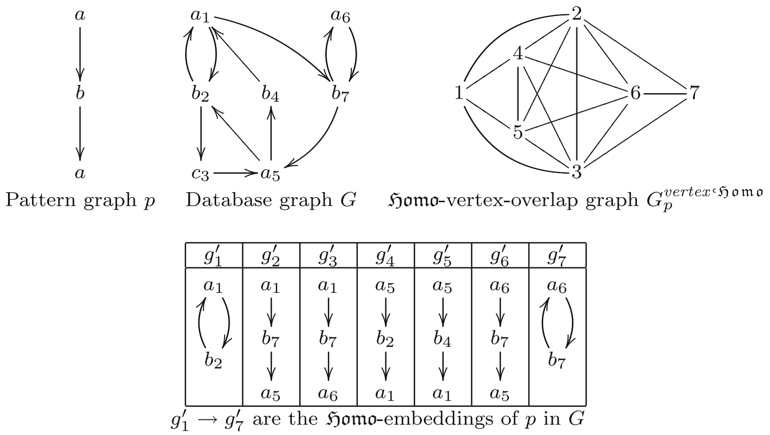
Definition 5 Let $P, G \in \mathcal{G}_\beta^\alpha$, $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$.
Two subgraphs g_1 and g_2 of G have a *vertex-overlap* if $V(g_1) \cap V(g_2) \neq \emptyset$ and an *edge-overlap* if $E(g_1) \cap E(g_2) \neq \emptyset$.

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

Note that $G_P^{\gamma,\mathfrak{K}}$ is always undirected and that the edges depend on the notion of overlap used. For example, $G_P^{\gamma,\mathfrak{K}}$ will be denser for vertex-overlap than for edge-overlap because the latter implies the former. For an illustration of the different overlap graphs, see Fig. 4. For an extra illustration of a $\mathfrak{Homo}$-vertex-overlap graph, see Fig. 5. These two figures together will be used later on as a running example.

Let $p, P, G \in \mathcal{G}_\beta^\alpha, \gamma \in \{\text{vertex, edge}\}$, $\alpha \in \{\rightarrow, \leftrightarrow\}, \beta \in \{\lambda, \bullet\}$, and $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Iso}, \mathfrak{Homeo}\}$. We use the following notation throughout the article:

[3] Vanetik et al. (2006) 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. Therefore we follow the terminology of Kuramochi and Karypis (2005) to avoid confusion.



$g'_1 \rightarrow g'_7$ are the $\mathfrak{Homo}$-embeddings of p in G

Fig. 5 $\mathfrak{Homo}$-vertex-overlap graph of a subpattern p of the pattern P in Fig. 4 in the same database graph G

P represents the (super)pattern, $p \subseteq P$ the subpattern and G the database graph, a single graph. $G_P^{\gamma,\mathfrak{K}}$ ($G_p^{\gamma,\mathfrak{K}}$) is the $\mathfrak{K}$-γ-overlap graph of P (p) in G respectively.

Vanetik et al. (2006) consider three operations on the overlap graph $G_P^{\gamma,\mathfrak{K}}$: clique contraction, edge removal and vertex addition, as defined below.

Definition 6 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 the subgraph $K \subseteq G$ is replaced by a new vertex $k \notin V$ adjacent to $\{w \mid \forall v \in V(K) : \{v, w\} \in E\}$:

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

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

These three operations and how they can be used to transform one overlap graph into another is illustrated by Fig. 6.

The rationale behind these operations is that the $\mathfrak{K}$-γ-overlap graph of a pattern P can be transformed into the $\mathfrak{K}$-γ-overlap graph of a subpattern p of P by means of these operations.

Property 7 *(Vanetik et al. 2006) Let G be a database graph, $p \subseteq P$ two patterns, $\mathfrak{K} \in \{\mathfrak{Homo}, \mathfrak{Homeo}, \mathfrak{Iso}\}$ a morphism type and $\gamma \in \{edge, vertex\}$. $G_P^{\gamma,\mathfrak{K}}$ can be transformed into $G_p^{\gamma,\mathfrak{K}}$ with a sequence of* CC, VA*, and* ER*-operations.*

Proof Technically speaking, the proof of Vanetik et al. (2006) only concerns $\mathfrak{Iso}$,edge-overlap graphs, yet the arguments of Vanetik et al. (2006). also directly apply to the

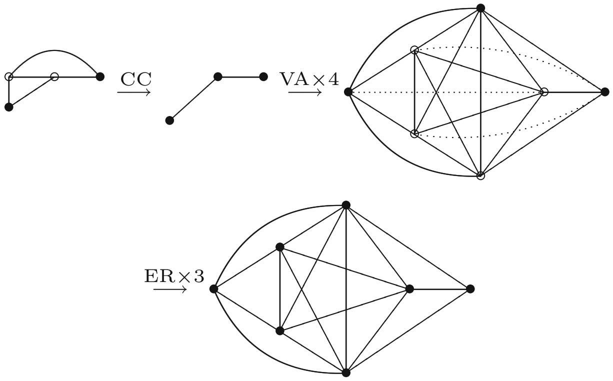


Fig. 6 A series of operations to transform the $\mathfrak{Homo}$-vertex-overlap graph of P in G of Fig. 4 into the $\mathfrak{Homo}$-vertex-overlap graph of p in G of Fig. 5. Vertices that are contracted in a CC-operation as well as the vertices resulting from a VA-operation are depicted as open nodes and edges that are removed by an ER-operation are *dotted lines*

other settings as well, as we will illustrate next. We will use the $\mathfrak{Homo}$-vertex-overlap graphs of Figs. 4 and 1, which represent respectively the overlap graph $G_P^{vertex,\mathfrak{Homo}}$ of a pattern P in a database graph G and the overlap graph $G_p^{vertex,\mathfrak{Homo}}$ a subpattern p of P in the same database graph G. With the three operations, $G_P^{vertex,\mathfrak{Homo}}$ can be transformed into $G_p^{vertex,\mathfrak{Homo}}$.

The property now follows from the next two observations:

1. Any $\mathfrak{K}$-image of P contains a $\mathfrak{K}$-image of p.
2. Let g_1, g_2 be two $\mathfrak{K}$-images of P and $g_1' \subseteq g_1$ and $g_2' \subseteq g_2$ be two $\mathfrak{K}$-images of p. If g_1' and g_2' have a γ-overlap, so do g_1 and g_2.

Indeed, for our running example we have:

Image of P	Contains image of p
g_1	g_2', g_4'
g_2	g_2', g_5'
g_3	g_1'
g_4	g_7'

Since g_1' and g_4' have a vertex-overlap, and g_1 contains g_4' and g_3 contains g_1', g_1 and g_3 overlap as well. Since g_3 and g_4 do not overlap, none of the images of p contained in g_3 can overlap with any of the images of p contained in g_4.

These conditions hold for all settings considered in this article. We quickly sketch the main ideas of the transformation process and refer to Vanetik et al. (2006) for the full details. For reasons of simplicity we assume that p contains at least one edge.

Let $g' \subseteq G$ be a image of p, and let $super(g')$ be all images of P in G containing g'. For our running example this gives:

g'	$super(g')$
g'_1	g_3
g'_2	g_1, g_2
g'_3	
g'_4	g_1

g'	$super(g')$
g'_5	g_2
g'_6	
g'_7	g_4

Because of Observation 1, every image g of P in G must be in at least one $super(g')$. Because of Observation 2, $super(g')$ forms a clique in $G_P^{\gamma,\mathfrak{K}}$, as they all overlap on g'. Furthermore, if there is an edge $\{g'_1, g'_2\}$ in $G_p^{\gamma,\mathfrak{K}}$, there is an edge between any two $g_1 \in super(g'_1)$ and $g_2 \in super(g'_2)$ in $G_P^{\gamma,\mathfrak{K}}$. As such, an induced subgraph of $G_P^{\gamma,\mathfrak{K}}$ can be formed by subsequently contracting the cliques $super(g')$ until for all $g' \in G_p^{\gamma,\mathfrak{K}}$, either $super(g')$ is empty, or a singleton. In our running example, these contractions lead to the second overlap graph of Fig. 6. It is easy to see that one can go from an induced subgraph of $G_P^{\gamma,\mathfrak{K}}$ to $G_p^{\gamma,\mathfrak{K}}$: first add all vertices not in the induced subgraph with vertex additions, and then remove spurious edges with edge removals. Again, for the running example this is illustrated the third and the fourth overlap graphs of Fig. 6. □

3.2 Overlap support measure

We now formally describe what is meant by an overlap support measure.

Definition 8 A *graph measure* is a function $\hat{f} : \mathcal{G}_\bullet^{\leftrightarrow} \to \mathbb{R}$. 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 9 Let $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$, $\gamma \in \{\text{vertex, edge}\}$ 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 $\forall P, G \in \mathcal{G}_\beta^\alpha : f(P, G) = \hat{f}(G_P^{\gamma,\mathfrak{K}})$.

Informally, an overlap support measure is a support measure that only depends on the overlap graph. Consider, e.g., the following measure based on the *maximal independent set* of the overlap graph. An *independent set* of a graph G is a subset I of $V(G)$ such that $\forall v, w \in I : \{v, w\} \notin E(G)$. A *maximal independent set* (MIS) of G is an independent set of maximal cardinality and its size is notated as $mis(G)$. The MIS-based overlap support measure assigns to every pattern P the size of the maximal independent set (MIS) (Vanetik et al. 2006) of $G_P^{\gamma,\mathfrak{K}}$; that is, the support is the maximal number of images that fit in G without overlap. This measure is anti-monotonic. One of the main results of this article is the generalization of the following theorem of Vanetik et al. (2006).

Theorem 10 *(Vanetik et al. 2006)*
Let $\alpha \in \{\rightarrow, \leftrightarrow\}$. Any $\mathfrak{Iso}$-edge-overlap support measure f on $\mathcal{G}^{\alpha}_{\lambda}$ is anti-monotonic if and only if the associated graph measure $\hat{f}$ is increasing under clique contraction, edge removal and vertex addition.

In this paper, we extend this result to the complete space defined by the parameters α, β, $\mathfrak{K}$ and γ. More formally:

Theorem 11 *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 $\mathcal{G}^{\alpha}_{\beta}$ 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 $G^{\gamma,\mathfrak{K}}_{P}$ can be transformed into $G^{\gamma,\mathfrak{K}}_{p}$ by these operations (cfr. Property 7).

To prove necessity, Vanetik et al. (2006) construct for every unlabeled graph H and every operation o, a triple (P, p, G) (where P is a superpattern, p a subpattern and G a database graph. such that $G^{\gamma,\mathfrak{K}}_{P} \cong H$ and $G^{\gamma,\mathfrak{K}}_{p} \cong o(H)$. Henceforth, if f would be non-increasing under some $o \in \{\mathsf{CC}, \mathsf{ER}, \mathsf{VA}\}$, then 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 will follow the same approach in Sect. 5.

4 Minimal and maximal overlap support measures

In this section we show that there exist natural minimal and maximal support measures such that every *normalized* overlap support measure is always between these two extremes. A normalized support measure is defined as follows.

Definition 12 Let $G \in \mathcal{G}^{\leftrightarrow}_{\bullet}$ be an undirected graph and $\overline{K_k}$ the graph composed of k isolated vertices.

We call an overlap support measure f *normalized* if it is anti-monotonic and assigns the frequency k to k non-overlapping images, i.e., $\hat{f}(\overline{K_k}) = k$.

Up to now, all normalized overlap support measures f we are aware of are *MIS*-measures, i.e., the support of $f(P, G) = mis(G^{\gamma,\mathfrak{K}}_{P})$. *MIS* was introduced and proven to be anti-monotonic in Vanetik et al. (2006). A more compact proof can be found in Fiedler and Borgelt (2007).

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 complete graph in G. A *minimum clique partition* (MCP) is a clique partition of minimum cardinality. Its cardinality is denoted $mcp(G)$.

The *MCP-measure* is defined by $MCP(P, G) : (P, G) \mapsto mcp(G_P^{\gamma, \mathfrak{K}})$.

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* $\mathcal{G}_\beta^\alpha$.

Proof By Theorem 11, it is sufficient to show that mcp is not decreasing under CC, VA, ER. Let $H \in \mathcal{G}_\bullet^{\leftrightarrow}$ and $\{V_1, \ldots, V_k\}$ be an MCP for H.

Clearly, removing an edge (ER) can only not decrease $mcp(H)$. $mcp(H)$ is invariant under VA: as the new vertex $v \notin V(H)$ is adjacent to all other vertices, it can be added to any partition of an MCP for H or removed from any partition of an MCP for $\mathsf{VA}(H)$ to obtain an MCP of the same size for $\mathsf{VA}(H)$ respectively H.

Let K be a clique in H that is contracted into $v_k \notin V(H)$, the new vertex in $H' = \mathsf{CC}(H, K)$. Assume that $\{V_1', \ldots, V_{k'}'\}$ is an MCP for H'. Without loss of generality we can assume that $v_k \in V_1'$. Since V_1' is a complete graph in H', so is $V_1^* = V_1' \setminus \{v_k\} \cup K$ in H. Moreover, $V_2', \ldots, V_{k'}'$ are also cliques in H. So $\{V_1^*, V_2', \ldots, V_k'\}$ is a MCP for H, and therefore $mcp(H) \leq mcp(H')$. We can conclude that mcp is not decreasing under clique contraction. □

It is interesting to compare MCP with MIS. Let $\chi(G)$ be the *chromatic number* of G, i.e., the minimal number of colors needed 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 and Yellen 2004, Sect. 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 image of V_i, MIS forces us to choose either no image or exactly one image v_i, which must be independent of all chosen $v_j \in V_j$. MCP, however, allows us to count a image in V_i as soon there is *a* image in V_i which does not overlap with *a* image in V_j. That is, we can make another choice for each (V_i, V_j) pair.

Example 15 Let us look at an example: consider pattern P and the graph G as shown in Fig. 7. The 5 $\mathfrak{Iso}$-images 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 $G_P^{edge, \mathfrak{Iso}}$ of P in G is shown on the right in Fig. 7 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 $G_P^{edge, \mathfrak{Iso}}$. Hence, if we count image 1 with MIS, we can only take image 3 or image 4 as second independent image, because 3 and 4 overlap, leading to a MIS-support of 2. This is a bit unnatural, because each of the 3 images of the triangle can be extended to a image of P in a way that they do not overlap with each other, which would lead to a support of 3 of P.

