Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 Vol I,
IMECS 2010, March 17 - 19, 2010, Hong Kong

An Expressive Outerplanar Graph Pattern Class and its Efficient Pattern Matching Algorithm

Hitoshi Yamasaki* Takashi Yamada* Takayoshi Shoudai*

Abstract— An outerplanar graph is a planar graph which can be embedded in the plane in such a way that all of vertices lie on the outer boundary. Many chemical compounds are known to be expressed by outerplanar graphs. An externally extensible outerplanar graph pattern (*eeo-graph pattern* for short) represents a graph pattern common to a finite set of outerplanar graphs like a dataset of chemical compounds. The eeo-graph pattern can express a substructure common to blocks which appear in outerplanar graph structured data. In this paper, we propose a polynomial time algorithm of deciding whether or not a given eeo-graph pattern matches a given connected outerplanar graph.

Keywords: pattern discovery, graph structured pattern, outerplanar graph, graph mining, chemical dataset

1 Introduction

Large amount of data having graph structures, such as map data, CAD, biomolecular, chemical molecules, the World Wide Web, are stored in databases. For example, HTML/XML documents can be expressed by ordered trees, and almost chemical compounds in the NCI dataset [5], which is one of the popular graph mining datasets, are known to be expressed by outerplanar graphs. Outerplanar graphs are planar graphs embedded in the plane in such a way that all of vertices lie on the outer boundary. In Fig. 1, we give graphs g_1, g_2, g_3, G, and H as examples of outerplanar graphs. Many researchers are interested in knowledge discovery from data having structures such as sequences, trees, or graphs [1, 2, 6]. Horváth et al. [3] proposed a frequent subgraph mining algorithm for outerplanar graphs. For graph mining algorithms, graph pattern matching algorithms play a key role throughout the computations. In this paper, we present polynomial time matching algorithms for graph patterns having outerplanar graph structures.

For a graph G, we call a maximal biconnected subgraph of G a *block* of G if it has at least three vertices. For example, in Fig. 1, $B_1, \ldots, B_8$ are blocks. A block of an outerplanar graph has a unique planar embedding up to the mirror image. Thus the edges of an outerplanar graph are classified into two types, *external* and *internal*. External edges lie on the external face, while internal edges do not lie on the external face. An external edge which does not belong to any block is called a *bridge*. Because an internal edge must belong to a block, it is also called a *diagonal* of the block.

For an integer $d \geq 0$, an outerplanar graph is said to be d-tenuous if each of its blocks contains at most d diagonals. Horváth et al. [3] proposed an Apriori-like algorithm for enumerating all frequent d-tenuous outerplanar subgraphs in a finite set of outerplanar graphs. Their algorithm works in incremental polynomial time (i.e., in polynomial time with respect to the combined size of the input and the output so far computed). In [8], we introduced a graph-structured pattern, called a *block preserving outerplanar graph pattern* (*bpo-graph pattern* for short), which is an outerplanar graph with structured variables, and proposed a refinement-based technique for enumerating all maximal frequent bpo-graph patterns in a finite set of outerplanar graphs. A bpo-graph pattern effectively represents connection patterns between different blocks.

Our final object of this research is to propose an efficient data mining method for extracting more expressive outerplanar graph-structured patterns, which simultaneously represent connection patterns and internal structured patterns common to different blocks. First, we introduce a new graph pattern having structured variables, called an *externally extensible outerplanar graph pattern* (*eeo-graph pattern* for short). A variable of an eeo-graph pattern is a vertex pair that does not currently exist in the graph and that becomes an external edge if added as an edge to the graph. Then a variable of an eeo-graph pattern is called an *external variable*. External variables can be replaced with arbitrary connected outerplanar graphs having at least two vertices. In Fig. 1, a graph pattern p is an example of eeo-graph patterns, and an outerplanar graph G is obtained from p by removing variables h_1, h_2, and h_3, and identifying vertex pairs $[u_1, u_2]$, $[u_3, u_4]$, and $[u_5, u_6]$ of p with $[v_1, v_2]$ of g_1, $[w_1, w_2]$ of g_2, and $[z_1, z_2]$ of g_3, respectively. In this paper, we propose a polynomial time algorithm for deciding whether or not there is such a variable replacement by which a given connected outer-

*Department of Informatics, Kyushu University, 744 Motooka, Nishi-Ku, Fukuoka 819-0395, Japan, Email: {h-yama@i, takashi.yamada@inf, shoudai@inf}.kyushu-u.ac.jp

