Designing and Learning Substitutable Plane Graph Grammars

Rémi Eyraud, Jean-Christophe Janodet, Tim Oates, Frédéric Papadopoulos

Fundamenta Informaticae XX (2015) 1–1025
DOI 10.3233/FI-2012-0000
IOS Press

Designing and Learning Substitutable Plane Graph Grammars

Rémi Eyraud
Qarma Team
LIF Marseille, France

Jean-Christophe Janodet
IBISC lab
University of Evry, France

Tim Oates
CoRaL lab, CSEE department
University of Maryland, Baltimore County, USA

Frédéric Papadopoulos
IBISC lab
University of Evry, France

Abstract. Though graph grammars have been widely investigated for 40 years, few learning results exist for them. The main reasons come from complexity issues that are inherent when graphs, and *a fortiori* graph grammars, are considered. The picture is however different if one considers *drawings* of graphs, rather than the graphs themselves. *E.g.*, it has recently been proved that the isomorphism and pattern searching problems could be solved in polynomial time for *plane graphs*, that is, planar embedding of planar graphs. In this paper, we introduce the *Plane Graph Grammars* (PGG) and detail how they differ to usual graph grammar formalisms while at the same time they share important properties with string context-free grammars. In particular, the parsing of a graph with a given PGG is polynomial for languages with appropriate restrictions. We demonstrate that PGG are well-shaped for learning: we show how recent results on string grammars can be extended to PGG by providing a learning algorithm that identifies in the limit the class of *substitutable* plane graph languages. Our algorithm runs in polynomial time assuming the same restriction used for polynomial parsing, and the amount of data needed for convergence is comparable to the one required in the case of strings.

1. Introduction

Graph Grammars have been defined and studied for four decades from a language-theoretical standpoint (see [29] for an overview), but the learning of these formalisms is known to be intricate and has hardly been investigated in the literature yet. Most contributions concern heuristics tailored for graphs involved in restricted application domains. This is the case of both most famous algorithms, *Subdue* [5] and *FFSM* [21], and their extensions [27, 23, 25].

Address for correspondence: R. Eyraud, 39 rue F. Joliot Curie 13453, Marseille cedex 13, France

On the theoretical side, it appears that learnability results are even rarer and often provide us with preliminary results, rather than effective learning procedures. For instance, E. Jeltsch and H.-K. Kreowski [22] give an algorithm that generates the set of grammars consistent with a given set of graphs. R. Brijder and H. Blockheel [2] investigate the inference of a grammar consisting of a single production rule, given a graph and a distinguished pattern with many occurrences. Few necessary conditions for the learning of graph grammars have also been established under unrestricted Gold's paradigm [6]. Also note that new learnability results are anticipated in the framework of recognizable series of (hyper-)graphs [1]. The situation is a bit different in other branches of Machine Learning: several techniques have been proposed to tackle classification problems over graphs in the scope of social network analysis, biological network analysis or image analysis [18]. However, they generally hide the complexity of the graph structures into abstract numerical structures such as graph kernels [32].

There are many reasons for this, ranging from the profusion of incomparable graph grammar formalisms to the hardness of the model itself. Concerning the latter, many basic problems, such as the search for a subgraph in a graph, and thus the possibility to parse a graph with a grammar, are generally $\mathcal{NP}$-complete [15]. Nevertheless, the main reason for this absence of positive learning results is probably that no kind of graph grammars was designed with the aim of learning. Indeed, two main characteristics have to be shared by the representation formalism if one wants to use it as a model for inference. On the one hand, the graph isomorphism problem needs to be efficiently solvable: the key point to learn graph grammars is to extract knowledge about the structure of the graphs in the learning sample, and thus the names of the vertices are irrelevant for instance. If less importance is given to the understanding of the structure, general machine learning methods can be applied to graph data with great success [32].