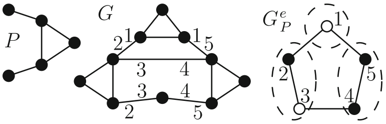


Fig. 7 *Left*: a pattern P and a graph G. The 5 $\mathfrak{Iso}$-images of P in G are indicated by the image in G of the edges outside the *triangle*. *Right*: the $\mathfrak{Iso}$-edge-overlap graph $G_P^{edge,\mathfrak{Iso}}$ with a *MCP* (*dashed ellipses*) and a *MIS* (*white vertices*)

This more natural notion of counting independent images is exactly what *MCP*-support allows us to do: we do not count individual images, but groups of images of P sharing a image of a subpattern p (a triangle) and allow to "switch" images 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 image 3 (respectively image 4) and the group {2, 3} is independent of the group {4, 5} because, for instance, image 2 and image 5 do not overlap.

4.2 Bounding theorem

Interestingly, *MIS* and *MCP* turn out to be the minimal and the maximal possible normalized overlap measures. The following theorem is one of our main results:

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 normalized $\mathfrak{K}$-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 11 to show both the minimality of *MIS* and the maximality of *MCP*.

Let $H = G_P^{\gamma,\mathfrak{K}}$, 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 normalized, 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 did, then joining the two cliques that were contracted to two adjacent vertices of H_{cc} would give us a smaller clique partition. Because f is anti-monotonic, $\hat{f}$ is not decreasing under CC and thus

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

This bounding theorem still leaves a lot of room to define support measures, as there can be an arbitrarily large gap between MIS and MCP (Brimkov 2004).

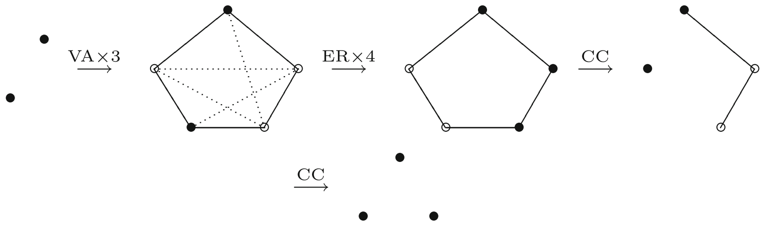


Fig. 8 Illustration of a sequence of operations to move from a maximal independent set to the overlap graph of Fig. 7 to the minimal clique partition

Example 17 Consider again the example given in Fig. 7. $mis(G_P^{edge,\mathfrak{Iso}}) = 2$ and $mcp(G_P^{edge,\mathfrak{Iso}}) = 3$. Hence, every normalized $\mathfrak{Iso}$, edge based support measure must assign a value between 2 and 3 for P in G. Indeed, as illustrated in Fig. 8, $\overline{K_2}$ can be transformed into $G_P^{edge,\mathfrak{Iso}}$ and $G_P^{edge,\mathfrak{Iso}}$ can on its turn be transformed into $\overline{K_3}$.

4.3 A polynomial time computable support measure

In the previous section we have established upper and lower bounds for normalized overlap measures. However, 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 normalized 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 (Knuth 1994). There are several equivalent characterizations of this function. Here, we will use a characterization which is convenient for our proof that it is a normalized measure.

In the following, we will associate a vector with each vertex of a graph. For notational convenience, we will assume that $V(G) = \{1, \ldots, n\}$ so that we can use a vertex as an index of a vector. We will denote cell i of a vector x by x_i.

If x is a vector, then $c(x) = 0$ if $x = 0$ (the nul-vector), else

$$c(x) = x_1^2 / \sum_i x_i^2$$

An orthogonal labeling a of a graph G is a function from $V(G)$ to vectors such that the vectors associated with u and v are orthogonal (i.e. $a(u) \cdot a(v) = 0$) whenever $\{u, v\} \notin E(G)$.

Let G be an undirected unlabeled graph. One can now define $\theta(G)$ as

$$\theta(G) = \max \left\{ \sum_u c(a(u)) \mid a \text{is an orthogonal labeling of } \overline{G} \right\}$$

Example 18 Consider the class of cycles of odd length n. The triangle is a clique and hence its MIS, MCP and θ all equal 1. For odd $n \geq 5$, the size of the minimum clique partition equals $(n+1)/2$ and a maximum independent set has size $(n-1)/2$. For any odd integer $n \geq 3$, the theta function of a cycle of length n is (Crespi 2004)

$$n\frac{\cos(\pi/n)}{1+\cos(\pi/n)}.$$

One can see that for small n, θ is close to MIS while for large n, $\theta - n/2$ converges to 0, so θ is halfway between MIS and MCP for large n.

We now prove the following result:

Theorem 19 *θ is a normalized overlap measure.*

Proof By Theorem 11, it is sufficient to show that θ is not decreasing under CC, VA and ER to prove that θ is anti-monotonic.

Let $G \in \mathcal{G}_{\bullet}^{\leftrightarrow}$ be an arbitrary graph. For all three operations CC, VA, and ER we will show a construction that transforms an orthogonal labeling b for G into a labeling b' for the result G' of the operation on G such that $\sum_{u\in V(G)} c(b(u)) = \sum_{u\in V(G')} c(b(u))$. Since θ is defined as the maximum over all orthogonal labelings a this construction effectively proves that $\theta(G') \geq \theta(G)$.

VA Let $G' = \mathsf{VA}(G, v')$. Let $b : V(G) \to \mathbb{R}^n$ be an orthogonal labeling of $\overline{G}$ maximizing $\sum_{u\in V(G)} c(a_u)$. Then, let b' be defined by

$$\begin{aligned} \forall v \in V(G) : b'(v) &= b(v) \\ b'(v') &= 0_n \end{aligned}$$

where 0_n is a $n \times 1$ vector of zeroes. b' is an orthogonal labeling of $\overline{G'}$, since for each $v \in V(G') \setminus \{v'\}$, $\{v, v'\} \notin E(\overline{G'})$. Therefore $\theta(G') \geq \theta(G)$.

ER Let $G' = \mathsf{ER}(G, e) = (V(G), E(G) \setminus \{e\})$. Let b be an orthogonal labeling of $\overline{G}$ maximizing $\sum_u c(b(u))$. b is an orthogonal labeling of $\overline{G'}$, as for every edge $\{u, v\} \notin E(\overline{G'})$, we have $\{u, v\} \in E(G')$, $\{u, v\} \in E(G)$ and $\{u, v\} \notin E(\overline{G})$ and therefore $b(u) \cdot b(v) = 0$. We can conclude that $\theta(G') \geq \sum_{u\in V(G)} c(b(u)) = \theta(G)$.

CC Let $C = \{v_1 \dots v_n\}$ be a clique of G. Let $G' = CC(G, C, v)$, i.e. C is contracted into v. Let b be a orthogonal labeling of $\overline{G}$ that maximizes $\theta(G)$, so

$$\theta(G) = \sum_{u\in V(G)} c(b(u))$$

We can assume $\|b(u)\| = 1$ for all v (scaling a vector does not change $c(b(v))$). Now consider the matrix $C_b = [b(v_1) \dots b(v_n)]$. We can see that:

- for all pairs of distinct vertices v_i and v_j of C, $v_i, v_j \in C$ are adjacent, i.e. $\{v_i, v_j\} \notin V(\overline{G})$ and hence $b(v_i) \cdot b(v_j) = 0$
- for all $v_i \in C$, we have $b(v_i) \cdot b(v_i) = \|b(v_i)\| = 1$.
- Therefore, $C_b^T \cdot C_b = I_{|C|}$.

C_b is a $d \times n$ matrix with $d \geq n$. C_b can be extended into a square $d \times d$ unitary matrix $Q = [C_b\ C_b^{\perp}]$, i.e. $Q^T Q = I_d = QQ^T$. Let $C_b^{\perp} = [b(v_{n+1}) \dots b(v_d)]$ (for some new $v_{n+1} \dots v_d$). Let e be the unit vector with $e_1 = 1$ and $e_i = 0$ for $i \neq 0$. Notice that $c(x) = (x \cdot e)^2$ for any vector x with $\|x\| = 1$. Let

$$p_C = \sum_{w \in C} (b(w) \cdot e).b(w)$$

and

$$e_C = \frac{p_C}{\|p_C\|}$$

It is clear that e_C is the unit vector in the subspace spanned by C_b for which the first component is maximal (because p_C is the projection of e in this space spanned by C_b).

Moreover, we have:

$$\begin{aligned} c(e_C) &= (e_C \cdot e)^2 \\ &= ((p_C/\|p_C\|) \cdot e)^2 \\ &= (p_C \cdot e)^2/\|p_C\|^2 \end{aligned}$$

where

$$\begin{aligned} p_C \cdot e &= \left(\sum_{w \in C} (b(w) \cdot e).b(w) \right) \cdot e \\ &= \sum_{w \in C} (b(w) \cdot e)^2 \\ &= \sum_{w \in C} c(b(w)) \end{aligned}$$

and also (since the $b(w)$ are orthonormal)

$$\begin{aligned} \|p_C\|^2 &= \| \sum_{w \in C} (b(w) \cdot e).b(w) \|^2 \\ &= \sum_{w \in C} (b(w) \cdot e)^2 \\ &= \sum_{w \in C} c(b(w)) \end{aligned}$$

Therefore,

$$\begin{aligned} c(e_C) &= (p_C \cdot e)^2/\|p_C\|^2 \\ &= \frac{\left(\sum_{w\in C} c(b(w))\right)^2}{\sum_{w\in C} c(b(w))} \\ &= \sum_{w\in C} c(b(w)) \end{aligned}$$

Let b' be the labeling of $\overline{G'}$ where $b'(u) = b(u)$ if $u \notin C$ and

$$b'(v) = e_C = \sum_{w\in C} \frac{b(w)\cdot e}{\|p_C\|} b(w)$$

with v the new vertex of G'. So we have $c(b'(v)) = \sum_{w\in C} c(b(w))$.
b' is an orthogonal labeling of $\overline{G'}$. Indeed, for $u_1, u_2 \in V(G)\setminus C$, we have $b'(u_1)\cdot b'(u_2) = b(u_1)\cdot b(u_2) = 0$ whenever $\{u_1, u_2\} \notin E(\overline{G})$. For $u \in V(G)\setminus C$, $\{u, v\} \notin E(\overline{G'})$ iff $\{u, w\} \notin E(\overline{G})$ for all $w \in C$. In that case, $b(u)\cdot b(w) = 0$ for all $w \in C$, and hence

$$\begin{aligned} b(u)\cdot b(v) &= b(u)\cdot \sum_{w\in C} \frac{b(w)\cdot e}{\|p_C\|} b(w) \\ &= 0 \end{aligned}$$

Therefore, b' is an orthogonal labeling for $\overline{G'}$. Moreover,

$$\begin{aligned} \sum_{u\in V(G')} c(b'(u)) &= \sum_{u\in V(G)\setminus C} c(b(u)) + c(b'(v)) \\ &= \sum_{u\in V(G)\setminus C} c(b(u)) + \sum_{u\in C} c(b(u)) \\ &= \theta(G) \end{aligned}$$

and hence $\theta(G') \geq \theta(G)$.

Since θ is sandwiched between *mis* and *mcp*, and $mis(\overline{K_k}) = mcp(\overline{K_k}) = k$, it follows immediately that θ is normalized. □

5 Necessity for other morphisms and graph settings

In the previous sections we considered graph properties on overlap graphs that are not decreasing under the operations VA, ER and CC. Property 7 showed that these conditions are sufficient for the support measure based on this graph property to be an overlap-graph measure to be anti-monotonic. Vanetik et al. (2006) showed, for the case of labeled graphs with isomorphic images and edge-overlap that this condition is also necessary. In this section, we generalize this result to all 24 combinations of morphism type $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, overlap type $\gamma \in \{\mathit{edge}, \mathit{vertex}\}$, directed/undirected $\alpha \in \{\rightarrow, \leftrightarrow\}$, and labeled/unlabeled $\beta \in \{\bullet, \lambda\}$. We will denote such a combination, or case, by $(\mathfrak{K}, \gamma, \alpha, \beta)$.

For some cases we will use a direct proof, based on the following notion of completeness.

Definition 20 A case $(\mathfrak{K}, \gamma, \alpha, \beta)$ is called *complete* if and only if for every graph $H \in \mathcal{G}_{\bullet}^{\leftrightarrow}$, and every operation $o \in \{\mathsf{ER}, \mathsf{VA}, \mathsf{CC}\}$, there exist $p, P, G \in \mathcal{G}_{\beta}^{\alpha}$ such that $p \subseteq P$, $G_P^{\mathfrak{K},\gamma} = H$ and $G_p^{\mathfrak{K},\gamma} = o(H)$.

If $(\mathfrak{K}, \gamma, \alpha, \beta)$ is a complete case, any $\mathfrak{K}, \gamma$ overlap-graph based support measure f on $\mathcal{G}_{\beta}^{\alpha}$ that is anti-monotone will be non-decreasing under the three operations $\mathsf{ER}, \mathsf{VA}, \mathsf{CC}$. This is easy to see as follows: given H and o, let p, P, G be as in the definition. Then,

$$\hat{f}(H) = \hat{f}(G_P) = f(P, G) \leq f(p, G) = \hat{f}(G_p) = \hat{f}(o(H)),$$

where $\hat{f}$ denotes the (overlap) graph measure associated with f. This strategy was used by Vanetik et al. (2006) for the cases $(\mathfrak{Iso}, \mathit{edge}, \leftrightarrow, \lambda)$ and $(\mathfrak{Iso}, \mathit{edge}, \rightarrow, \lambda)$. In Sect. 5.2, we will use this same strategy to show, in a generic way, the necessity for all cases $(\mathfrak{K}, \gamma, \alpha, \beta)$ with $\alpha = \lambda$ and $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}\}$.

Another key construction in the proofs of this section will be that of a *reduction*. A reduction maps a triplet (p, P, D) for one case to a triplet (p', P', D'), maintaining subpattern relations and overlap graphs. All cases to which a complete case can be reduced will be complete as well, and hence, the necessity condition holds.