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

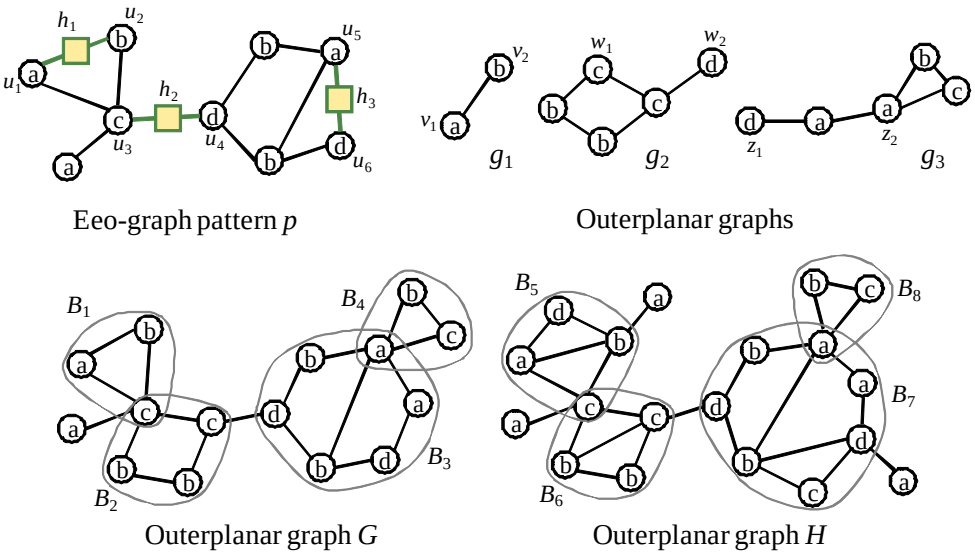

Figure 1: An eeo-graph pattern p and outerplanar graphs g_1, g_2, g_3, G, and H over a vertex label set $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$: A variable is drawn by a box with lines to its elements.

planar graph is obtained from a given eeo-graph pattern.

This paper is organized as follows. In Sec. 2, we give a formal definition of eeo-graph patterns. In Sec. 3, we propose a polynomial time pattern matching algorithm for eeo-graph patterns. In Sec. 4, we conclude this paper with our future work.

2 Externally Extensible Outerplanar Graph Patterns

This paper deals with undirected simple graphs whose vertices and edges are labeled. Let Λ and Δ be two alphabets each of whose elements is called a *vertex label* and an *edge label*, respectively. A list is denoted by a collection of elements enclosed in parentheses, e.g. $[u_1, u_2, u_3]$. The k-th element in a list σ is denoted by $\sigma[k]$. A *graph pattern* is defined as a graph-structured pattern with internal variables, which represents characteristic common structures in graph-structured data. In [7], Uchida et al. introduced a general graph-structured pattern, called a *term graph pattern*, in order to design efficient algorithms for computational problems on graphs.

Definition 1 *Let $G = (V, E)$ be a graph whose vertices and edges are labeled with elements in Λ and Δ, respectively. Let X be a finite alphabet whose elements are called variable labels. A variable of G is a list $[u_1, u_2, \ldots, u_\ell]$ of distinct ℓ vertices of V where $\ell \geq 1$. Variables are labeled with elements in X so that variables of different length have different variable labels. Then, a triple $p = (V, E, H)$ is called a graph pattern if (V, E) is a graph and H is a set of variables*

For a graph pattern p, $V(p)$, $E(p)$ and $H(p)$ denote the sets of all vertices, edges and variables of p, respectively. And $\lambda_p(u)$, $\delta_p(e)$ and $x_p(h)$ denote the vertex label of $u \in V(p)$, the edge label of $e \in E(p)$ and the variable label of $h \in H(p)$, respectively.

Definition 2 *Let p and q be graph patterns. We say that p is isomorphic to q if there exists a bijection $\psi : V(p) \to V(q)$ satisfying the following conditions: (1) ψ is a graph isomorphism from $(V(p), E(p))$ to $(V(q), E(q))$, (2) $[v_1, \ldots, v_\ell] \in H(p)$ if and only if $[\psi(v_1), \ldots, \psi(v_\ell)] \in H(q)$, where $\ell \geq 1$, and (3) for any two variables $[v_1, \ldots, v_\ell]$ and $[v'_1, \ldots, v'_\ell] \in H(p)$, $x_p([v_1, \ldots, v_\ell]) = x_p([v'_1, \ldots, v'_\ell])$ if and only if $x_q([\psi(v_1), \ldots, \psi(v_\ell)]) = x_q([\psi(v'_1), \ldots, \psi(v'_\ell)])$.*

A graph pattern p' is said to be a subgraph pattern of p if $V(p') \subseteq V(p)$, $E(p') \subseteq E(p)$, and $H(p') \subseteq H(p)$. We say that p is subgraph isomorphic to q if p is isomorphic to a subgraph pattern of q.

In an outerplanar embedding of an outerplanar graph G, an edge of G is *external* if it has a border with the outer face, that is, it is not a diagonal of any block. We note that a bridge of G is an external edge, and an external edge which is not a bridge belongs to the outer cycle of a block.

Definition 3 *A graph pattern p is said to be an externally extensible outerplanar graph pattern (eeo-graph pattern for short) if p satisfies the following conditions.*

1. *Any variable has just 2 vertices.*

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

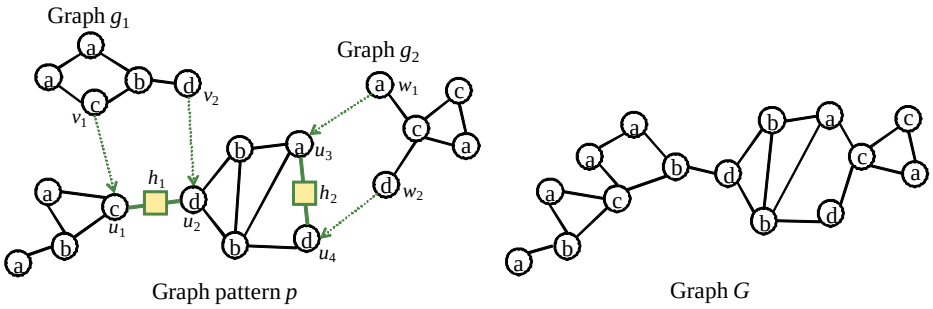

Figure 2: A graph pattern p and graphs g_1, g_2, G over a vertex label set $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$.