The second important characteristics is that the grammar formalism has to capture properties that are observable in a set of data. The most obvious kind of observable properties concerns sub-structures, *e.g.* the frequency or the relative positions of the subgraphs in the graphs of the sample. But standard graph grammars of different types are not designed for the inference from the observation of properties of sub-structures. For instance, in the framework of Hyperedge-Replacement Grammars (HRG) [8], we can compute from a sample the *set* of external nodes for each sub-hypergraph. However, this set of vertices must be transformed into a *sequence* during the inference stage of a HRG, as this sequence is necessary for the embedding mechanism that is used when a rule is applied to rewrite a hypergraph. In other words, an essential piece of information for the inference of a HRG is not observable in the sample.

From a general standpoint, one way to tackle the difficulties raised by the learning of generative devices (grammars) consists in restricting the languages. That is, the successful approach in Grammatical Inference is often to determine features that are learnable, which usually correspond to observable properties in any set of examples, and then to focus only on the languages that share these characteristics.

Hence, in the case of graph languages, we should first determine which kind of graphs are likely to be learnable, and then choose the kind of grammars to use. For reasons that will be developed in this paper, a promising candidate is the class of *plane graphs*, that is, planar graphs embedded in the plane (see Fig. 1 for an example). Note that a *planar* graph has a set X of vertices and a set E of edges as usual, but as soon as this graph is embedded in the plane, it also has a set F of faces. A planar graph may have several incomparable drawings, so we define a plane graph by fixing the embedding. More formally, a plane graph stands for an isotopy class of planar embeddings for a given planar graph [14]. A plane graph is thus a planar graph that is embedded in the plane without edge-crossing and up to continuous deformations. Given a planar embedding of a planar graph, Fáry [13] proved that it is always possible to move the vertices, within the same isotopy class, so that the edges are drawn with straight-line segments.

We shall use such straight-line drawings in the following.

Now, as no common graph grammar formalism captures the specificities of such plane graphs, we choose not to use existing general graph-grammar formalisms, but propose in this paper a new type of grammar, called the *Plane Graph Grammars (PGG)*. These grammars can be seen as face-replacement grammars, thus constitute an interesting alternative to standard node-replacement or hyperedge-replacement grammars. Indeed, their rules replace one face by a new plane graph, which is sewn in the mother graph using a syntactic gluing law. We provide theoretical results about these grammars regarding the possibility to efficiently parse a plane graph, and compare them with other types of graph grammars.

We then investigate the learning of PGG, and prove that one can get formal learnability results in this setting. We believe that this is quite an interesting improvement *w.r.t.* the state of the art. Concerning the difficulties, notice that when one is trying to learn from graphs, negative data are usually not available. We know since the work by E. M. Gold [17] that it is not possible to identify in the limit any superfinite class[1] of languages from positive data, and thus need to restrict ourselves to a subclass of plane graph languages. The recent successes of distributional learning for string grammars [3] and tree grammars [24] motivate us to define an analogue of substitutable context-free languages [4] for plane graph languages.

Notice that a preliminary version of this paper appeared in the Proceedings of ICGI'12 [12], but the present paper is substantially different: in [12] we tackled the problem of learning *Binary* Plane Graph grammars, a restricted type of PGG where the production rules had binary right hand-sides and were thus similar to Chomsky normal forms. Moreover, we omitted the study of their properties. In this paper, we consider general PGG, improve the definition of the rewriting mechanism and propose new conditions for the parsing problem to be achievable in polynomial time. We also improve the learning algorithm and thus establish a more general learnability result for substitutable plane-graph languages.

Preliminaries about plane graphs are given in Section 2. The definition of Plane-Graph Grammars as well as the rewriting mechanism is detailed in Section 3, where we also compare them with node-replacement grammars and hyperedge-replacement grammars. We prove formal properties of these grammars in Section 4, in the scope of the parsing problem. Next Section 5 is devoted to the learning of PGG, and is thus the core of the paper: the substitutability property is first adapted, then the learning procedure is described, and a learnability result is finally proved for substitutable plane graph language. We conclude the paper with a discussion in Section 6.

2. On Plane Graphs

We have introduced the plane graphs using the notion of embeddings, *i.e.*, functions that map vertices to points, and edges to curves. However, this mathematical approach is quite unsuitable for designing algorithms. As the set of faces is the corner stone to describe plane graphs, we introduce *plane graph systems* [20] below, which allow us to describe any *connected* plane graph through its faces.