The structure of this section is now as follows: first, we formally introduce the notion of a reduction and show that it effectively carries over completeness from one case to another. As an easy start we show how to reduce unlabeled to labeled cases, and undirected to directed cases. After that, it will be shown that the labeled homo- and isomorphisms cases are complete, and finally the proofs are completed by giving and proving various reductions making all 24 cases reachable from at least one complete case. Figure 9 serves as a road map for this section, giving a complete picture of the different theorems and proofs in this section.

5.1 Reductions

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

Definition 21 Let $\mathfrak{K}, \mathfrak{K}' \in \{\mathfrak{Iso}, \mathfrak{Homo}, \mathfrak{Homeo}\}$, $\alpha, \alpha' \in \{\rightarrow, \leftrightarrow\}$, $\beta, \beta' \in \{\bullet, \lambda\}$, and $\gamma, \gamma' \in \{\text{edge}, \text{vertex}\}$.

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

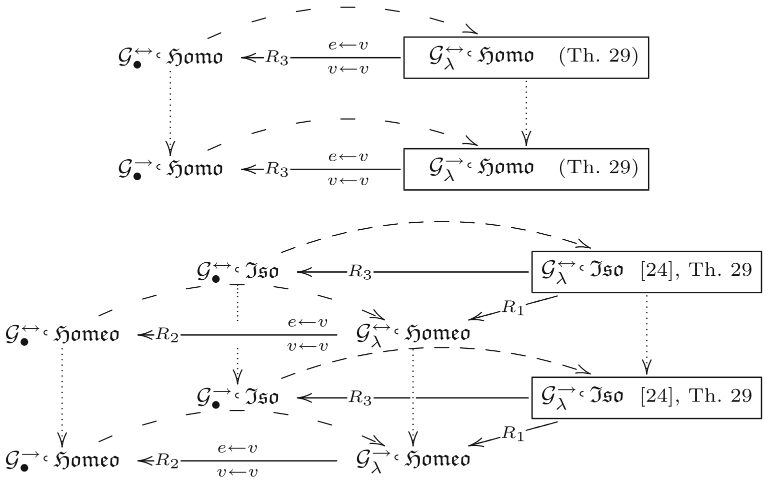

$\mathcal{G}_{\bullet}^{\leftrightarrow} \subseteq \mathfrak{Homo}$
R_3
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}_{\lambda}^{\leftrightarrow} \subseteq \mathfrak{Homo}$ (Th. 29)
$\mathcal{G}_{\bullet}^{\rightarrow} \subseteq \mathfrak{Homo}$
R_3
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}_{\lambda}^{\rightarrow} \subseteq \mathfrak{Homo}$ (Th. 29)
$\mathcal{G}_{\bullet}^{\leftrightarrow} \subseteq \mathfrak{Iso}$
R_3
$\mathcal{G}_{\lambda}^{\leftrightarrow} \subseteq \mathfrak{Iso}$ [24], Th. 29
R_1
$\mathcal{G}_{\bullet}^{\leftrightarrow} \subseteq \mathfrak{Homeo}$
R_2
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}_{\lambda}^{\leftrightarrow} \subseteq \mathfrak{Homeo}$
$\mathcal{G}_{\bullet}^{\rightarrow} \subseteq \mathfrak{Iso}$
R_3
$\mathcal{G}_{\lambda}^{\rightarrow} \subseteq \mathfrak{Iso}$ [24], Th. 29
R_1
$\mathcal{G}_{\bullet}^{\rightarrow} \subseteq \mathfrak{Homeo}$
R_2
$e \leftarrow v$
$v \leftarrow v$
$\mathcal{G}_{\lambda}^{\rightarrow} \subseteq \mathfrak{Homeo}$

Fig. 9 Overview of the reductions. R_1, R_2, and R_3 are proven respectively in Theorems 33, 41, and 50; the *dashed* and the *dotted arrows* in Proposition 23. *Arrows* representing reductions changing γ-overlap into γ'-overlap are labeled with $\gamma \to \gamma'$

(1) $p \subseteq P$ if and only if $p' \subseteq P'$, and
(2) $G_p^{\gamma,\mathfrak{K}} \cong G'^{\gamma',\mathfrak{K}'}_{p'} \wedge G_P^{\gamma,\mathfrak{K}} \cong G'^{\gamma',\mathfrak{K}'}_{P'}$.

Recall that the overlap graph is defined w.r.t. images of the pattern in the database graph. Therefore, 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 importance of reductions lies in the following theorem, stating that the reductions carry over completeness, and hence prove the necessity condition:

Theorem 22 *Let R be a $(\mathfrak{K}, \gamma, \alpha, \beta)$ to $(\mathfrak{K}', \gamma', \alpha', \beta')$ reduction. If $(\mathfrak{K}, \gamma, \alpha, \beta)$ is complete, then $(\mathfrak{K}', \gamma', \alpha', \beta')$ is complete as well.*

Proof Let $H \in \mathcal{G}_{\bullet}^{\leftrightarrow}$, and $o \in \{\mathsf{ER}, \mathsf{VA}, \mathsf{CC}\}$. Since $(\mathfrak{K}, \gamma, \alpha, \beta)$ is complete, there exist $p, P, G \in \mathcal{G}_{\beta}^{\alpha}$ such that $p \subseteq P$, $G_P^{\mathfrak{K},\gamma} = H$ and $G_p^{\mathfrak{K},\gamma} = o(H)$. Since R is a reduction, for $(p', P', G') = R(p, P, G)$, also $p' \subset P'$ and $G'^{\mathfrak{K}',\gamma'}_{P'} = H$ and $G'^{\mathfrak{K}',\gamma'}_{p'} = o(H)$. Since H and o are arbitrary, this shows that $(\mathfrak{K}', \gamma', \alpha', \beta')$ is complete as well. □

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

Property 23 *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:*

1. *from $(\mathfrak{K}, \gamma, \alpha, \bullet)$ to $(\mathfrak{K}, \gamma, \alpha, \lambda)$, and*
2. *from $(\mathfrak{K}, \gamma, \leftrightarrow, \beta)$ to $(\mathfrak{K}, \gamma, \rightarrow, \beta)$.*

Proof 1. (unlabeled to labeled) is straightforward; the function that labels every node in all graphs in the unlabeled triplet (p, P, G) with the same label clearly is a reduction.

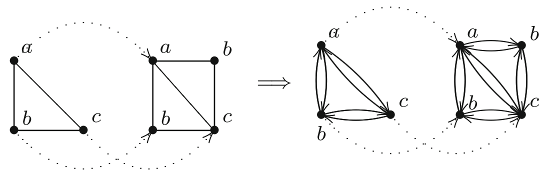


Fig. 10 Illustration of the reduction in Property 23 from undirected to directed for iso- and homomorphism. The *left* figure shows a homeomorphic embedding of a triangular pattern graph into the data graph (undirected setting) and the corresponding homeomorphic embedding in the directed setting

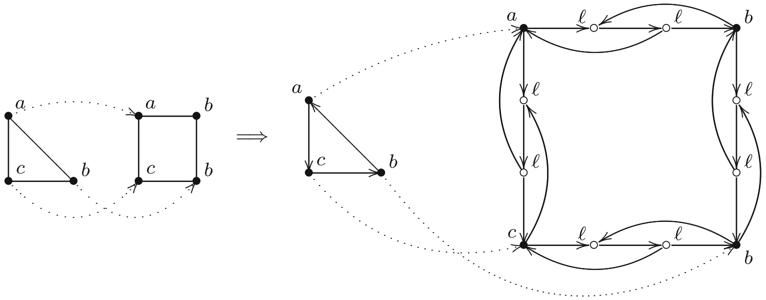


Fig. 11 Illustration of the reduction in Property 23 from undirected to directed for homeomorphism. The left figure shows a homeomorphic embedding of a triangular pattern graph into the data graph (undirected setting) and the corresponding homeomorphic embedding in the directed setting

2. (undirected to directed) We discriminate two cases. $\mathfrak{K} \in \{\mathfrak{Iso}, \mathfrak{Homo}\}$: This reduction is illustrated in Fig. 10. Let G be an undirected graph, and let G^d denote the following directed graph: $V(G^d) = V(G)$, $\lambda_{G^d} = \lambda_G$ in case G is labeled, and $E(G_d) = \{(v, w), (w, v) \mid \{v, w\} \in E(G)\}$; hence, every edge $\{v, w\}$ is replaced by the pair of arcs (v, w) and (w, v). The function that maps a triplet of undirected graphs (p, P, G) to (p^d, P^d, G^d) is a reduction. This follows easily from the fact that any iso- and homomorphism π is completely determined by the image of the vertices alone, and that there are arcs (v, w) and (w, v) in G^d if and only there is an edge $\{v, w\}$ in G.

$\mathfrak{K} = \mathfrak{Homeo}$: For homeomorphisms, the reduction requires somewhat more work, as an homeomorphism is not completely determined by the image of the vertices, as edges (arcs) are mapped to paths. If we keep the same construction as above, before the reduction we can map the edge $a - c$ to the path $a - b - c$, yet the two edges $a \rightarrow c, c \rightarrow a$ in the reduced pattern cannot be mapped to the paths $a \rightarrow b \rightarrow c$ and $c \rightarrow b \rightarrow a$ at the same time, as this would result in the inner vertex b being shared by both paths. Only keeping one direction for the pattern and maintaining both directions for the database graph does not result in a valid reduction either as edge-overlap may be lost between two images using the edge $a \rightarrow b$ and $b \rightarrow a$ respectively. A construction that does work is given below and is illustrated in Fig. 11.

We will show the reduction only for labeled graphs, because later on, we will show how to simulate labeled graphs with unlabeled graphs (The proofs there do not depend on the property we are proving now). For an undirected, labeled graph G, let G' denote the following graph: for every edge $e \in E(G)$, we introduce two new nodes v_e^1 and

v_e^2. $V(G') = V(G) \cup \bigcup_{e \in E(G)} \{v_e^1, v_e^2\}$. Let ℓ be a new label, not used in G. For all $v \in V(G)$, $\lambda_{G'}(v) = \lambda(v)$, and $\lambda_{G'}(v_e^i) = \ell$, for all $e \in E(G)$, $i = 1, 2$. The set of edges $E(G')$ is defined as

$$\bigcup_{e=\{v,w\} \in E(G)} \{(v, v_e^1), (w, v_e^1), (v_e^1, v_e^2), (v_e^2, v), (v_e^2, w)\}.$$

Hence, G' is G in which every edge $\{v, w\}$ has been replaced by 5 arcs as follows:

$$\begin{tikzcd} v \arrow[r] & v_e^1 \arrow[r] & v_e^2 \arrow[r] \arrow[ll, bend left] & w \arrow[ll, bend right] \end{tikzcd}$$

The newly introduced nodes can be recognized by their new label ℓ. The patterns p and P are transformed into p' and P' by replacing every edge $\{v_i, v_j\}, i < j$ by the arc (v_i, v_j). (We assume an arbitrary order of the vertices of the patterns which is consistent with their subpattern relationship.)

The function that maps (p, P, G) to (p', P', G') is a reduction: because of the uniqueness of the new label ℓ, a homeomorphism from p' (P') to G' must map nodes in p' (P') to nodes in $V(G)$. Furthermore, $q = (v_1, v_2, \ldots, v_p)$ is a path in G if and only if $q' = (v_1, v_{e_1}^1, v_{e_1}^2, v_2, v_{e_2}^1, v_{e_2}^2, v_3, \ldots, v_{e_{p-1}}^1, v_{e_{p-1}}^2, v_p)$, with $e_i = (v_i, v_{i+1}), i = 1 \ldots p-1$ is a path in G'. Also, all paths in G' between nodes in $V(G)$ are of the form

$$(v_1, v_{e_1}^1, v_{e_1}^2, v_2, v_{e_2}^1, v_{e_2}^2, v_3, \ldots, v_{e_{p-1}}^1, v_{e_{p-1}}^2, v_p),$$

with $e_i = (v_i, v_{i+1}), i = 1 \ldots p-1$. Therefore, π is a homeomorphism from p to G if and only if the following π' is a homeomorphism from p' to G': for all $v \in V(p'), \pi'(v) = \pi(v)$, and for all arcs $(v, w) \in p', \pi'((v, w)) = q'$ if $\pi(\{v, w\}) = q$. Hence, any image g from p in G corresponds uniquely to a image g' of p' in G'. Last but not least: two images g_1 and g_2 of p in G are vertex- (edge-) disjoint if and only if g_1' and g_2' are vertex (edge-) disjoint. For vertex-disjoint this follows from the fact that if two paths q_1 and q_2 overlap in $v \in V(p)$ if and only if q_1' and q_2' overlap in v, and q_1' and q_2' overlap in v_e^1 or v_e^2 if and only if q_1' and q_2' overlap in v and w, where $e = \{v, w\}$. For edge-disjointness, it suffices to notice that if two paths q_1 and q_2 edge-overlap in $e = \{v, w\}$, then q_1' and q_2' overlap on the arc (v_e^1, v_e^2). □

The reductions we will prove later on in this article are similar in spirit in the sense that we will replace vertices and/or edges with certain structured subgraphs to enforce certain properties in the transformed graphs which can be used to limit the potential images of vertices and/or edges. An overview of the reductions which follow from our proofs are shown in Fig. 9.

5.2 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 $\mathfrak{K}$-γ-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

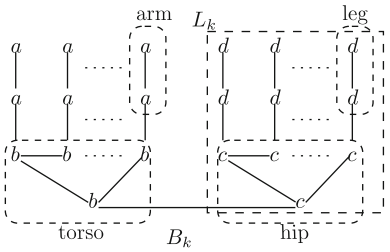


Fig. 12 The graphs B_k and L_k

not rely on reductions, but show the necessity directly for these cases. We show only the undirected case, as for 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 23.

The general idea of the proof is to construct a superpattern P composed of three parts: a collection of *arms*, a collection of *legs*, a *hip* and a *torso*. An example of such a pattern is given in Fig. 12. The subpattern p will be composed only of the hip and legs.