2. *$E(p) \cap E_H(p) = \emptyset$, where $E_H(p) = \{(u, v) \mid [u, v] \in H(p)\}$.*
3. *The graph $G_p = (V(p), E(p) \cup E_H(p))$ of p is a connected outerplanar graph, and all edges in $E_H(p)$ are external edges.*
4. *All variable labels of variables in $H(p)$ are mutually distinct.*

A variable h in p is called a block variable if h is not a bridge variable, that is, h is contained in a block of p.

Definition 4 *Let x be a variable label in X. Let p and q be eeo-graph patterns and $h = [v_1, v_2]$ a variable of p labeled with x. Let $\sigma = [u_1, u_2]$ be a list of distinct two vertices in q. The form $x := [q, \sigma]$ is called an eeo-binding for x if*

1. *$\lambda_q(u_1) = \lambda_p(v_1)$ and $\lambda_q(u_2) = \lambda_p(v_2)$, and*
2. *if h is a block variable of p, all edges on a path between u_1 and u_2 are bridges.*

The eeo-binding $x := [q, \sigma]$ is an operation for a variable h labeled with x in which we assign h with q by identifying $\sigma[i]$ with $h[i]$ for each i $(i = 1, 2)$ and removing the variable h from $H(p)$.

Let $p, q_1, \ldots, q_m$ be eeo-graph patterns. A eeo-substitution for p is a finite collection of eeo-bindings $\{x_1 := [q_1, \sigma_1], \ldots, x_m := [q_m, \sigma_m]\}$ where $x_1, \ldots, x_m$ are mutually distinct variable labels in X and each g_i $(1 \leq i \leq m)$ has no variable labeled with a variable label in $\{x_1, \ldots, x_m\}$. For an eeo-graph pattern p and an eeo-substitution θ for p, $p\theta$ denotes the eeo-graph pattern obtained from p and θ by applying all the eeo-bindings in θ to p simultaneously.

We give an example of eeo-substitutions. In Fig. 2, we give a graph pattern p having variables $h_1 = [u_1, u_2]$ and $h_2 = [u_3, u_4]$, so that the graph $p\{x_p(h_1) := [g_1, [v_1, v_2]], x_p(h_2) := [g_2, [w_1, w_2]]\}$ is isomorphic to G.

Let $\theta = \{x_1 := [q_1, \sigma_1], \ldots, x_m := [q_m, \sigma_m]\}$ and $\tau = \{y_1 := [r_1, \sigma'_1], \ldots, y_n := [r_n, \sigma'_n]\}$ be eeo-substitutions. Then the *composition* $\theta\tau$ of θ and τ is the eeo-substitution obtained from $\{x_1 := [q_1\tau, \sigma_1], \ldots, x_m := [q_m\tau, \sigma_m], y_1 := [r_1, \sigma'_1], \ldots, y_n := [r_n, \sigma'_n]\}$ by deleting (i) $x_i := [q_i\tau, \sigma_i]$ which is trivial, and (ii) $y_j := [r_j, \sigma'_j]$ with $y_j = x_k$ for some k $(1 \leq k \leq m)$. Then we have the following proposition.

Proposition 1 *Let θ, τ and γ be eeo-substitutions.*
(1) $\theta\tau$ is an eeo-substitution.
(2) $(p\theta)\tau = p(\theta\tau)$ for any eeo-graph pattern p.
(3) $(\theta\tau)\gamma = \theta(\tau\gamma)$.

The reason why we restrict the number of vertices in a variable is that the pattern matching problem for a graph pattern which has variables with more than 3 variables is hard to solve in polynomial time, because the pattern matching problem for a tree-structured term graph, called a term tree, is NP-complete [4].

Given a graph pattern p and a graph G, an operator which decides whether or not p matches G is called a matching operator. Because the subgraph isomorphism problem for outerplanar graphs is known to be NP-complete, it is hard to solve for deciding whether a graph pattern matches an outerplanar graph by using a subgraph isomorphism matching operator. In this paper, we employ a restricted subgraph isomorphism introduced by [3] as a matching operator.

Definition 5 *For two graphs G and H, we say that G is bridge and block preserving (BBP) subgraph isomorphic to H if there exists a subgraph isomorphism ψ from G to H which maps (i) the set of bridges of G to the set of bridges of H and (ii) different blocks of G to different blocks of H.*

We denote by $\mathcal{O}_{\Lambda,\Delta}$ the set of all connected outerplanar graphs over a vertex label set Λ and an edge label set Δ, and denote by $\mathcal{EOP}_{\Lambda,\Delta}$ the set of all eeo-graph patterns over a vertex label set Λ and an edge label set Δ.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Definition 6 *For an eeo-graph pattern $p \in \mathcal{EOP}_{\Lambda,\Delta}$ and an outerplanar graph $G \in \mathcal{O}_{\Lambda,\Delta}$, we say that p matches G if there is an eeo-substitution θ such that $p\theta$ is BBP subgraph isomorphic to G.*

For example, in Fig. 1, from p and $\theta = \{x_p(h_1) := [g_1, [v_1, v_2]], x_p(h_2) := [g_2, [w_1, w_2]], x_p(h_3) := [g_3, [z_1, z_2]]\}$, we obtain the outerplanar graph $p\theta$ which is isomorphic to G. The graph G is BBP subgraph isomorphic to H such that blocks B_1, B_2, B_3, and B_4 of G correspond to B_5, B_6, B_7, and B_8 of H, respectively. Therefore, p matches both G and H.