Let $X \subset \mathbb{N}$ be the alphabet of vertices. Let X^* be the set of all *strings* over X, and ϵ the empty string. Given a string $x = a_1 \dots a_n$, we denote $|x| = n$ its length and x^R the reverse string of x, that is to say $x^R = a_n \dots a_1$. We also define $\texttt{first}(x)$ to be a_1. A *circular string* is intuitively a string in which the last symbol is followed by the first; more precisely, there is neither a first nor a last symbol but a mapping associating to each symbol the next one. We denote a circular string by $[u]$, with the convention

[1] A class is superfinite if it contains all possible finite languages and at least one infinite language

that if u and v are two strings, then $[uv] = [vu]$. The set of all circular strings over X is denoted by $X^\top$. We set $[x]^R = [x^R]$. Finally, given an alphabet X, we can extend any function $\phi : X \to X$ to strings: $\forall x = a_1 \dots a_n \in X^*, \hat{\phi}(x) = \phi(a_1) \dots \phi(a_n)$, to circular strings: $\hat{\phi}([x]) = [\hat{\phi}(x)]$, to sets of strings: $\hat{\phi}(S) = \{\hat{\phi}(x) : x \in S\}$, and to sets of pairs of strings: $\hat{\phi}(C) = \{(\hat{\phi}(x), \hat{\phi}(x')) : (x, x') \in C\}$.

Now consider the plane graph of Figure 1. The outer face is f_1 and bounded (inner) faces are f_2 and f_3. Each face has only one boundary since the graph is connected. Such a boundary can be described by a circular string of vertices in which two consecutive vertices and the last and first vertices are linked by an edge. Conventionally, we follow a boundary by leaving it to the right. In other words, the bounded face is on the left of the walk. Hence, the boundary of face f_3 is [*53634*], or equivalently [*63453*], by circular permutation.

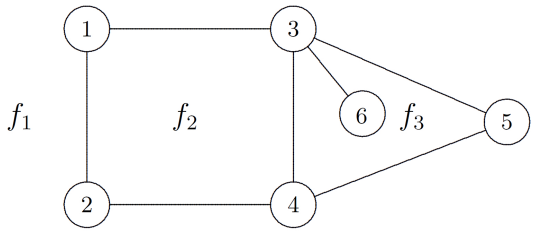

Figure 1. A plane graph with 3 faces.

We now introduce the following description system for connected plane graphs:

Definition 2.1. (Plane Graph System [20])
A *plane graph system* (PGS for short) is a tuple $S = \langle X, E, F, o, \mathcal{D}\rangle$ such that (1) $\langle X, E\rangle$ is a connected planar simple graph [16], (2) F is a finite nonempty set of symbols called the *faces*, (3) $o \in F$ is a special face called the *outer face* and (4) $\mathcal{D} : F \to X^\top$ is a function, called the *boundary descriptor*, that maps any face to its boundary. For sake of readability, we shall make no distinction between a face f and the description of its boundary $\mathcal{D}(f)$. In consequence, function $\mathcal{D}$ will be kept implicit and we simply denote by $\langle X, E, F, o\rangle$ the plane graph system S.

Note that every plane graph can be described with a plane graph system, but the converse does not hold in general. We thus introduce further conditions below:

Definition 2.2. (Valid PGS)
A PGS $S = \langle X, E, F, o, \mathcal{D}\rangle$ is said *valid* if:

1. For all $f \in F$ and $x, y \in X$ and $u \in X^*$, if $\mathcal{D}(f) = [xyu]$ then $\{x, y\} \in E$;
2. For all $e = \{x, y\} \in E$, there exist a unique face $f \in F$ such that $\mathcal{D}(f) = [xyu]$, and an unique face $f' \in F$ such that $\mathcal{D}(f') = [yxv]$, for some $u, v \in X^*$;
3. For all $f \in F$ and $x, y, z, z' \in X$ and $u, v \in X^*$, if $\mathcal{D}(f) = [xyzuxyz'v]$ or $\mathcal{D}(f) = [zxyuz'xyv]$ then $z = z'$;
4. Euler's formula holds, that is, $|X| - |E| + |F| = 2$.