We than construct a database graph G composed of a number of instances of P and p which overlap in some of these parts. The number of arms and legs is sufficiently high so that we can choose for any pair of images of the superpattern respectively subpattern to share an arm respectively a leg. In that way, we can choose for any pair of vertices of the overlap graph of P if they are connected by an edge. If we want the same edge to be present in the overlap graph of p we use a leg and otherwise we use an arm. If we want a k-clique in the overlap graph of P we let k instances of P overlap in one instance of p (hip and legs) but not in the torso and arms.

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 to force each subgraph homomorphism into an isomorphism:

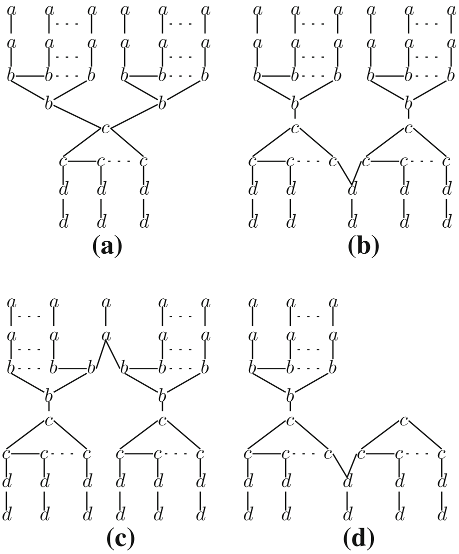
Proposition 24 (Hell and Nešetřil 2004) *If $H, G \in \mathcal{G}^{\alpha}_{\beta}$, $\alpha \in \{\rightarrow, \leftrightarrow\}$, $\beta \in \{\lambda, \bullet\}$ for which there exists a homomorphism $\pi : H \rightarrow G$, then*

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 from Proposition 24 implies that a cycle of length k, with k odd, can only be mapped by a homomorphism to a cycle of odd length at most k.

We will use the following graphs as patterns:

Fig. 13 The four overlap types: **a** lower body overlap, **b** leg overlap, **c** arm overlap, and **d** partial leg overlap



Definition 25 Let $k + 1$ be an odd integer. B_k denotes the graph in $\mathcal{G}_\lambda^{\leftrightarrow}$ defined by:

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

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

$L_k \in \mathcal{G}_\lambda^{\leftrightarrow}$ 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 Fig. 12.

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. Let $g \subseteq G$ be a subgraph of G. Let $super(g)$ denote the set of all images of P in G containing $g \subseteq G$. We will use four types of overlap (see Fig. 13):

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 $G_P^{\gamma,\mathfrak{K}}$ and a single vertex in $G_p^{\gamma,\mathfrak{K}}$,
leg overlap $P[1]$ and $P[2]$ share a leg, resulting in two adjacent vertices in $G_P^{\gamma,\mathfrak{K}}$ and two adjacent vertices in $G_p^{\gamma,\mathfrak{K}}$,
arm overlap $P[1]$ and $P[2]$ share an arm, resulting in two adjacent vertices in $G_P^{\gamma,\mathfrak{K}}$ and two independent vertices in $G_p^{\gamma,\mathfrak{K}}$,
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 $G_P^{\gamma,\mathfrak{K}}$ and two adjacent vertices in $G_p^{\gamma,\mathfrak{K}}$.

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. More formally:

Definition 26 (*overlap condition*) Let G be a graph composed of n overlapping instances $P[i]$ of $P = B_k$ and/or m overlapping instances $p[j]$ of $p = L_k$, such that no $p[j]$ is a subgraph of a $P[i]$, $1 \leq i \leq n$, $1 \leq j \leq m$, $k > 1$. We say that G obeys the *overlap condition* if any $P[i_1]$ and $P[i_2]$

- do not overlap, or
- overlap in exactly one complete instance of p, which might be shared by other instances $P[i_3]$, or
- overlap in exactly one leg, not shared by any other $P[i_3]$ or $p[j_1]$, or
- overlap in exactly one arm, not shared by any other instance $P[i_3]$,

and any $P[i_1]$ and any $p[j_1]$

- do not overlap, or
- overlap in exactly one leg, not shared by any other $P[i_2]$ or $p[j_2]$,

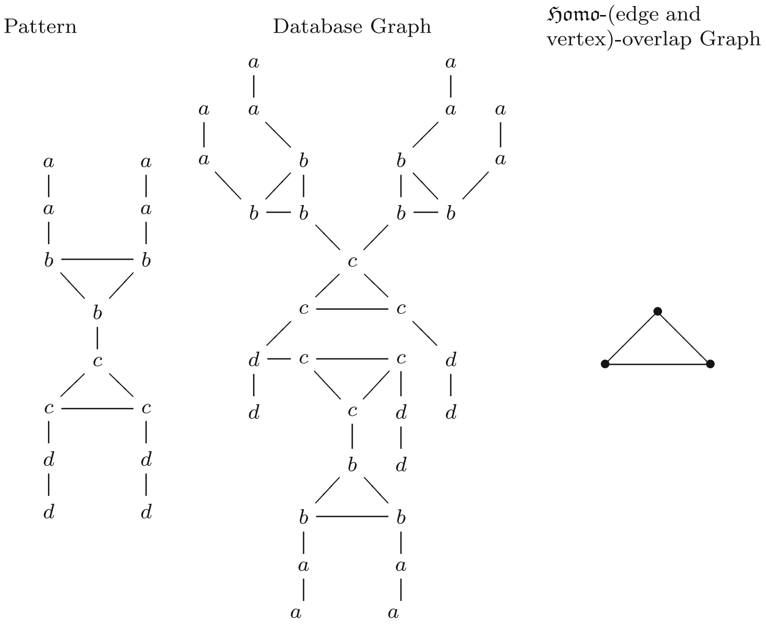
and any $p[j_1]$ and $p[j_2]$

do not overlap, or
overlap in exactly one leg, not shared by any other $p[j_3]$,

with $i_1, i_2, i_3 \in \{1, \ldots, n\}$ all distinct, and $j_1, j_2, j_3 \in \{1, \ldots, m\}$ all distinct.

Figure 14 shows a graph built up from instances of B_3 where all pairs of instances either do not overlap, or follow one of the four overlap types described in Definition 26. For such graphs the following lemma states that the only images of B_3 are exactly those that were used to build up the graph.

Lemma 27 *Let $G \in \mathcal{G}_\lambda^{\leftrightarrow}$ be a graph constructed from n overlapping instances $P[1], \ldots, P[n]$ of $P = B_k$ such that the overlap condition is obeyed. Then, the $P[1], \ldots, P[n]$ are the only $\mathfrak{Homo}$-images of P in G.*



Pattern

Database Graph

$\mathfrak{Homo}$-(edge and vertex)-overlap Graph

Fig. 14 Example of a database graph constructed of three images of B_3 with a lower body overlap and a leg overlap. The pattern graph B_3 has hence three overlapping $\mathfrak{Homo}$-images, resulting in a *triangle* as overlap graph

Proof First, note that the only cycles induced by vertices labeled b (c) are the hips (torsos) of the $P[i]$, $1 \leq i \leq n$. Moreover, because the torso is never shared, there are exactly n instances of the torso, which is a cycle of odd length. Consequently, by Proposition 24, they are exactly $n\mathfrak{Homo}$-images of the torso and they are necessarily $\mathfrak{Iso}$-images. Hence, any homomorphism π from P to G must map the torso of P to such an instance of the torso $T \subset G$, which is part of an instance of $P[i]$, for some $1 \leq i \leq n$.

Due to the nature of the overlap types, k vertices of T are connected to exactly one arm and the only vertex not connected to an arm is connected to a unique vertex v labeled c. By construction each of those arms and v are all part of $P[i]$. Because $d_P(a_i, b_i) = 1, \pi$ must map the arms injectively to the arms of $P[i]$. Using $d_P(c_i, d_i) = 1, 1 \leq i \leq k$, we can repeat the same argument for the legs connected to the unique instance of the hip to which v belongs. Hence, $\pi(P) = P[i]$ and π is necessarily a subgraph isomorphism.

Thus, there are exactly $n\mathfrak{Homo}$-images of P, one for each instance of the torso, and they are necessarily $\mathfrak{Iso}$-images. □

Theorem 28 *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* $H \cong G_P^{\gamma,\mathfrak{K}}$.

Proof Let $H \in \mathcal{G}_\bullet^\leftrightarrow$ be an arbitrary graph and let $V(H) = \{v_1, \ldots, v_n\}$.

Let $k = 2\lceil n/2 \rceil + 1$. As pattern P we take B_k and as database graph G the graph composed of n instances of $P[1], \dots, P[n]$ of P where $P[i]$ and $P[j]$ have an arm or leg overlap if and only if $\{v_i, v_j\} \in E(H)$, $1 \le i < j \le n$. The choice between arm or leg overlap is not important as long as the overlap condition is obeyed, which is always possible by our choice of k. Due to the nature of the overlaps, two images have vertex-overlap if and only if they have edge-overlap.

By Lemma 27 the only $\mathfrak{Homo}$-images of P in G are the $P[i]$, $1 \le i \le n$, and they are all $\mathfrak{Iso}$-images.

As such, if u_i is the vertex of $G_P^{\gamma,\mathfrak{K}}$ corresponding to $P[i]$, we see that the mapping $\varphi(u_i) = v_i$ is an isomorphism from $G_P^{\gamma,\mathfrak{K}}$ to H, $1 \le i \le n$. □

We are now ready to prove the main theorem that extends the result of Theorem 10 to homomorphic images. The constructions used in the proof are illustrated in Fig. 15.

Theorem 29 *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 not decreasing under clique contraction, edge removal and vertex addition.*

Proof Let $H \in \mathcal{G}_\bullet^\leftrightarrow$ be an arbitrary graph, with $V(H) = \{v_1, \dots, v_n\}$.

As pattern P we take B_k and as subpattern $p = L_k$, with $k = 2\lceil n/2 \rceil + 1$.

For each operation $o \in \{\mathsf{CC}, \mathsf{ER}, \mathsf{VA}\}$, we construct a database graph G, such that $G_P^{\gamma,\mathfrak{K}} \cong H$ and $G_p^{\gamma,\mathfrak{K}} \cong o(G_P^{\gamma,mk})$. The database graph G is always composed of n instances $P[1], \dots, P[n]$ of P, which overlap depending on the operation considered. We will tackle vertex- and edge-overlap at the same time, since in each overlaptype there is vertex-overlap if and only if there is edge-overlap. By Lemma 27 the only $\mathfrak{Homo}$-images of P in G are the $P[i]$, $1 \le i \le n$, and they are all $\mathfrak{Iso}$-images. As soon as the database graph G is constructed, we will implicitely assume an isomorphism φ from H to $G_P^{\gamma,\mathfrak{K}}$ such that $\varphi(v_i)$ corresponds with $P[i] \subseteq G$, as in the proof of Theorem 28, and use v_i as short hand for $\varphi(v_i)$, $1 \le i \le n$.

Clique contraction Let K be a clique in H, with $k = |V(K)|$. Without loss of generality we assume that $V(K) = \{v_1, \dots, v_k\}$. The overlap of the $P[i]$ in G is as follows. $P[1], \dots, P[k]$ have a lower body overlap. For any edge $\{v_i, v_j\} \in E(H)$, $k < i < j \le n$, we let $P[i]$ and $P[j]$ have a leg overlap. For any edge $\{v_i, v_j\} \in E(H)$, $1 \le i \le k < j \le n$, we let $P[i]$ and $P[j]$ have a leg overlap if v_j is adjacent to all vertices of K and an arm overlap otherwise. Note that edges corresponding with a leg overlap will remain in $G_p^{\gamma,\mathfrak{K}}$, while edges corresponding with an arm overlap will disappear, since the arms are not present in p. As such, $G_p^{\gamma,\mathfrak{K}} \cong \mathsf{CC}(G, K)$.

Edge removal Let e be an edge in H. Without loss of generality we can assume that $e = \{v_1, v_2\}$. For each edge $\{v_i, v_j\} \neq e \in E(H)$, we let $P[i]$ and $P[j]$ have a leg overlap and $P[1]$ and $P[2]$ have an arm overlap. As such, any edge in $G_P^{\gamma,\mathfrak{K}}$ except e will be present in $G_p^{\gamma,\mathfrak{K}}$ and thus $G_p^{\gamma,\mathfrak{K}} \cong \mathsf{ER}(G, e)$.

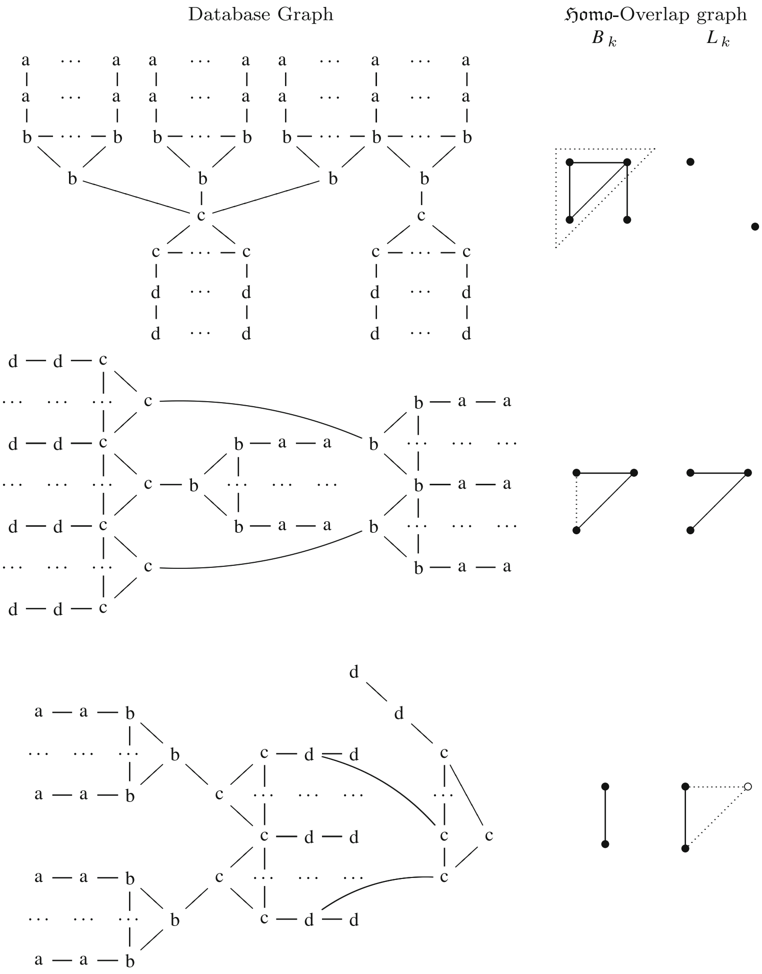


Fig. 15 Illustration of the constructions in Theorem 29. The database graph and the overlap graphs of respectively the pattern B_k and its subpattern L_k are shown for subsequently a clique contraction, an edge removal, and a vertex additions. The *dotted* edges and the *open* nodes are those that are removed/added by the operations on the overlap graphs

Vertex addition Let $P[i]$ and $P[j]$ have a leg overlap in G if $\{v_i, v_j\} \in E(H)$ and add an instance p' of p having a partial leg overlap with all $P[i]$, $1 \leq i \leq n$. Thus, since p' is not part of an instance of P it will have no corresponding vertex in $G_P^{\gamma,\mathfrak{K}}$. In $G_p^{\gamma,\mathfrak{K}}$ however, the vertex associated with p' will be adjacent to all other vertices, since it has a (partial) leg overlap with all instances of p contained in the $P[i]$. In other words, $G_p^{\gamma,\mathfrak{K}} \cong \mathsf{VA}(G)$, as desired.