3 A Pattern Matching Algorithm for EEO-Graph Patterns

In this section, we give a polynomial time matching algorithm for solving the following problem.

Matching Problem for $\mathcal{EOP}_{\Lambda,\Delta}$
Input: An eeo-graph pattern $p \in \mathcal{EOP}_{\Lambda,\Delta}$ and an outerplanar graph $G \in \mathcal{O}_{\Lambda,\Delta}$.
Problem: Decide whether or not p matches G.

For eeo-graph patterns p and q in $\mathcal{EOP}_{\Lambda,\Delta}$, and vertices $r \in V(p)$ and $r' \in V(q)$, we say that the rooted eeo-graph pattern p^r is BBP subgraph isomorphic to the rooted eeo-graph pattern $q^{r'}$ if there exists a BBP subgraph isomorphism $\psi : V(p) \to V(q)$ such that $\psi(r) = r'$. Let G be an outerplanar graph in $\mathcal{O}_{\Lambda,\Delta}$ and s a vertex in $V(G)$. For a rooted eeo-graph pattern p^r and a rooted outerplanar graph G^s, we say that p^r matches G^s if there exists an eeo-substitution θ such that $p^r\theta$ $(= (p\theta)^r)$ is BBP subgraph isomorphic to G^s.

The idea of our algorithm for the matching problem is similar to the matching algorithm for bpo-graph patterns in [8]. First we fix one vertex of p as its root. Next, for every vertex of G, we specify it as the root of G, and proceed to construct correspondences between all vertices of p and G in bottom up manner, that is, from the leaves to the root of G. If p is a bpo-graph pattern, in a process of the algorithm, we decide whether or not a block of G is isomorphic to a block of p. However, if p is an eeo-graph pattern, we have to decide whether or not a block of G is matched by a block of p having block variables. Therefore, in this paper, we focus on the following subproblem and give a polynomial time algorithm for solving it.

Matching Problem for Biconnected EEO-Graph Patterns
Input: A rooted biconnected eeo-graph pattern p^r and a rooted biconnected outerplanar graph G^s.
Problem: Decide whether or not p^r matches G^s.

A sequence of vertices $[u_1, u_2, \ldots, u_k]$ of a graph pattern p is called a path (resp. cycle) of p if it forms a path (resp. cycle) of the underlying graph of p. For a rooted biconnected eeo-graph pattern p^r, we identify vertices on the outer cycle of p^r (i.e., the boundary of an outerplanar embedding of the underlying graph of p^r) with numbers $1, \ldots, |V(p^r)|$ in the clockwise or counterclockwise order such that r is identified as 1.

Lemma 1 *Let p^r (resp. G^s) be a rooted biconnected eeo-graph pattern (resp. rooted biconnected outerplanar graph) with vertices identified with numbers $1, 2, \ldots, n$ (resp. $1, 2, \ldots, N$) such that r (resp. s) is identified as 1 and the vertex sequence $1, 2, \ldots, n$ (resp. $1, 2, \ldots, N$) forms the outer cycle of p^r (resp. G^s). If p^r matches G^s, there exists an eeo-substitution θ for p^r, and a subgraph isomorphism $\psi : V(p\theta) \to V(G)$ such that $\psi(1) = 1$ (i.e., $\psi(r) = s$) and the vertex sequence $\psi(2), \psi(3), \ldots, \psi(n)$ is either an increasing number sequence or a decreasing number sequence.*

Proof. Let C be a simple cycle (i.e., cycle with no repeated vertex) $[i_1, i_2, \ldots, i_m]$ of G^s such that $i_1 = 1$. We assume that there exists an index k such that $2 < k < m$, $i_2, \ldots, i_k$ is an increasing number sequence and $i_{k+1} < i_k$. Then, there exists an index h such that $1 \leq h < k$ and $i_h < i_{k+1} < i_{h+1}$. On the outer cycle of G^s, there exists a path reaching from i_h to i_{k+1}, denoted by P_1, and a path reaching from i_{k+1} to i_{h+1}, denoted by P_2. Let P_3 be the path $[i_{k+1}, \ldots, i_m, i_1]$ on C, and P_4 the path $[i_k, i_k + 1, \ldots, N, 1]$ on the outer cycle of G^s. Since the outer cycle of G^s and C are simple cycles, P_3 does not contain i_k and P_4 does not contain i_{k+1}. Since P_3 and P_4 share at least one vertex, let ℓ be the first cross point of P_3 and P_4. Let P_5 be a path consisting of $\ell, \ldots, N$ on the outer cycle of G^s and $i_1, \ldots, i_{h+1}$ on C.

Suppose that $k \neq h+1$. We note that P_5 does not contain i_k and i_{k+1}. Let P_6 be the path $[i_{h+1}, i_{h+2}, \ldots, i_k]$ on C. Then, a subgraph of G^s consisting of P_2, P_3, P_4, P_5, P_6 and the edge (i_k, i_{k+1}) is homeomorphic to K_4 (the complete graph on four vertices). It is well-known fact that any outerplanar graph contains no subgraph homeomorphic to K_4. Suppose that $k = h + 1$. We note that $i_h > i_1$ and $i_h \neq \ell$. Then, a subgraph of G^s consisting of P_1, P_3, P_4, P_5, and the edge (i_k, i_{k+1}) is homeomorphic to K_4. It contradicts the fact that G^s is an outerplanar graph. Therefore, $i_2, \ldots, i_m$ is either an increasing number sequence or a decreasing number sequence.