Following result establishes that a valid PGS denotes a plane graph. As the problem is a bit far from the core of this paper, the interested reader may consult its proof in Appendix A:

Theorem 2.3. Any valid PGS describes a plane graph.

Let $\mathbb{G}$ be the set of all plane graph systems. The *size* of a PGS $G = \langle X, E, F, o\rangle$ is $|G| = \sum_{f\in F} |f|$. Given any edge e, we denote by faces(e) the set of faces incident to edge e. Notice that faces(e) can contain either 1 or 2 faces (only one in the case of a *pendant edge*). We use $\texttt{nodes}(f)$ and $\texttt{edges}(f)$ for the set of vertices and edges along the boundary of face f, respectively.

For instance, consider the plane graph of Fig. 1. The corresponding PGS is $S = \langle X, E, F, o\rangle$ with $X = \{1, 2, 3, 4, 5, 6\}$, $E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \{4, 5\}\}$, $F = \{f_1, f_2, f_3\}$, $o = f_1$ and $f_1 = [\textit{13542}]$, $f_2 = [\textit{1243}]$, $f_3 = [\textit{34536}]$. Moreover, we have faces$(\{3, 4\}) = \{f_2, f_3\}$, faces$(\{3, 6\}) = \{f_3\}$, $\texttt{nodes}(f_3) = \{3, 4, 5, 6\}$ and $\texttt{edges}(f_3) = \{\{3, 4\}, \{4, 5\}, \{5, 3\}, \{3, 6\}\}$.

Definition 2.4. (Set of contiguous faces)
Let $S = \langle X, E, F, o\rangle$ be a PGS. Two distinct faces $f, f' \in F$ are *adjacent* if $\exists e \in E$: faces$(e) = \{f, f'\}$. The faces of a subset $K \subseteq F$ are *contiguous* if $\forall f, f' \in K$, a sequence $f = f_0, f_1, \dots f_n = f'$ of faces in K exists such that $\forall i \in \{0, 1, \dots, n-1\}$, f_i and f_{i+1} are adjacent.

Given a subset $K \subseteq F$ of contiguous inner faces, we denote by $\texttt{outer}(K)$ the (boundary of the) outer face of that set. For instance, on the PGS of Fig. 1, $\texttt{outer}(\{f_2, f_3\}) = f_1$ and $\texttt{outer}(\{f_3\}) = [354]$. Notice that $\texttt{outer}(K)$ can be computed in polynomial time using the *normalization* procedure introduced in [20].

Let us finally introduce the notion of subgraph that we will use throughout the rest of this paper:

Definition 2.5. (Pattern)
Given a PGS $G = \langle X, E, F, o\rangle$ and a set $F' \subseteq F \setminus \{o\}$ of contiguous faces, the PGS $G' = \langle \texttt{nodes}(F'), \texttt{edges}(F'), F' \cup \{\texttt{outer}(F')\}, \texttt{outer}(F')\rangle$ is called a *pattern* of G. By extension, any renaming of the vertices and edges and faces of G' will also be called a *pattern* of G.

For instance, the PGS G of Fig. 1 has 3 patterns: $\langle\{1, 2, 3, 4\}, \{(1, 2), \dots\}, \{[1243], [1342]\}, [1342]\rangle$, $\langle\{3, 4, 5, 6\}, \{(3, 4), \dots\}, \{[34536], [354]\}, [354]\rangle$, and G itself.

More general notions of subgraphs exist (based on subsets of vertices and edges, independently on faces), but they often induce intractable problems. In particular, whereas searching for general subgraphs in planar graphs is a $\mathcal{NP}$-complete problem, it was shown in [20] that searching for a pattern in a PGS is tractable in polynomial time. In the following, term *subgraph* will exclusively mean *pattern*.

2.1. Concatenation

The concatenation of two PGS is a basic operation that allows one to glue together two distinct plane graphs using their outer face.