Note again that we choose k sufficiently large such that in each case we can construct G in such a way that the overlap condition holds. □

5.3 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 Proposition 23.

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

$$\Delta_G(v) := \begin{cases} \#\{w \mid \{v, w\} \in E(G)\} & , \text{ if } \alpha = \leftrightarrow \\ \#\{w \mid (v, w) \in E(G) \vee (w, v) \in E(G)\} & , \text{ if } \alpha = \rightarrow \end{cases}$$

The *maximum degree* of G is then

$$\Delta(G) := max_{v \in V(G)} \Delta_G(v).$$

The *(vertex-)connectivity* $\kappa(G)$ of $G \in \mathcal{G}^\leftrightarrow$ is the minimum number of vertices that need to be removed to make G disconnected. The *local connectivity* $\kappa_G(v, w)$ between two vertices v and w of G equals the minimal number of vertices that need to be removed to disconnect v and w. It is well known that $\kappa_G(v, w)$ equals the number of pairwise disjoint paths between v and w in G (Menger's theorem, Diestel 2000):

$$\kappa_G(v, w) = \max\{|P| \ : \ P \subseteq P_G(v, w) \text{ s.t. paths in } P \text{ are pairwise disjoint}\}.$$

The following theorem is straightforward (see a.o. Bolobas 2001):

Lemma 30 *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. $\forall v, w \in V(H) : \kappa_H(v, w) \leq \kappa_G(\pi(v), \pi(w))$.

Proof $|V(H)| \leq |V(G)|$ follows directly from the fact that $\pi|_{V(H)} : V_H \rightarrow V_G$ is injective. Because $\forall e \neq e' \in E(H), \pi(e)$ and $\pi(e')$ are disjoint and contain each at least one edge, $|E(H)| \leq \sum_{e \in E(H)} |E(\pi(e))| \leq |E(G)|$.

Let v be a node in $V(H)$ with neighbors $v_1, \ldots, v_d$. By definition of homeomorphism, the edges (v, v_1), …, (v, v_d) are mapped to edge disjoint paths $p_1, \ldots, p_d$ of G, all with the same starting node $\pi(v)$. Consider now the second nodes $v'_1, \ldots, v'_d$ on each of these paths. These nodes are all different, as the paths p_i are mutually disjoint. Hence, $\pi(v)$ has a degree of at least d, because it is adjacent to $v'_1, \ldots, v'_d$.

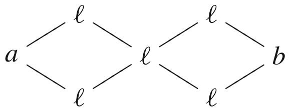
Let v, w be two nodes in $V(H)$ with $\kappa_H(v, w) = k$. Let $(v_1, \ldots, v_k) \oplus (v_k, \ldots v_n)$ denote $(v_1, \ldots, v_n)$; i.e., $\oplus$ concatenates two paths for which the endpoint of the first path is the starting point of the second path, without duplicating the common node. Let now $p = (v, v_1, \ldots, v_l, w)$ be a path between v and w. Then,

$$\pi(p) \;:=\; \pi((v, v_1)) \oplus \pi((v_1, v_2)) \oplus \ldots \oplus \pi((v_l, w))$$

is a path between $\pi(v)$ and $\pi(w)$ in G. First of all, by definition of a homeomorphism, π maps (v_i, v_{i+1}) to a path between $\pi(v_i)$ and $\pi(v_{i+1})$, and thus the precondition for applying $\oplus$ is satisfied. Secondly, no nodes are repeated in p. Similarly, if $p_1, \ldots, p_k$ are pairwise vertex-disjoint paths between v and w, so are $\pi(p_1), \ldots, \pi(p_k)$. Therefore, $\kappa_G(\pi(v), \pi(w)) \geq k$. □

5.3.1 Labeled isomorphisms to labeled homeomorphisms

We first present the labeled case, i.e. a $(\mathfrak{Iso}, \gamma, \leftrightarrow, \lambda)$ to $(\mathfrak{Homeo}, \gamma, \leftrightarrow, \lambda)$ 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. The construction that replaces an edge between a node with label a and a node with label b is as follows:



The two pairs of disjoint paths guarantee that edges are mapped to edges and not to longer paths. If we would simply replace an a-b edge with an a-ℓ-b path, it would be possible to map an a-ℓ-b path to, for instance, an a-ℓ-c-ℓ-b path.

Definition 31 Formally, let $G \in \mathcal{G}_\lambda^{\leftrightarrow}$ with label alphabet Σ. Let $R_1 : G \rightarrow R_1(G) \in \mathcal{G}_\lambda^{\leftrightarrow}$ and $e = \{u, w\} \in E(G)$. We define (V_e, E_e) as follows:

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

Let $e' = \{u', w'\} \in E(G)$. For $e' \neq e$, we make sure that:

$$\begin{aligned} V_e \cap V_{e'} &= \{u, w\} \cap \{u', w'\}, \\ E_e \cap E_{e'} &= \emptyset. \end{aligned}$$

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

$$V(R_1(G)) = V(G) \cup \bigcup_{e \in E(G)} V_e,$$

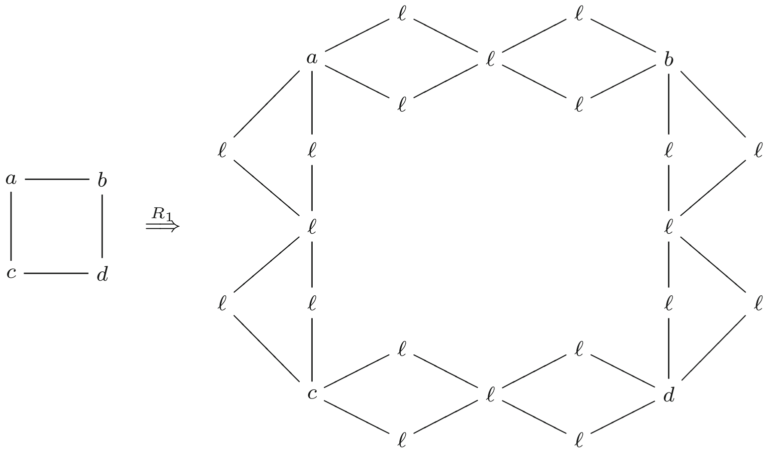


Fig. 16 Illustration of the function R_1 defined in Definition 31

$$E(R_1(G)) = \bigcup_{e \in E(G)} E_e,$$

$$\lambda_{R_1(G)}(u) = \begin{cases} \lambda_G(u), & \forall u \in V(G), \\ \ell \notin \Sigma, & \forall v_e \in V(R_1(G)) \setminus V(G) \end{cases}$$

In Fig. 16 the function R_1 is illustrated.

We will prove the following lemma and theorem only for $\alpha = \leftrightarrow$. The reader can easily check that both proofs are completely analogous for $\alpha = \rightarrow$.

Lemma 32 *Let $G, H \in \mathcal{G}_\lambda^{\leftrightarrow}$. Any surjective homeomorphism π from $R_1(G)$ to $h' \subseteq R_1(H)$ is an isomorphism and the restriction of π to $V(G)$ defines an isomorphism from G to $h \subseteq H$ with $h' = R_1(h)$.*

Proof First note that π must map vertices from G to vertices from H since all vertices in $V(R_1(H)) \setminus V(H)$ are labeled $\ell \notin \Sigma$. Note also that if v is an isolated vertex in G, it is also an isolated vertex in $R_1(G)$. Hence, by the surjectivity of π, $\pi(v)$ must also be isolated in h'.

Let $e = \{u, w\} \in E(G)$ and consider the image of v_e^3 under π. By Lemma 30, $\pi(v_e^3)$ must be a vertex labeled ℓ of degree at least 4. Hence, by construction, $\pi(v_e^3) = v_{e'}^3$ for some $e' \in E(H)$ as these are the only vertices in $R_1(H)$ having this property. Moreover, Lemma 30 also dictates that

$$\kappa_{R_1(H)}(\pi(u), \pi(v_e^3)) \geq \kappa_{R_1(G)}(u, v_e^3) = 2$$
$$\kappa_{R_1(H)}(\pi(v_e^3), \pi(w)) \geq \kappa_{R_1(G)}(v_e^3, w) = 2.$$

The only vertices reachable by two disjoint paths from $v_{e'}^3$ are the end vertices of e'. Hence, the restriction of π to V_e is an isomorphism from (V_e, E_e) to $(V_{e'}, E_{e'})$.

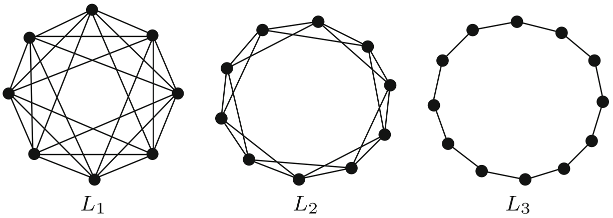


Fig. 17 Label graphs for $n = 3$

Let $E' = \{e' \mid \exists e : \pi(V_e) = V_{e'}\}$ and let W be the set of isolated vertices of G. Combining all the above, we see that

$$V(h') = \pi(W) \cup \bigcup_{e' \in E'} V_{e'} \text{ and } E(h') = \bigcup_{e' \in E'} E_{e'}.$$

Thus, as $E(R_1(G)) = \cup_{e \in E(G)} E_e$ and isolated vertices are mapped to isolated vertices, π itself must be an isomorphism from $R_1(G)$ to $h' \subseteq R_1(H)$. Consequently,

$$\{u, w\} \in E(G) \iff \{\pi(u), \pi(w)\} \in E(h)$$

and thus, by construction, $h' = R_1(h)$ with h the subgraph of H with vertex set $\pi(V(G))$ and edgeset E'. □

Theorem 33 *For all* $\gamma \in \{$*vertex, edge*$\}$, $(P, p, G) \rightarrow (R_1(P), R_1(p), R_1(G))$ *is a* $(\mathfrak{Iso}, \gamma, \leftrightarrow, \lambda)$ *to* $(\mathfrak{Homeo}, \gamma, \leftrightarrow, \lambda)$ *reduction.*

Proof Because R_1 preserves both vertex- and edge-overlap, the theorem follows immediately from the previous lemma as it establishes a bijection between $\mathfrak{Iso}$-images of P (p) in G and $\mathfrak{Homeo}$-images of $R_1(P)$ $(R_1(p))$ in $R_1(G)$. □

5.3.2 Labeled to unlabeled homeomorphisms

We now show the reduction from the labeled case to the unlabeled one; i.e., from vertex-overlap of $\mathcal{G}_\lambda^{\leftrightarrow}$ to γ-overlap of $\mathcal{G}_\bullet^{\leftrightarrow}$ for all $\gamma \in \{$vertex, edge$\}$.

We will use the following special label graphs L_i^n, depicted in Fig. 17, to replace the labels $\Sigma = \{l_1, \ldots, l_n\}$. The L_i^n are based on circulant graphs—a class of graphs for which Muzychuk proved that the isomorphism problem can be solved in polynomial time (Muzychuk 2004).

Definition 34 Let $1 \leq s < k$ be integers. $Ci_k(1, \ldots, s) \in \mathcal{G}_\bullet^{\leftrightarrow}$ denotes the circulant graph with nodes $\{c_0, \ldots, c_{k-1}\}$ and edges

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

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

Hence, a circulant graph $Ci_k(1, \ldots, s)$ is a cycle of length k, with additional edges: every node is adjacent 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 Fig. 17, 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 35 *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 $i = j$.*

Proof Suppose there exists a vertex-disjoint subgraph homeomorphism from L_i^n to L_j^n. The number of vertices of $L_i^n = Ci_{2(n+i)}(1, \ldots, n-i+1)$ equals $2(n+i)$, and the number of edges is equal to $2(n+i) \cdot (n-i+1) = 2n^2 + 2n - 2(i^2 - i)$, because every vertex has degree $2(n-i+1)$. Similarly, for L_j^n, the number of vertices is $2(n+j)$, and the number of edges $2(n^2 + n - (j^2 - j))$. Because of Lemma 30, both the number of vertices and the number of edges of L_i^n must be dominated by, respectively, the number of vertices and the number of edges of L_j^n. This gives:

$$2(n+i) \leq 2(n+j)$$
$$2n^2 + 2n - 2(i^2 - i) \leq 2n^2 + 2n - 2(j^2 - j)$$

and thus,

$$i \leq j$$
$$i^2 - i \geq j^2 - j$$

which is only possible if $i = j$, since $i, j \geq 1$, and $x^2 - x$ is strictly increasing for $x > 0.5$. □

5.3.3 The construction

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 $\mathcal{G}_\lambda^{\leftrightarrow}$ 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 $L^n_{\iota(\lambda(v_i))}$: $V(L[v_i]) = \{c_j^i \mid c_j \in V(L^n_{\iota(v_i)})\}$, and $E(L[v_i]) = \{\{c_j^i, c_k^i\} \mid \{c_j, c_k\} \in E(L^n_{\iota(v_i)})\}$. As such, any two $L[v]$ and $L[w]$ are disjoint whenever $v \neq w$, even if $\lambda(v) = \lambda(w)$.