From this fact, for any biconnected outerplanar subgraph g of G^s containing s, the outer cycle of g has the vertex sequence as an increasing number sequence. From the definition of eeo-graph patterns, since the order of vertices of the outer cycle of p^r is not changed by any eeo-substitution, consequently, the lemma holds. □

Hereafter, we suppose that all vertices in p^r and G^s are identified as $1, 2, \ldots, n$ and $1, 2, \ldots, N$, respectively,

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

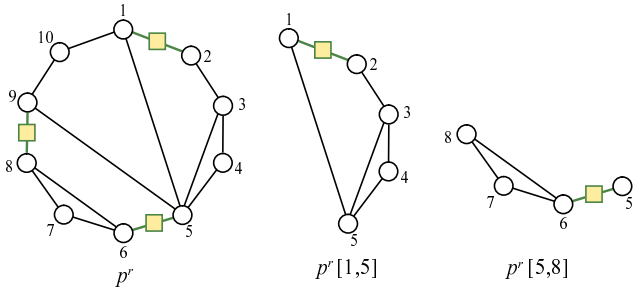

Figure 3: A rooted biconnected eeo-graph pattern and its subgraph patterns.

in the clockwise order or counterclockwise order of boundaries of outerplanar embeddings of p^r and G^s.

For vertices i and i' $(i < i')$ in p^r, $p^r[i, i']$ denotes the subgraph pattern obtained from the induced subgraph $p[\{i, i+1, \ldots, i'\}]$ by removing a variable $[i, i'] \in H(p)$ if it exists. For example, in Fig. 3 we give a rooted biconnected eeo-graph pattern p^r and its subgraph patterns $p^r[1, 5]$ and $p^r[5, 8]$. For vertices i and i' $(i < i')$ in p^r, and vertices j and j' $(j < j')$ in G^s, we say that $p^r[i, i']$ matches $G^s[j, j']$ if there exists an eeo-substitution θ for $p^r[i, i']$ and a subgraph isomorphism $\psi : V(p^r[i, i']\theta) \to V(G^s[j, j'])$ such that $\psi(i) = j$ and $\psi(i') = j'$. For vertices i and i' $(i < i')$ in p^r, the *correspondence-set* (*C-set* for short) of the pair (i, i'), denoted by $CS(i, i')$, is the set of all pairs (j, j') of vertices in G^s such that $p^r[i, i']$ matches $G^s[j, j']$.

Lemma 2 *For a vertex i in p^r where $1 \leq i < n$, and vertices j and j' $(j < j')$ in G^s, $(j, j') \in CS(i, i+1)$ if and only if either of the following conditions holds.*

(1) $(i, i+1) \in E(p^r)$, $(j, j') \in E(G^s)$, $\lambda_p(i) = \lambda_G(j)$, $\lambda_p(i+1) = \lambda_G(j')$, *and* $\delta_p((i, i+1)) = \delta_G((j, j'))$.

(2) $[i, i+1] \in H(p^r)$, $\lambda_p(i) = \lambda_G(j)$, *and* $\lambda_p(i+1) = \lambda(j')$.

Proof. If $(i, i+1) \in E(p^r)$, it is easy to see that statement (1) holds. If $[i, i+1] \in H(p^r)$, $p^r[i, i+1]$ consists of two vertices i and $i+1$, and has no edge and no variable. Therefore, statement (2) holds. □

Statement (2) of the above lemma means a variable $[i, i+1]$ supplements any subgraph $G^s[j, j']$ such that $\lambda_p(i) = \lambda_G(j)$ and $\lambda_p(i+1) = \lambda(j')$. We note that any variable in p^r forms either $[i, i+1]$ $(1 \leq i < n)$ or $[1, n]$.

Let i and i' be vertices in p^r such that $i' - i > 1$ and they border on a same inner face, in other words, if an edge (i, i') exists, (i, i') keeps outerplanarity of p^r. The *chordless path* of $p^r[i, i']$, denoted by $P(p^r[i, i'])$, is a simple path $[i_1, i_2, \ldots, i_m]$ such that $i_1, i_2, \ldots, i_m$ is an increasing number sequence satisfying that $m > 2$, $i_1 = i$, $i_m = i'$, and, except for edges and variables constituting the path, there is possibly only one edge (i_1, i_m) or variable $[i_1, i_m]$ between all vertices in the path.

For example, in Fig. 3, we can see that chordless paths of $p^r[1, 5]$ and $p^r[5, 8]$ are $[1, 2, 3, 5]$ and $[5, 6, 8]$, respectively.

Proposition 2 *Let i and i' be vertices in p^r such that $i' - i > 1$ and they border on a same inner face, and let $P(p^r[i, i']) = [i_1, i_2, \ldots, i_m]$. If $(i, i') \notin E(p^r)$, then the vertices $i_1, i_2, \ldots, i_m$ are cutpoints of $p^r[i, i']$.*

Lemma 3 *Let i and i' be vertices in p^r such that $i' - i > 1$ and they border on a same inner face, and j and j' vertices in G^s such that $j < j'$. Let $P(p^r[i, i']) = [i_1, i_2, \ldots, i_m]$. Then, $(j, j') \in CS(i, i')$ if and only if the following conditions hold.*

(1) If $(i, i') \in E(p^r)$, $(j, j') \in E(G^s)$ and $\delta_p((i, i')) = \delta_G((j, j'))$.