Definition 2.6. Let $G_1 = \langle X_1, E_1, F_1, o_1\rangle$ and $G_2 = \langle X_2, E_2, F_2, o_2\rangle$ be two PGS, and ϕ : $\texttt{nodes}(o_1) \to \texttt{nodes}(o_2)$ a partial bijective function. We say that G_1 and G_2 are *concatenable* following ϕ iff

- $F_1 \cap F_2 = \emptyset$,

- $(X_1 \setminus \texttt{nodes}(o_1)) \cap (X_2 \setminus \texttt{nodes}(o_2)) = \emptyset$,
- $\exists k > 1$: $\hat{\phi}(X_1) \cap X_2 = \{\phi(a_1), \ldots, \phi(a_k)\}$, and $o_1 = [a_1 \ldots a_k y]$ and $o_2 = [\phi(a_k) \ldots \phi(a_1) z]$ with $y \in (X_1 \setminus X_2)^*, z \in (X_2 \setminus X_1)^*$ and $|yz| \geq 1$.

Intuitively, two PGS are concatenable following ϕ if they can be glued together by merging pairwise nodes of their outer face following ϕ. This requires that they can only share nodes of their outer face, and that consecutive edges of one outer face correspond to reverse consecutive edges in (the image by ϕ of) the other outer face. In consequence, the gluing stage does not modify the inner faces.

Definition 2.7. (Concatenation)
Let $G_1 = \langle X_1, E_1, F_1, o_1 \rangle$ and $G_2 = \langle X_2, E_2, F_2, o_2 \rangle$ be two PGS concatenable following a function ϕ. The *concatenation* of G_1 and G_2 following ϕ, written $G_1 \diamond_\phi G_2$, is the PGS $G = \langle X_1 \cup X_2 \setminus \{a_1, \ldots, a_k\}, E_1 \cup E_2 \setminus \{(a_i, a_{i+1}) : 1 \leq i < k\}, \hat{\phi}(F_1 \setminus \{o_1\}) \cup (F_2 \setminus \{o_2\}) \cup \{o\}, o \rangle$ with $o = [\phi(a_k) y \phi(a_1) z]$.

If the function ϕ is the identity, we will write $G_1 \diamond G_2$ instead of $G_1 \diamond_{id.} G_2$.

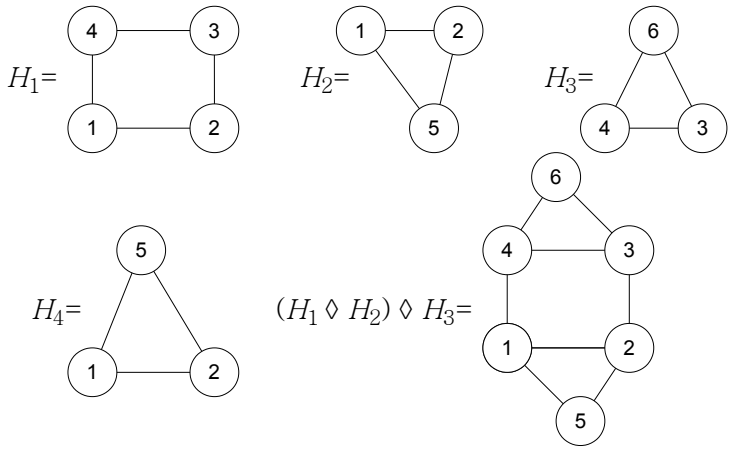

Figure 2. Example of concatenation: H_1 and H_2 are concatenable following the identity function and so is H_2 and H_4. The same occurs for $H_1 \diamond H_2$ and H_3. H_2 and H_3 are not concatenable following the identity function, which is also the case of H_1 and H_4.

Concatenation is well-defined, that is, if G_1 and G_2 are concatenable PGS, then $G_1 \diamond_\phi G_2$ is necessarily a PGS. Indeed, the conditions on the external faces ensure that no new face is created by concatenation, but the outer one which is modified; moreover, if $\langle X_1, E_1 \rangle$ and $\langle X_2, E_2 \rangle$ are connected planar simple graphs then $\langle X_1 \cup X_2 \setminus \{a_1, \ldots, a_k\}, E_1 \cup E_2 \setminus \{(a_i, a_{i+1}) : 1 \leq i < k\} \rangle$ is a connected planar simple graph.