We are now ready to define the reduction, parameterized by c.

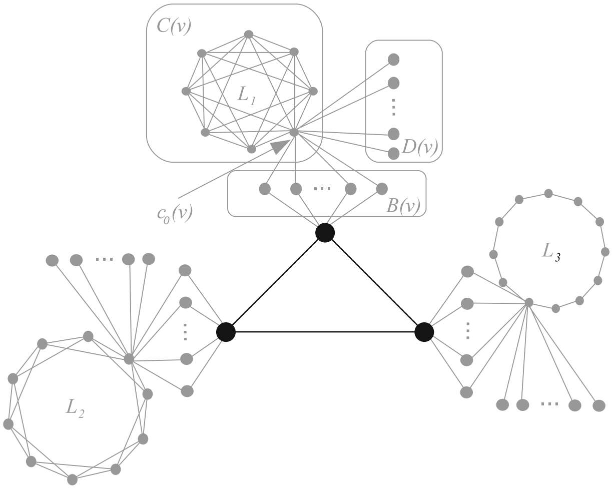


Fig. 18 Reduction for removing labels in the case of homeomorphisms. The triangle in the middle is the original graph G. The labels of the top, $left$, $right$ nodes were respectively l_1, l_2, and l_3

Definition 36 For every $v_i \in V(G)$, let the following sets of vertices be given.

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

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 $\mathcal{G}^{\leftrightarrow}_{\bullet}$:

- The set of vertices is:

$$V(R^c_2(G)) \ := \ V(G) \ \cup \ \bigcup_{v \in V} (C(v) \cup D(v) \cup B(v))$$

- The set of edges $E(R^c_2(G))$ is:

$$\begin{array}{l} E(G) \ \cup \ \bigcup_{v \in V} E(L[v]) \cup \ \bigcup_{v \in V} \{\{c_0(v), d\} \mid d \in D(v)\} \\ \qquad \cup \ \bigcup_{v \in V} \{\{v, b\}, \{b, c_0(v)\} \mid b \in B(v)\} \end{array}$$

An example of the reduction has been given in Fig. 18. Intuitively, the rationale behind the reduction is as follows: the subgraphs $L[v]$ 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 local connectivity arguments to show that in an image always label graphs will be associated with the right node.

Lemma 37 *For all integers* $1 \leq s, k > 2s, \kappa(Ci_k(1, \ldots, s)) = 2s$. *Hence, for all* $L_i^n, \kappa(L_i^n) = 2n - 2i + 2 \leq 2n$.

Proof It is easy to see that $2s$ is an upper bound, as every node has degree $2s$. For the lower bound, we show that there are $2s$ disjoint paths between node c_0 and node c_{s+1}. There are s paths using only nodes $c_0, c_1, \ldots, c_s$: (c_0, c_i, c_s), for $i = 1 \ldots s$. Then there are also s paths using the nodes $c_0, c_s, c_{s+1}, \ldots, c_k$: for every $m = 0, \ldots, s-1$, there is one path using only the nodes $\{c_i \mid s < i \leq k, i \bmod s = m\}$ as intermediate nodes; i.e., for $k = qs + r, r < s$, the paths:

$$\begin{aligned}
&(c_0, c_{qs}, c_{(q-1)s}, \ldots, c_s)\\
&(c_0, c_{(q-1)s+1}, c_{(q-2)s+1}, \ldots, c_{s+1}, c_s)\\
&\ldots\\
&(c_0, c_{(q-1)s+r-1}, c_{(q-2)s+r-1}, \ldots, c_{s+r-1}, c_s)
\end{aligned}$$

The second claim follows directly form the definition of L_i^n as $Ci_{2(n+i)}(1, \ldots, n - i + 1)$. □

The next lemma states some basic properties of the construction. The proof is straightforward and only provided for completeness.

Lemma 38 *Let* G *be a graph in* $\mathcal{G}_\lambda^{\leftrightarrow}$, *with* $V(G) = \{v_1, \ldots, v_k\}$. *Let* $c > \max(\Delta(G), 2n)$. *Then,*

1. $\{x \in V(R_2^c(G)) \mid \Delta_{R_2^c(G)}(x) \geq 2c + 2\} = c_0(V(G))$;
2. $\forall v, w \in V(G) : P_G(v, w) = P_{R_2^c(G)}(v, w)$;
3. $\kappa(R_2^c(G)) = c$, *and*, $\forall v \in V(G) : \{x \in V(R_2^c(G)) \mid \kappa_{R_2^c(G)}(c_0(v), x) \geq c\} = \{v\}$;

Proof For 1 it suffices to notice the degrees of the different node types. Let $v \in V(G)$.

- v has degree $\Delta_G(v_i) + c$ in $R_2^c(G)$: it is still adjacent to the $\Delta_G(v)$ nodes it was adjacent to in G, plus the c nodes in $B(v)$.
- Every node c in $C(v) \setminus \{c_0(v)\}$ is only adjacent to nodes in $C(v)$ and hence has degree $2(n - \iota(v) + 1)$, which is at most $2n$.
- Every node $c_0(v)$ has degree at least $2 + 2c$; first of all, it is adjacent to at least 2 nodes in $L[v]$. Secondly, it is adjacent to the $2c$ nodes in $B(v) \cup D(v)$.
- The nodes $b \in B(v)$ all have degree 2; b is adjacent only to $c_0(v)$ and v.
- The nodes $d \in D(v)$ all have degree 1; d is adjacent only to $c_0(v)$.

For 2, notice that any path between node $v \in V(G)$, and a node in $x \in C(v') \cup D(v') \cup B(v')$, with $v' \in V(G)$, has to contain v'. Therefore, for any path between v and w in $V(G)$, it is not possible that it contains any of the nodes $x \in C(v') \cup D(v') \cup B(v')$ for any $v' \in V(G)$, as this would imply that node v' would be on the path twice; v' must be once on the subpath from v to x, and once again on the second part of the path from x to w. By definition of a path, this is impossible. As such, any path between v and w uses only nodes of $V(G)$, and hence $P_{R_2^c(G)}(v, w) \subseteq P_G(v, w)$. For the other direction of the inclusion, it suffices to notice that $R_2^c(G)$ contains G as a subgraph.

For 3, we give the local connectivity for all pairs of nodes $x, y \in V(R_2^c(G))$. From this, these claims follow directly. Let v, w be two distinct and arbitrary nodes in $V(G)$.

- $(x, y) = (v, c_0(v))$: $c_0(v)$ is connected to v by the $c > \Delta(G) \geq \kappa(G)$ disjoint paths $(v, b, c_0(v)), b \in B(v)$. There are no other paths, as v and $c_0(v)$ become disconnected if all $b \in B(v)$ are removed from $R_2^c(G)$. Hence, the local connectivity is c.
- $(x, y) = (v, c)$, with $c \in C(v) \setminus \{c_0(v)\}$: local connectivity is 1, as v is connected to c, but all paths have to pass through $c_0(v)$.
- $(x, y) = (c_1, c_2)$, with $c_1 \neq c_2 \in C(v)$: the local connectivity is at most $\kappa(L[v]) = 2n - 2\iota(v) + 2$ (Lemma 37), which is at most $2n$, and at least 2, as $L[v]$ always contains a cycle.
- $(x, y) = (v, b)$, with $b \in B(v)$: local connectivity is at most 2 because b has degree 2. There exist two disjoint paths: $((v, b))$ and $((v, b', c_0(v), b))$ for any $b' \in B(v) \setminus \{b\}$. Hence the local connectivity is 2.
- $(x, y) = (b, c_0(v)), b \in B(v)$: local connectivity is 2; the degree of b is 2, and $(b, v, b', c_0(v))$ with $b' \in B(v) \setminus \{b\}$ and $(b, c_0(v))$ are disjoint.
- $(x, y) = (b, c), b \in B(v), c \in C(v) \setminus \{c_0(v)\}$: the local connectivity is 1, as b is connected to c, but all paths between them have to pass through $c_0(v)$.
- (x, d) with $x \in \{v\} \cup C(v) \cup B(v) \cup D(v)$, and $d \in D(v)$: local connectivity is 1; the degree of d is 1, and d is connected to all presented choices for x.
- $(x, y) = (v, w)$: the local connectivity between v and w is at most $\kappa(G) \leq \Delta(G) < c$, since because of point 2, the paths between v and w in $R_2^c(G)$ are exactly the same as the paths between them in G.
- (x, y), with $x \in \{v\} \cup B(v) \cup C(v) \cup D(v)$, $y \in B(w) \cup C(w) \cup D(w)$: the local connectivity is 0 or 1; every path between x and y has to pass through w, so the local connectivity is at most 1. It is 1 if and only if v and w are adjacent in G.

The following lemma will be crucial in the proof of correctness of the reduction:

Lemma 39 *Let $P, G \in \mathcal{G}_\lambda^{\leftrightarrow}$, and let π be a homeomorphism from the unlabeled graph $R_2^c(P)$ to the unlabeled graph $R_2^c(G)$, with $c > \max\{\Delta(G), \Delta(P), 2n\}$. Then, $\pi|_{V(P)\cup E(P)}$ is a label-preserving homeomorphism from the labeled graph P to the labeled graph G. Furthermore, let $v \in V(P)$, $e \in E(P)$, and $e' \in E(R_2^c(P)) \setminus E(P)$, with $e' = \{x, y\}$. It holds that:*

$$\begin{array}{ll} \pi(v) \in V(G) & \\ \pi(e) \in P_G & \pi(C(v)) = C(\pi(v)) \\ \pi(c_0(v)) = c_0(\pi(v)) & \pi(D(v)) = D(\pi(v)) \\ \pi(B(v)) = B(\pi(v)) & \pi(e') = (\pi(x), \pi(y)) \end{array}$$

Proof From Lemma 38 (point 1), we know that for all $v \in V(P)$, $c_0(v)$ has a degree of at least $2c + 2$ in $R_2^c(P)$. According to Lemma 30, these nodes must hence be mapped to nodes in $R_2^c(G)$ that also have a degree of at least $2c + 2$. Because of Lemma 38 (point 1), only the nodes in $c_0(V(G))$ satisfy this condition, and hence, $\pi(c_0(V(P))) \subseteq c_0(V(G))$. According to Lemma 38 (point 3), the connection strength in $R_2^c(P)$ between v and $c_0(v)$ is c, for all $v \in V(P)$. Henceforth, because of Lemma 30, the local connectivity between $\pi(v)$ and $\pi(c_0(v))$ must be at least c as well. Furthermore, we already established that $\pi(c_0(v))$ must be equal to $c_0(w)$ for a w in $V(G)$. Also because of Lemma 38 (point 3), the only node that satisfies this local connectivity condition with $c_0(w)$ in $R_2^c(G)$ is w. Hence, for all $v \in V(P), \pi(v) = w \in V(G)$, and $\pi(c_0(v)) = c_0(w) = c_0(\pi(v))$.

As such, any path in P is mapped to a path in $R_2^c(G)$ between two nodes of $V(G)$. Since Lemma 38 (point 2) states that every path in $R_2^c(G)$ between nodes in $V(G)$ is also a path in $G, \pi|_{V(P)\cup E(P)}(E(P)) \subseteq P_G$. All paths in $\pi|_{V(P)\cup E(P)}(E(P))$ are disjoint, as π is a homeomorphism. So, $\pi|_{V(P)\cup E(P)}(E(P))$ is a homeomorphism between P and G.

We still have to show that $\pi|_{V(P)\cup E(P)}$ is label-preserving. Let $v \in V(P)$. In $R_2^c(P)$, the set of nodes that are adjacent to both $c_0(v)$ and v is the set $B(v)$. These nodes must be mapped to nodes that are connected to both $c_0(\pi(v))$ and $\pi(v)$ through disjoint paths. The only nodes in $R_2^c(G)$ that satisfy this condition are the nodes in $B(\pi(v))$. Therefore, $\pi(B(v)) \subseteq B(\pi(v))$. Because π is injective and both sets have the same cardinality $c, \pi(B(v)) = B(\pi(v))$.

In $R_2^c(G)$, all nodes of $C(v) \setminus c_0(v)$ have a local connectivity of at least 2 with $c_0(v)$ (proof of Lemma 38) Hence, also all nodes $\pi(C(v)) \setminus \{c_0(\pi(v))\}$ must have connection strength of at least 2 with $c_0(\pi(v))$, and thus $\pi(C(v) \setminus \{c_0(\pi(v))\}) \subseteq C(\pi(v))$, because these are the only nodes that have a local connectivity of 2 with $c_0(\pi(v))$, except for the nodes in $B(\pi(v))$ and $\pi(v)$, but these nodes are already in the image of the nodes $B(v) \cup \{v\}$, and π is injective. Hence, $\pi|_{V(L[v])\cup E(L[v])}$ is a subgraph homeomorphism from a $L[v]$ to $L[\pi(v)]$). Because of Lemma 35, this implies that $\iota_P(v) = \iota_G(\pi(v))$, and thus $\lambda_P(v) = l_{\iota_P(v)} = l_{\iota_G(\pi(v))} = \lambda_G(\pi(v))$. Henceforth, $\pi|_{V(P)\cup E(P)}$ is also label-preserving.