(2) There exists a vertex j'' in G^s such that $j < j'' < j'$, $(j, j'') \in CS(i_1, i_{m-1})$ and $(j'', j') \in CS(i_{m-1}, i_m)$.

Proof. Suppose that the if-statement holds. There exists an eeo-substitution θ_1 for $p^r[i_1, i_{m-1}]$ (resp. θ_2 for $p^r[i_{m-1}, i_m]$), and a subgraph isomorphism ψ_1 from g_1 to $G^s[j, j'']$ where $g_1 = p^r[i_1, i_{m-1}]\theta_1$ (resp. ψ_2 from g_2 to $G^s[j'', j']$ where $g_2 = p^r[i_{m-1}, i_m]\theta_2$) such that $\psi_1(i_1) = j$ and $\psi_1(i_{m-1}) = j''$ (resp. $\psi_2(i_{m-1}) = j''$ and $\psi_2(i_m) = j'$). Let θ be the union of θ_1 and θ_2, and $g = p^r[i, i']\theta$. Let ψ be the injection from $V(g)$ to $V(G^s[j, j'])$ such that $\psi(k) = \psi_1(k)$ for each $k \in V(g_1)$ and $\psi(k) = \psi_2(k)$ for each $k \in V(g_2)$. Because statement (1) holds and, from Proposition 2, there is possibly only one edge (i, i') between $V(g_1) \setminus \{i_{m-1}\}$ and $V(g_2) \setminus \{i_{m-1}\}$, we see that for each vertices k and k' in g, if $(k, k') \in E(g)$ then $(\psi(k), \psi(k')) \in E(G^s)$ where $\delta_g((k, k')) = \delta_G((\psi(k), \psi(k')))$. Therefore, ψ is a subgraph isomorphism from g to $G^s[j, j']$.

Conversely, suppose that there exists an eeo-substitution θ for $p^r[i, i']$, and a subgraph isomorphism ψ from g to $G^s[j, j']$ where $g = p^r[i, i']\theta$, such that $\psi(i) = j$ and $\psi(i') = j'$. It is easy to see that statement (1) holds. By assuming $(i, i') \in E(p^r)$, we regard g and $G^s[j, j']$ as biconnected outerplanar graphs. We suppose that all vertices on the outer cycle of g are identified as $1, 2, \ldots, \ell$ such that i and i' are identified as 1 and ℓ, respectively, and i_{m-1} is identified as ℓ' in g. From Lemma 1, the vertex sequence $\psi(1), \psi(2), \ldots, \psi(\ell)$ is an increasing number sequence. Let $j'' = \psi(\ell')$, $g[1, \ell']$ and $g[\ell', \ell]$ are subgraph isomorphic to $G^s[j, j'']$ and $G^s[j'', j']$, respectively. Let θ_1 and θ_2 be subsets of θ which consist of all

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

eeo-bindings for variables in $p^r[i_1, i_{m-1}]$ and $p^r[i_{m-1}, i_m]$, respectively. Then, $g[1, \ell']$ (resp. $g[\ell', \ell]$) is obtained from $p^r[i_1, i_{m-1}]$ and θ_1 (resp. $p^r[i_{m-1}, i_m]$ and θ_2). Accordingly, $p^r[i_1, i_{m-1}]$ matches $G^s[j, j'']$ and $p^r[i_{m-1}, i_m]$ matches $G^s[j'', j']$. □

Lemma 4 *Let $(i, i') \in E(p^r)$ be a diagonal in p^r such that $i < i'$, and j and j' vertices in G^s such that $j < j'$. Let $P(p^r[i, i']) = [i_1, i_2, \ldots, i_m]$. Then, $(j, j') \in CS(i, i')$ if and only if $(j, j') \in E(G^s)$ where $\delta_p((i, i')) = \delta_G((j, j'))$ and there exists an increasing number sequence $j_1, j_2, \ldots, j_m$ satisfying that $j_1 = j$, $j_m = j'$, and $(j_k, j_{k+1}) \in CS(i_k, i_{k+1})$ for each k $(1 \leq k < m)$.*

Proof. It is immediate from Lemma 3. □

Lemma 5 *Let p^r (resp. G^s) be a rooted biconnected eeo-graph pattern (resp. rooted biconnected outerplanar graph) with vertices on the outer cycle identified with numbers $1, 2, \ldots, n$ (resp. $1, 2, \ldots, N$) such that $r = 1$ (resp. $s = 1$). Then, p^r matches G^s and the vertex sequence $1, 2, \ldots, n$ is mapped to an increasing number sequence by an subgraph isomorphism $\psi : V(p^r\theta) \to V(G^s)$ for an eeo-substitution θ if and only if $(1, i) \in CS(1, n)$ for a vertex $i \in V(G^s)$.*

Proof. If $(1, n) \in E(p^r)$, it is easy to see that the lemma holds. If $[1, n] \in H(p^r)$, the variable $[1, n]$ can supplement the path $[i, i+1, \ldots, N, 1]$, therefore, p^r matches G^s. □