Examples of concatenable and non-concatenable PGS are given in Fig. 2. For instance H_1 and H_2 are concatenable following the identity function but it is not the case of H_1 and H_4 since the third requirement of Definition 2.6 is not met. These examples show also that the associativity of graph

concatenation is not ensured: $(H_1 \diamond H_2)$ and H_3 are concatenable following the identity function, but H_2 and H_3 are not concatenable following the identity and thus $H_1 \diamond (H_2 \diamond H_3)$ is not defined. In absence of brackets in a sequence of concatenations, we will consider the left to right organization: $H_1 \diamond H_2 \diamond H_3$ is to be read as $(H_1 \diamond H_2) \diamond H_3$.

2.2. Plane isomorphism

We finally need a way to compare two PGS:

Definition 2.8. (Plane isomorphism [20])
Let $G_1 = \langle X_1, E_1, F_1, o_1 \rangle$ and $G_2 = \langle X_2, E_2, F_2, o_2 \rangle$ be two PGS. We say that G_1 and G_2 are *plane-isomorphic*, written $G_1 \cong_p G_2$, if there exist a 1-to-1 mapping $\chi : X_1 \to X_2$ over the vertices and a 1-to-1 mapping $\xi : F_1 \to F_2$ over the faces such that (1) the outer face is preserved: $\xi(o_1) = o_2$ and (2) the boundaries are preserved: $\forall f_1 \in F_1, f_2 \in F_2$, if $\xi(f_1) = f_2$ and $f_1 = [a_1 \dots a_n]$ then $f_2 = [\chi(a_1) \dots \chi(a_n)]$.

Plane-isomorphism is decidable in $\mathcal{O}(|E_1| \cdot |E_2|)$ time [20]. The key property to prove this result is that, given a PGS G and an edge e, we can define an ordering over all the other edges which is unique and computable in linear time thanks to a traversal of the PGS. So the isomorphism algorithm consists in finding two edges e_1 in G_1 and e_2 in G_2, generating the ordering of all the edges in both PGS and using them to generate possible isomophism functions. A similar strategy is used to check whether a given PGS is a pattern of any other PGS, that is, searching for patterns is also tractable in polynomial time [20].

We can now define the notion of plane graph language:

Definition 2.9. A (possibly infinite) set L of PGS is a *plane graph language* if it is closed under plane-isomorphism: for all $G_1, G_2 \in \mathbb{G}$ such that $G_1 \cong_p G_2$, we have $G_1 \in L \iff G_2 \in L$.

The following technical but crucial lemma deals with the link between concatenation and plane-isomorphism. Informally, if two graphs are concatanable then graphs that are plane-isomorphic to them are also concatenable (using a different function). Moreover, the concatenated graphs are plane-isomorphic.

Lemma 2.10. Let G_1, G_1', G_2 and G_2' be four PGS such that $G_1 \cong_p G_1'$ and $G_2 \cong_p G_2'$. Suppose that G_1 and G_2 are concatenable following a function ϕ. Then there exist a PGS G_1'' such that $G_1'' \cong_p G_1'$, and a function ϕ' such that G_1'' and G_2' are concatenable following ϕ' and $G_1 \diamond_\phi G_2 \cong_p G_1'' \diamond_{\phi'} G_2'$.

Proof:
As $G_1 \cong_p G_1'$ (resp. $G_2 \cong_p G_2'$), there exist two 1-to-1 mapping $\chi_1 : X_{G_1} \to X_{G_1'}$ (resp. $\chi_2 : X_{G_2} \to X_{G_2'}$) and $\xi_1 : F_1 \to F_1'$ (resp. $\xi_2 : F_2 \to F_2'$) fulfilling the conditions of plane-isomorphism. As G_1 and G_2 are concatenable following ϕ, there exist vertices $a_1, \dots, a_k$ and (possibly empty) sequences of nodes y and z such that $o_{G_1} = [a_1 \dots a_k y]$ and $o_{G_2} = [\phi(a_k) \dots \phi(a_1) z]$. Moreover we have $o_{G_1'} = [\chi_1(a_1) \dots \chi_1(a_k) \hat{\chi_1}(y)]$ and $o_{G_2'} = [\chi_2 \circ \phi(a_k) \dots \chi_2 \circ \phi(a_1) \hat{\chi_2}(z)]$.