To show $\pi(D(v)) = D(\pi(v))$, notice that every $d \in D(v)$ is adjacent to $c_0(v)$. Therefore, $\pi((c_0(v), d))$ must be a path from $\pi(c_0(v))$ to $\pi(d)$, disjoint from all nodes $\pi(V(R_2^c(P)))$. The only nodes to which there still is a disjoint path possible, are the nodes in $D(\pi(v))$. Both $D(\pi(v))$ and $\pi(D(v))$ have the same cardinality and thus $D(v) = D(\pi(v))$. □

Also the other direction obtains:

Lemma 40 *Let $P, G \in \mathcal{G}_\lambda^{\leftrightarrow}$, and let $c > \max\{\Delta(G), \Delta(P), 2n\}$. Let π be a subgraph homeomorphism from P to G. Then there exists a (not necessarily unique) subgraph homeomorphism ψ from $R_2^c(P)$ to $R_2^c(G)$ such that $\psi|_{V(P)\cup E(P)} = \pi$. Nevertheless, the image ext(g) in $R_2^c(G)$ through ψ is unique and g equals the subgraph of ext(g) induced by V_G.*

Proof The first part on the existence of ψ is straightforward; from Lemma 39 we know that for all $v \in V(P), \pi(v) \in V(G)$, and $\lambda_P(v) = \lambda_G(\pi(v))$, and thus $L[v]$ and $L[\pi(v)]$ are isomorphic, $B(v)$ can be mapped to $B(\pi(v))$, and $D(v)$ to $D(\pi(v))$. Any edge (v, w) in $E(R_2^c(P)) \setminus E(P)$ is mapped to the path with as only edge $(\pi(v), \pi(w))$. The resulting function is a homeomorphism from $R_2^c(P)$ to $R_2^c(G)$.

The second part on the uniqueness of the image now follows from Lemma 39: the mapping of the vertices and the edges $V(P) \cup E(P)$ completely determines the mapping of $R_2^c(P)$ in $R_2^c(G)$; the image of $C(v)$ is $C(\pi(v))$, the image of $B(v)$ is

$B(\pi(v))$, and the image of $D(v)$ is $D(\pi(v))$. Also the edges in $\pi(E(R_2^c(P)))$ are uniquely determined by the mapping of the vertices and the edges $V(P) \cup E(P)$; let $e = (x, y)$ be an edge in $E(R_2^c(P)) \setminus E(P)$. Then at least one of x, y must be in $B(v) \cup C(v) \cup D(v)$ for a $v \in V(P)$. We assume without loss of generality that $x \in B(v) \cup C(v) \cup D(v)$. Then, $\pi(x) \in B(\pi(v)) \cup C(\pi(v)) \cup D(\pi(v))$. Notice, however, that all vertices adjacent to x are in $\pi(V(R_2^c(P)))$. Therefore, the path $\pi(x, y)$ cannot contain any of the nodes adjacent to $\pi(x)$ as an internal node. This is only possible if $\pi(x, y) = (\pi(x), \pi(y))$. As we already established that $\pi(x)$ and $\pi(y)$ are completely determined by $\pi|_{V(P) \cup E(P)}$, so is $\pi(e)$.

Hence, the image of π is completely determined by $\pi|_{V(P) \cup E(P)}$. □

We will denote the unique extension of a $\mathfrak{Homeo}$-image g of P in G to an $\mathfrak{Homeo}$-image of $R_2^c(P)$ in $R_2^c(G)$ by *ext*(g).

Theorem 41 *Let $P, G \in \mathcal{G}_\lambda^{\leftrightarrow}$, let $c > \max\{\Delta(G), \Delta(P), 2n\}$, and let ext$(g)$ and ext(g') be two $\mathfrak{Homeo}$ images of $R_2^c(P)$ in $R_2^c(G)$.*

The mapping from $\mathfrak{Homeo}$-image of P in G to $\mathfrak{Homeo}$-image of $R_2^c(P)$ in $R_2^c(G)$ that maps g to ext(g) is a bijection.

Let g, g' be two $\mathfrak{Homeo}$-images of P in G. The following are equivalent:

1. *g and g' vertex-overlap.*
2. *ext(g) and ext(g') vertex-overlap;*
3. *ext(g) and ext(g') edge-overlap;*

Proof Suppose *ext*$(g) =$ *ext*(g'). Then, g and g' are equal on $V(P)$. Because of Lemma 40, this implies that $g = g'$.

The second part follows directly from Lemma 40 and the fact that when two homeomorphisms from $R_2^c(P)$ to $R_2^c(G)$ overlap, they must also overlap on G. This follows from Lemma 39; R_2^c is such that in a homeomorphism π must have $\pi(B(v)) = B(\pi(v))$, $\pi(C(v)) = C(\pi(v))$, and $\pi(D(v)) = D(\pi(v))$. Hence, if two images overlap in any node of $B(\pi(v)) \cup C(\pi(v)) \cup D(\pi(v))$, they must also overlap at $\pi(v)$. For the edge overlap it suffices to notice that if two mappings overlap in $\pi(v)$, they also overlap in, e.g., $(\pi(v), \pi(c_0(v)))$. □

From the theorem, the proof of Corollary 42 is immediate.

Corollary 42 *Let $p, P, G \in \mathcal{G}_\lambda^{\leftrightarrow}$, and let $c = \max\{\Delta(G), \Delta(P), \Delta(p), 2n\} + 1$. The function R_2 that maps (p, P, G) to $(R_2^c(p), R_2^c(P), R_2^c(G))$ is a $(\mathfrak{Homeo}, vertex, \alpha, \lambda)$ to $(\mathfrak{Homeo}, \gamma, \alpha, \bullet)$ reduction, for all $\alpha \in \{\rightarrow, \leftrightarrow\}$ and $\gamma \in \{vertex, edge\}$.*

5.4 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 43 *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.*

Proof It is sufficient to show that for any two different vertices v and w of K_k, $\pi(v) \neq \pi(w)$. If $\pi(v) = \pi(w)$ would be true, the loop $\{\pi(v), \pi(w)\}$ would belong to $E(G)$, which would be a contradiction with the assumption that G does not contain loops. Therefore, π is a subgraph isomorphism mapping. □

Apart from the notations introduced in Definition 45, we will also use

$$V_j^w = \{v_{w,j}, v_{w,j+1 \bmod \sigma(w)}, \dots v_{w,j+K \bmod \sigma(w)}\}.$$

The subgraphs attached to the original vertices to represent the labels are isomorphic to the graphs $C_K^{\sigma(w)}$ as in Definition 34 and are illustrated in Fig. 17.

We now formalize the encoding of labels with undirected subgraphs:

Definition 44 Let $G \in \mathcal{G}_\lambda^{\leftrightarrow}$. Let k be a strict upper bound on the size of the largest clique of G. A Schema for Labeling with Unlabeled Subgraphs (SLUS) for G is a pair (K, σ) such that

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

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

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 45 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, $j = 1 \dots \sigma(w)$ are new vertices.
- the edges of $R_3^{K,\sigma}(G)$ are

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

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

$$E^w = \Big\{\{v_{w,j}, v_{w,j+i \bmod \sigma(w)}\} \mid \\ 0 \leq j < \sigma(w) \wedge 1 \leq i \leq K\}$$

The subgraphs attached to the original vertices to represent the labels are isomorphic to the graphs $C_K^{\sigma(w)}$ as in Definition 34 and are illustrated in Fig. 17.

Lemma 46 *Let $G \in \mathcal{G}_\lambda^{\leftrightarrow}$, (K, σ) be a SLUS for G and $w \in V(G)$. Then, the set of $(K+1)$-cliques of $R_3^{K,\sigma}(G)$ is exactly the set of subgraphs induced by V_j^w, $w \in V(G), 0 \leq j < \sigma(w)$.*

Proof This follows directly from the construction of $les(G, K, \sigma)$. □

In the remainder of this section, we will assume that $G, P \in \mathcal{G}_\lambda^{\leftrightarrow}$ and that (K, σ) is a SLUS for G and for P. We will prove that there is a homomorphism π from P to G iff there is a homomorphism π' from $R_3^{K,\sigma}(P)$ to $R_3^{K,\sigma}(G)$, and that in this case $\pi' \supset \pi$.

Lemma 47 *Let π' be a subgraph homomorphism from $R_3^{K,\sigma}(P)$ to $R_3^{K,\sigma}(G)$. Let w be a vertex of P. Then, there is a $w' \in V(G)$ such that $\pi'(V^w) \subseteq V^{w'}$.*

Proof Consider the graph induced by V_j^w in $R_3^{K,\sigma}(P)$ for some $0 \leq j < \sigma(w)$. It is clearly a $(K+1)$-clique from the definition of $E_{K,\sigma}^{lab}(P)$. From Lemma 43, we know that the image under π' should be a $(K+1)$-clique of $R_3^{K,\sigma}(G)$.

Suppose that $\pi'(e) \in E(G)$, and hence that $\pi'(e)$ is part of a $(K+1)$-clique. Then, as no element of $E_{K,\sigma}^{lab}(G)$ is adjacent to two different vertices in $V(G)$, no vertices of $E_{K,\sigma}^{lab}(G)$ can be in this $(K+1)$-clique, and the $(K+1)$-clique must be a subgraph of G itself. This is impossible as the size of the largest clique in G is at most $k < K$. Therefore, we must conclude that $\pi'(e) \in E_{K,\sigma}^{lab}(G)$.

Moreover, as all (V^v, E^v) are disjoint graphs, the endpoints of $\pi'(e)$ are in the same (V^v, E^v). We have $\pi'(w) = \pi'(w_0) = \pi'(w_1)$ and by induction we can see that all $\pi'(w_j)$ are in the same $V^{w'}$ for some w'. □

Theorem 48 *Assume that G and P have no two adjacent vertices with identical labels. Let π' be a subgraph homomorphism mapping from $R_3^{K,\sigma}(P)$ to $R_3^{K,\sigma}(G)$. Let w be a non-isolated vertex of P. Then, for any $0 \leq j < \sigma(w)$, π' maps w, j on a vertex $w'_{j'}$ where $\lambda(w) = \lambda(w')$ and $j = j'$ or $j + j' = \sigma(w)$.*

Proof We will first introduce some notations. Consider a fixed homomorphism π' from P to G. Consider also a fixed $w \in V(P)$. According to Lemma 47 there is a $w' \in V(G)$ such that every w_j is mapped by π' on some vertex $w'_{j'}$ for some j'. Now let f be a function that maps every j on $f(j)$ such that

$$\pi'\left(w_{j \bmod \sigma(w)}\right) = w'_{f(j)}$$

and $0 \leq f(j) < \sigma(w')$.

For any j, as V_j^w induces a $(K+1)$-clique in $R_3^{K,\sigma}(P)$, its image under π' is also a $(K+1)$-clique, and according to Lemma 43, π' restricted to V_j^w is a bijection. Hence, $\pi'(V_j^w) = V_{j'}^{w'}$ for some j'. Let F be the function that maps every j on $F(j)$ such that

$$\pi'\left(V_{j \bmod \sigma(w)}^w\right) = V_{F(j)}^{w'}$$

and $0 \leq F(j) < \sigma(w')$.

We start with making a number of observations. First, from the structure of $les_{K,\sigma}(G)$, we can see that

$$F(j) - F(j+1) \bmod \sigma(w') \in \{\sigma(w') - 1, 0, 1\} \tag{2}$$

Indeed, $w_{j+i \bmod \sigma(w)}$ with $1 \leq i \leq K$ are all distinct, and so are there images under π'. They belong to the intersection of V_j^w and $V_{j+1 \bmod \sigma(w)}^w$. So the intersection of their images should also contain at least K elements.

Second, as the image of V_j^w contains $\pi'(w, j)$, we have

$$0 \leq f(j) - F(j) \bmod \sigma(w) \leq K \tag{3}$$

Third, corresponding to the three possible ways in which $\pi'\left(V_j^w\right)$ and $\pi'\left(V_{j+1 \bmod \sigma(w)}^w\right)$ overlap, we have

$$(f(j) - f(j+K+1)) \bmod \sigma(w') \in \{\sigma(w') - K - 1, 0, K+1\}. \tag{4}$$

It also holds that

$$\begin{aligned} 1 \leq |j_1 - j_2| \leq K \Rightarrow 1 \leq |f(j_1) - f(j_2)| \leq K \\ \vee \sigma(w') - K \leq |f(j_1) - f(j_2)| \leq \sigma(w') - 1 \end{aligned} \tag{5}$$

Let $\sigma(w) = 4(K+1) + r$ with $0 < r < K+1$ integer. Now consider

$$\begin{aligned} \Delta_{fj} &= f(j) - f(j + 4(K+1)) \\ &= f(j) - f(j + (K+1)) \\ &\quad + f(j + (K+1)) - f(j + 2(K+1)) + f(j + 2(K+1)) \\ &\quad - f(j + 3(K+1)) + f(j + 3(K+1)) - f(j + 4(K+1)) \\ &= k_1(\sigma(w') - (K+1)) + k_2(K+1) \\ &= k_1\sigma(w') + (k_2 - k_1)(K+1), \end{aligned}$$

for some $k_1, k_2 \in \mathbb{N}$.

Clearly,

$$|\Delta_{fj}| = n'(K+1) \vee |\Delta_{fj}| = \sigma(w') - n'(K+1) \tag{6}$$

for some integer n' with $0 < n' \leq 4$.