Using Lemmas 1 – 5, we give an algorithm for solving the matching problem for biconnected eeo-graph patterns. For any vertex i in p^r, any vertex j in G^s corresponding to i satisfies $i \leq j \leq i + N - n$. Therefore, in order to decide whether or not p^r matches G^s, we only need to compute subsets of C-set, $CS_d(i, i') = \{(j, j') \in CS(i, i') \mid i \leq j \leq i + N - n$ and $i' \leq j' \leq i' + N - n\}$ for each $(i, i') \in E(p^r)$ and $[i, i'] \in H(p^r)$. Hence, we give a pattern matching algorithm MATCHBLOCK-$\mathcal{EOP}_{\Lambda,\Delta}$ in Fig. 4. for computing the sets $CS_d(i, i')$ for all edges $(i, i') \in E(p^r)$ and variables $[i, i'] \in H(p^r)$ by using a dynamic programming manner. The algorithm assigns all C-sets of pair $(i, i+1)$ where $1 \leq i < n$ first, and it assigns each C-set of pair (i, i') such that $(i, i') \in E(p^r)$ is a diagonal and for its chordless path $[i_1, i_2, \ldots, i_m]$, the C-sets of all pairs (i_k, i_{k+1}) have been already assigned for all k $(1 \leq k < m)$. For an eeo-subgraph pattern $p^r[i, i']$ and an outerplanar subgraph $G^s[j, j']$, Procedure MATCHSUBBLOCK in Fig. 5 computes a subset of $CS_d(i, i')$, the set $\{(j, k) \in CS_d(i, i') \mid k \leq j'\}$. The assignment of C-sets terminates when the C-set of pair $(1, n)$ is assigned, and if $(1, i) \notin CS_d(1, n)$ for all $i \in V(g^s)$, we identify vertices of p^r in reverse and assign C-sets again.

Then, we have the following lemma.

Algorithm: MATCHBLOCK-$\mathcal{EOP}_{\Lambda,\Delta}$;
Input: a biconnected eeo-graph pattern p^r and a biconnected outerplanar graph G^s;
Output: TRUE or FALSE;
begin
1: Construct outerplanar embeddings of p^r and G^s;
2: Identify vertices of the boundaries of embeddings with numbers $1, \ldots, n(= |V(p^r)|)$ and $1, \ldots, N(= |V(G^s)|)$, respectively, in the clockwise order such that $r = 1$ and $s = 1$;
3: **foreach** $(i, i') \in V(p^r) \times V(p^r)$ s.t. $i < i'$ and $(i, i') \in E(p^r)$ or $[i, i'] \in E(p^r)$ **do begin**
4: $CS_d(i, i') := \emptyset$;
5: **if** $i' = i + 1$ and $(i, i') \in E(p^r)$ **then begin**
6: **foreach** $(j, j') \in E(G^s)$ s.t. $j < j'$, $i \leq j \leq i + N - n$, $i' \leq j' \leq j' + N - n$, $\lambda_p(i) = \lambda_G(j)$, $\lambda_p(i') = \lambda_G(j')$ and $\delta_p((i, i')) = \delta_G((j, j'))$ **do** add (j, j') to $CS_d(i, i')$
7: **end**
8: **else if** $i' = i + 1$ and $[i, i'] \in H(p^r)$ **then**
9: **foreach** $(j, j') \in V(G^s) \times V(G^s)$ s.t. $j < j'$, $i \leq j \leq i + N - n$, $i' \leq j' \leq j' + N - n$, $\lambda_p(i) = \lambda_G(j)$, and $\lambda_p(i') = \lambda_G(j')$ **do** add (j, j') to $CS_d(i, i')$
10: **end**
11: **end**;
12: Let S is an empty stack;
13: **for** $i' := n - 1$ **downto** 3 **do begin if** $(1, i') \in E(p^r)$ **then** push($(1, i')$, S) **end**;
14: **for** $i := 2$ **to** $n - 1$ **do begin**
15: **for** $i' := n$ **downto** $i + 2$ **do begin if** $(i, i') \in E(p^r)$ **then** push((i, i'), S) **end**
16: **end**;
17: **while** S is not empty **do begin**
18: $(i, i') :=$ pop(S);
19: **foreach** $(j, j') \in E(p^r)$ s.t. $i \leq j \leq i + N - n$ and $j' = \max\{k \mid j < k,\ i' \leq k \leq i' + N - n$ and $(j, k) \in E(G^s)\}$ **do begin**
20: $C :=$MATCHSUBBLOCK$(p^r[i, i'], G^s[j, j'])$;
21: **foreach** $k \in C$ **do** add (j, k) to $CS_d(i, i')$
22: **end**
23: **end**;
24: $C :=$MATCHSUBBLOCK$(p^r[1, n], G^s[1, N])$;
25: **if** $C \neq \emptyset$ **then return** TRUE;
26: Identify vertices of p^r with $1, \ldots, n$ in the counterclockwise order;
27: Do lines 3 – 25;
28: **return** FALSE
end.