We will now prove that $\lambda(w) = \lambda(w')$. Suppose that $\lambda(w) \neq \lambda(w')$. We will show that this leads to a contradiction. There are a number of possibilities to consider:

- A first case is the situation where there is a $0 \leq j' < \sigma(w')$ such that there is no j such that $f(j) = j'$. Without loss of generality we can assume that $j' = \sigma(w') - 1$ and $|\Delta_{fj}| = n'(K+1)$ for some n'. However, $f(j + n(K+1)) = f(j - r)$ and hence $\Delta_{fj} = f(j) - f(j-r)$. Equation 5 then contradicts $|\Delta_{fj}| = n'(K+1)$.
- Therefore, π' restricted to V^w is a surjection on $V^{w'}$. Hence, $\sigma(w) > \sigma(w')$ (as we assume $\lambda(w) \neq \lambda(w')$).
  So the function f takes all values from 0 to $\sigma(w') - 1$, and from Eq. 3, for each $0 \leq f(j) < \sigma(w')$, $F(j) \leq f(j) \leq F(j) + K$. Moreover, from Eq. 2 we have that for each j,

$$F(j+1) - F(j) \bmod \sigma(w') \in \{\sigma(w') - 1, 0, +1\}.$$

  These three facts together imply that for every $0 \leq p < \sigma(w')$, there is a j_p such that $F(j_p) = p$. If this would not be the case, e.g. if for no value of j, $F(j)$ would be $\sigma(w') - 1$, $F(j)$ would go (with increasing j) from a value $0 \leq F(j_K) \leq K$ when $f(j_K) = K$ up to a value $\sigma(w') - K - 2 \leq F(j_{\sigma(w')-2}) \leq \sigma(w') - 2$ when $f(j_{\sigma(w')-2}) = \sigma(w') - 2$ (which takes at least $(\sigma(w') - K - 2) - K = \sigma(w') - 2K - 2$ steps if we can not pass $\sigma(w') - 1$), and then down again to $0 \leq F(j_K) \leq K$ (which again takes at least $\sigma(w') - 2K - 2$ steps). Going up and down with increasing j would require $2(\sigma(w') - 2K - 2) = 2\sigma(w') - 4K - 4$ steps, but $2\sigma(w') - 4K - 4 > \sigma(w')$ due to the definition of $\sigma(w')$. So there are not enough steps available to go all the way up and down from $F(j_K)$ to $F(j_{\sigma(w')-2})$ and back.
  Now we can see that this means that for some j^*, $F(j^*) = F(j^*+1)$. Suppose this would not be true. Then, for every j, $F(j+1) = F(j) + 1 \bmod \sigma(w')$ or $F(j+1) = F(j) - 1 \bmod \sigma(w')$. If $F(0)$ is even, $F(1)$ is odd, $F(2)$ is even, etc. and since $\sigma(w')$ is odd, $F(\sigma(w'))$ would be even and we could conclude that $F(0)$ is odd. Similarly, if $F(0)$ is odd, we could show by going round once that $F(0)$ is also even. Hence, for some j^*, $F(j^*) = F(j^*+1)$.
  This means that $V^w_{j^*} \setminus V^w_{j^*+1} = \{w_{j^*}\}$ and $V^w_{j^*+1} \setminus V^w_{j^*} = \{w_{j^*+K+1}\}$ have the same image under π', and hence $f(j^*) = f(j^*+K+1)$. For each $1 \leq i \leq K$, we have now by Eq. 5 that

$$\begin{aligned} &f(j^*+i) - f(j^*) \bmod K+1 \in \pm K \\ &f(j^*-K-1+i) - f(j^*) \bmod K+1 \in \pm K \\ &f(j^*+K+1+i) - f(j^*) \bmod K+1 \in \pm K \end{aligned}$$

  where

$$\pm K = \{\sigma(w') - K, \ldots, \sigma(w') - 1, 0, 1, \ldots, K\}$$

  Because all the pairwise differences between $f(j^*+i)$, $f(j^*-K-1+i)$ and $f(j^*+K+1+i)$ (mod $K+1$) are multiples of $K+1$, for each i, two should be equal. Therefore, in total at least $K + 1$ vertices of $V^{w'}$ are the images of two different elements of V^w under π'.

Earlier we concluded that for all $0 \leq p < \sigma(w')$, for some j we have $f(j) = p$. If now for at least $K + 1$ of these values of p we should have two distinct values of j such that $f(j) = p$, we need at least $\sigma(w') + K + 1$ vertices in V^w. But $\sigma(w) < \sigma(w') + K + 1$, which is a contradiction.

Therefore, we must reject our assumption that $\lambda(w) \neq \lambda(w')$.

We have now shown that $\lambda(w) = \lambda(w')$. Now consider the hypothesis that an edge from E_{orig} with endpoints w_1 and w_2 is mapped on an edge of E_{lab} by π'. As no edge from E_{lab} is mapped on an edge of E_{orig}, this would mean that the subgraphs of $les(G, K, \sigma)$ induced by V^{w_1} and by V^{w_2} are mapped on the same subgraph of $les(G, K, \sigma)$ induced by $V^{w'}$ for some w'. As we know that no edge of G has endpoints with identical labels, we cannot have both $\lambda(w_1) = \lambda(w')$ and $\lambda(w_2) = \lambda(w')$. Therefore we should reject the possibility that an edge from E_{orig} is mapped on an edge of E_{lab} by π'.

Therefore, every vertex w_0 is mapped on some vertex w'_0 by π'. The rest of the proof is straightforward. □

Theorem 49 *Assume that P and G have no two adjacent vertices with identical labels and that P and G have no isolated vertices. Then, P is subgraph homomorphic to G if and only if $R_3^{K,\sigma}(P)$ is subgraph homomorphic to $R_3^{K,\sigma}(G)$.*

Proof Straightforward from Theorem 48, noting the relation between a label-preserving homomorphism $\pi : P \to G$ and a homomorphism $\pi' : les_{K,\sigma}(P) \to les_{K,\sigma}(G)$:

$$\begin{aligned} &\exists \pi', \forall w \in V(P), \begin{cases} \pi'(w_0) = w'_0 \\ \pi'(w_i) = w'_i, \quad 1 \leq i \leq \sigma(w) = \sigma(w') \end{cases} \\ &\iff \exists \pi, \forall w \in V(P), \pi(w) = w', \end{aligned}$$

Theorem 50 *Let $p, P, G \in \mathcal{G}_\lambda^{\leftrightarrow}$, 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 an $\mathfrak{Homo}$-vertex-overlap on $\mathcal{G}_\lambda^\alpha$ to $\mathfrak{Homeo}$-γ-overlap on $\mathcal{G}_\bullet^\alpha$ reduction, for all $\alpha \in \{\to, \leftrightarrow\}$ and $\gamma \in \{vertex, edge\}$.*

Proof Armed with Theorem 49, we only have to proof that the restriction to (sub)graphs with no isolated vertices or edges between vertices with the same label can be circumvented. Indeed, consider the disjoint union G' of two labeled graphs P and G without isolated vertices or edges between vertices with the same label. By Theorem 49, there exists a homomorphism π from P to a subgraph g of G iff there exists a homomorphism π' from $R_3^{K,\sigma}(P)$ to $R_3^{K,\sigma}(G)$, with (K, σ) an SLUS for G'.

Now, consider a mapping f that maps a labeled graph $H = (V, E, \lambda)$ with label alphabet Σ and $E(H) = \{e_1 = \{u_1, w_1\}, \ldots, e_m = \{u_m, w_m\}\}$ to a labeled graph $f(H) = (V_f, E_f, \lambda_f)$ with label alphabet $\Sigma \cup \{\alpha, \beta_1, \beta_2, \beta_3\}, \alpha, \beta_1, \beta_2, \beta_3 \notin \Sigma$, defined by:

$$\begin{aligned}
V_f &= V \cup \{v_*, v_1, \ldots, v_{5m}\},\\
E_f &= \{\{v*, w\} | w \in V\} \cup \cup_{i=1}^{m} \{\\
&\qquad \{u_i, v_{5i-4}\}, \{v_{5i-4}, v_{5i-3}\}, \{v_{5i-3}, w_i\},\\
&\qquad \{u_i, v_{5i-2}\}, \{v_{5i-2}, w_i\},\\
&\qquad \{u_i, v_{5i-1}\}, \{v_{5i-1}, v_{5i}\}, \{v_{5i}, w_i\}\}\\
\lambda_f(u) &= \lambda(u), \quad \forall u \in V,\\
\lambda_f(v_*) &= \alpha\\
\lambda_f(v_{5i-4}) = \lambda_f(v_{5i}) &= \beta_1, \quad 1 \leq i \leq m,\\
\lambda_f(v_{5i-3}) = \lambda_f(v_{5i-1}) &= \beta_2, \quad 1 \leq i \leq m,\\
\lambda_f(v_{5i-2}) &= \beta_3, \quad 1 \leq i \leq m,
\end{aligned}$$

with $v_*, v_1, \ldots, v_{5m} \notin V$. Clearly, $f(H)$ has no isolated vertices and no adjacent vertices with identical labels.

We show that P is homomorphic to a subgraph g of G iff $f(P)$ is homomorphic to $f(g)$, which in turn is a subgraph of $f(G)$. Let v_*^P, v_i^P and v_*^g, v_j^g be the new vertices of $f(P)$ and $f(g)$, $1 \leq i \leq 5|E(P)|$, $1 \leq j \leq 5|E(g)|$.

Assume a homomorphism π from P to g, then π_f mapping each $u \in V(P)$ to $\pi(u)$, each v_{5i-k}^P associated with an edge $e_i^P = \{u_i, w_i\}$ of P to v_{5j-k}^g associated with the edge $e_j^g = \{\pi(u_i), \pi(w_i)\}$ of g, $0 \leq k \leq 4$, v_*^P to v_*^g, is a homomorphism from $f(P)$ to $f(g)$.

On the other hand, if π_f is a homomorphism from $f(P)$ to $f(g)$ then $\pi(v_*^P) = v_*^g$ because there is only one vertex in each graph labeled α. Consider $u, w \in V(P)$, if $\{u, w\} = e_i \in E(P)$, for some i, then $u, v_{5i-4}^P, v_{5i-3}^P, w$ is a path in $f(P)$ in which every edge connects vertices with different labels. Consequently, it must be mapped by π_f to a path of equal length with the same label sequence. Due to label restrictions, $\pi_f(u) = u'$, for some $u' \in V(g)$, and $\pi_f(v_{5i-4}^P) \in \{v_{5j-4}^g, v_{5j}^g\}$, for some j such that $e_j^g = \{u', w'\} \in E(g)$. If $\pi_f(v_{5i-4}^P) = v_{5j-4}^g$ it immediately follows that $\pi_f(v_{5i-3}^P) = v_{5j-3}^g$ and $\pi_f(v_{5i-4}^P) = v_{5j}^g$ implies $\pi_f(v_{5i-3}^P) = v_{5j-1}^g$. In either case, $\pi_f(w)$ must be w'. Thus, if $\{u, w\} \in E(P)$ then $\{\pi_f(u), \pi_f(w)\} \in E(g)$. In other words, the restriction of π_f to $V(P)$ defines a homomorphism from $P \rightarrow g$. □

6 Discussion and conclusion

We extended the results of Vanetik et al. (2006) 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 images. We considered images to be all vertices (edges) of the embedding of a pattern in the database graph.

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 $|V(G)|$. Bringmann and Nijssen (2007) 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 also exist different notions of overlap. e.g. Fiedler and Borgelt (2007) 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 also be easily adapted to generalize to the harmful overlap notion to the considered combinations of directedness, labeledness and morphism choice.

This work focusses on the mathematical properties of anti-monotonic support measures on graphs, an important first condition for mining efficiently. It does not completely solve the complexity question, as in the general case pattern matching remains NP-complete and the number of matches may be exponential. There is a large literature on exact and approximate graph matching methods and in practical applications the number of patterns may be tractable. E.g. pattern matching is possible in polynomial time if the morphism type is homomorphism and the patterns have bounded treewidth. E.g. the size of the overlap graph is bounded by a linear function of the database size if plane graphs and plane embedding are considered. These elements offer good perspectives to build on our theory and develop a practical pattern mining system for large graphs in future work.

Another interesting research direction involves random graphs (Bolobas 2001). These graphs satisfy statistical regularities, and sampling only partially the occurrences of a pattern may be sufficient to make a good estimation of some useful support measures (Kashtan et al. 2004; Furer and Kasiviswanathan 2008).

Acknowledgements This work is supported by the FWO project ‘Graph logic: Representation, Inference and Learning’. Jan Ramon is supported by ERC Starting Grant 240186 ‘MiGraNT: Mining Graphs and Networks, a Theory-based approach’ and by K.U. Leuven GOA project ‘Probabilistic Logic Languages’.

References

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

Bolobas B (2001) Random graphs. Cambridge University Press, Cambridge

Brimkov Valentin E (2004) Clique, chromatic, and lovasz numbers of certain circulant graphs. Electron Notes Discr Math 17:63–67

Bringmann B, Nijssen S (2007) What is frequent in a single graph? In: Proceedings of mining and learning with graphs (MLG), Florence, Italy

Crespi V (2004) Exact formulae for the lovasz theta function of sparse circulant graphs. SIAM J Discr Math 17(4):670–674

De Raedt L, Kramer S (2001) The levelwise version space algorithm and its application to molecular fragment finding. In: Nebel B (ed) Proceedings of the 17th international joint conference on artificial intelligence. Morgan Kaufmann, CA, pp 853–862

Diestel Reinhard (2000) Graph theory. Springer, New York

Fiedler M, Borgelt C (2007) Support computation for mining frequent subgraphs in a single graph. In: Proceedings of the fifth workshop on mining and learning with graphs (MLG'07), Florence

Furer M, Kasiviswanathan S Prasad (2008) Approximately Counting Embeddings into Random Graphs. In: Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop, RANDOM 2008 on approximation, randomization and combinatorial optimization: algorithms and techniques, Boston, MA, USA, pp 416–429

Gross JL, Yellen J (2004) Handbook of graph theory. CRC Press, Boston

Grunewald et al (2007) Qnet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. Mole Biol Evol 24(2):532–538

He H, Singh AK (2007) Efficient algorithms for mining significant substructures in graphs with quality guarantees. In: IEEE international conference on data mining, Omaha, Nebraska, pp 163–172

Hell P, Nešetřil J (2004) Graphs and homomorphisms. Oxford University Press, Oxford

Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 22;20(11):1746–1758

Knuth Donald E (1994) The sandwich theorem. Electron J Combin 1:48

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

LaPaugh AS, Rivest RL (1978) The subgraph homeomorphism problem. In: STOC '78. ACM Press, New York, NY, USA, pp 40–50

Mcglohon M, Leskovec J, Faloutsos C, Hurst M, Glance N (2007) Finding patterns in blog shapes and blog evolution. In: Proceedings of the international conference on weblogs and social media, Boulder, CO, USA, pp 26–28

Muggleton S, De Raedt L (1994) Inductive logic programming : theory and methods. J Logic Prog 19, 20:629–679

Muzychuk M (2004) A solution of the isomorphism problem for circulant graphs. Proc Lond Math Soc 3:1–41

Papadimitriou CH (1994) Computational complexity. Addison-Wesley, Boston

Ramon J, Francis T, Blockeel H (2000) Learning a Tsume-Go heuristic with Tilde. In: Proceedings of CG2000, the second international conference on computers and games, Hamamatsu, Japan. Lecture Notes in Computer Science, vol 2063. Springer, NY, pp 151–169

Tong H, Faloutsos C, Gallagher B, Eliassi-Rad T (2007) Fast best-effort pattern matching in large attributed graphs. In: KDD '07: proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, pp 737–746

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