Figure 4: A pattern matching algorithm for biconnected eeo-graph patterns.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Procedure: MATCHSUBBLOCK;
Input: an eeo-subgraph pattern $p^r[i,i']$ and an outerplanar subgraph $G^s[j,j']$;
Output: the set $\{k \in V(G^s) \mid j < k \leq j'$ and $p^r[i,i']$ matches $G^s[j,k]\}$;
begin
1: $P(p^r[i,i']) := [i_1, i_2, \ldots, i_m]$; // $i_1 = i,\ i_m = i'$.
2: $CS_{i,j}(i_1) := \{j\}$;
3: **for** $\ell := 2$ **to** m **do begin**
4: $CS_{i,j}(i_\ell) := \emptyset$; // $CS_{i,j}(i_\ell)$ means the set $\{k \in V(G^s) \mid (j,k) \in CS_d(i_1, i_\ell)\}$.
5: **foreach** $k \in CS_{i,j}(i_{\ell-1})$ **do begin**
6: **foreach** $(k,k') \in CS_d(i_{\ell-1}, i_\ell)$ **do begin**
7: **if** $\ell < m$ **then** add k' to $CS_{i,j}(i_\ell)$
8: **else if** $(i_1, i_m) \in E(p^r)$ and $(j,k') \in E(G^s)$ s.t. $\delta_p((i_1,i_m)) = \delta_G((j,k'))$ **then**
9: add k' to $CS_{i,j}(i_m)$
10: **else if** $[i_1, i_m] \in H(p^r)$ **then** add k' to $CS_{i,j}(i_m)$
11: **end**
12: **end**
13: **end**;
14: **return** $CS_{i,j}(i')$
end.

Figure 5: A procedure for computing a subset of $CS_d(i,i')$ w.r.t. an outerplanar subgraph $G^s[j,j']$.

Lemma 6 *For a rooted biconnected eeo-graph pattern p^r and a rooted biconnected outerplanar graph G^s, the problem of deciding whether or not p^r matches G^s is correctly solvable in $O(nN^2)$ time, where $n = |V(p)|$ and $N = |V(G)|$.*

Proof. Since $|E(p)| + |H(p)| \leq 2|V(p)| - 3$, we need $O(n(N-n)^2)$ time at the initial stage (lines 3 – 11) of MATCHBLOCK-$\mathcal{ECOP}_{\Lambda,\Delta}$. Let $p^r[i,i']$ be a subgraph pattern of p^r and $P(p^r[i,i']) = [i_1, i_2, \ldots, i_m]$. In Procedure MATCHSUBBLOCK, each vertex i_ℓ $(1 \leq \ell \leq m)$ is assigned at most $N - n$ vertices of G^s, and therefore, MATCHSUBBLOCK works in $O(m(N-n))$ time. Since the total lengths of chordless paths of diagonals in p^r is $|E(p)| + |H(p)| - 1$, we need $O(nN(N-n))$ time at lines 17 – 23 of MATCHBLOCK-$\mathcal{ECOP}_{\Lambda,\Delta}$. Therefore, the total time is $O(nN^2)$ time. □

Theorem 1 *The matching problem for an eeo-graph pattern $p \in \mathcal{ECOP}_{\Lambda,\Delta}$ and an outerplanar graph $G \in \mathcal{O}_{\Lambda,\Delta}$ is computable in $O(nN^3)$ time, where n and N are the numbers of vertices of p and G, respectively.*

Proof. Let u and v be vertices of rooted eeo-graph pattern p^r and rooted outerplanar graph $G^s[v]$, respectively. From Lemma 6 and the proof of Corollary 1 in [8], in order to decide whether or not $p^r[u]$ matches $G^s[v]$, we need $O(c_u c_v^2)$ time where c_u and c_v are the numbers of children of u and v, respectively. Then, the runtime for deciding whether or not p^r matches G^s is $O(nN^2)$ time, and consequently, the total time is $O(nN^3)$ time. □

4 Conclusions

In this paper, we gave a polynomial time algorithm for deciding whether or not a given eeo-graph pattern matches a given connected outerplanar graph. By using this algorithm, in [9], we gave a refinement-based frequent mining algorithm for eeo-graph patterns and evaluated its performance by experiments on subsets of the NCI dataset. As future works, toward efficient graph mining systems for real-world databases, we are studying more expressive graph pattern classes based on graph transformation systems.

References

[1] D.J. Cook and L. Holder. *Mining Graph Data*. WILEY-INTERSCIENCE, 2007.

[2] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[3] T. Horváth, J. Roman, and S. Wrobel. Frequent subgraph mining in outerplanar graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 197–206, 2006.

[4] T. Miyahara, T. Shoudai, T. Uchida, K. Takahashi, and H. Ueda. Polynomial time matching algorithms for tree-like structured patterns in knowledge discovery. *Proc. PAKDD 2000, LNAI 1805*, pp. 5–16, 2000.

[5] National Cancer Institute. Chemical dataset. *http://cactus.nci.nih.gov/*.

[6] Y. Suzuki, T. Shoudai, T. Uchida, and T. Miyahara. Ordered term tree languages which are polynomial time inductively inferable from positive data. *Theor. Comput. Sci.*, 350, pp. 63–90, 2006.

[7] T. Uchida, T. Shoudai, and S. Miyano. Parallel algorithm for refutation tree problem on formal graph systems. *IEICE Transactions on Information and Systems*, E78-D(2):99–112, 1995.

[8] H. Yamasaki, Y. Sasaki, T. Shoudai, T. Uchida and Y. Suzuki. Learning Block-Preserving Graph Patterns and its Application to Data Mining. *Machine Learning*, 76(1):137–173, 2009.

[9] H. Yamasaki and T. Shoudai. Mining of Frequent Externally Extensible Outerplanar Graph Patterns. *Proc. ICMLA 2008*, pages 871–876, 2008.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)