Let $\phi' = \chi_2 \circ \phi \circ \chi_1^{-1}$. Let us denote $a_i' = \chi_1(a_i)$ for all $1 \leq i \leq k$. Clearly we have $o_{G_1} = [a_1' \dots a_k' \hat{\chi_1}(y)]$ and $o_{G_2} = [\phi'(a_k') \dots \phi'(a_1') \hat{\chi_2}(z)]$. Moreover, $|\hat{\chi_1}(y) \hat{\chi_2}(z)| = |yz| \geq 1$. Finally, if $X_{G_1'} \setminus \texttt{nodes}(o_{G_1'}) \cap X_{G_2'} \setminus \texttt{nodes}(o_{G_2'}) \neq \emptyset$, we can rename the inner nodes of G_1' to create a new PGS

G_1'' such that none of its inner nodes share the same name with a node of G_2', and $G_1'' \cong_p G_1'$. Therefore $G_1 \cong_p G_1''$ and G_1'' and G_2' are concatenable following ϕ'.

We need to show that $G_1 \diamond_\phi G_2 \cong_p G_1'' \diamond_{\phi'} G_2'$. Let χ_1' and ξ_1' the functions defining the plane-isomorphism between G_1 and G_1''. Let $\chi : X_{G_1 \diamond_\phi G_2} \to X_{G_1'' \diamond_{\phi'} G_2'}$ be the function such that $\chi(v) = \chi_1'(v)$ if $v \in X_{G_1}$ and $\chi(v) = \chi_2(v)$ if $v \in X_{G_2}$. As χ_1' and χ_2 are 1-to-1 mappings over distinct domains, so is χ. Now let $\xi : F_{G_1 \diamond_\phi G_2} \to F_{G_1' \diamond_{\phi'} G_2'}$ be the function such that $\xi(f) = \xi_1'(f)$ if $f \in F_{G_1}$, and $\xi(f) = \xi_2(f)$ if $f \in F_{G_2}$, and $\xi(o_{G_1 \diamond_\phi G_2}) = o_{G_1'' \diamond_{\phi'} G_2'}$. As ξ_1' and ξ_2 are 1-to-1 mapping and $F_{G_1} \cap F_{G_2} = \emptyset$, we deduce that ξ is a 1-to-1 mapping. In addition, as (χ_1', ξ_1') and (χ_2, ξ_2) preserve the boundaries of the faces, it is also the case of (χ, ξ) by construction. □

3. The Grammars for Plane Graph Languages

The idea underling the new graph grammar formalism we are defining here is the following: a grammar consists of rules that explain how to replace a given face into a pattern that can be made of several faces but whose outer face is the same than the one being replaced.

We first need to introduce the analogue to non-terminal symbols for string grammars. The idea is to have a ranked alphabet [10] of non-terminals: in a grammar rule, a non-terminal will be attached to a single face whose number of nodes is equal to its rank.

Definition 3.1. (Plane non-terminal)
A *plane non-terminal* is a couple (N, r) where N is a symbol and r an integer greater than 2 called the rank of the non-terminal. In the following, we assume that any symbol has a unique rank, that is, if (N, r_1) and (N, r_2) are plane non-terminals, then $r_1 = r_2$. We will slightly abuse the notations and write $\texttt{rank}(N)$ for the rank of the non-terminal (N, r).

Intuitively, a non-terminal (N, r) is not a PGS, but *denotes* a PGS, with a unique inner face called N, and no pendant edge, and a boundary delimited with r vertices.

We now define the two types of grammar rules used in our grammars.

Definition 3.2. (Plane graph lexical rule)
A *plane graph lexical rule* is a pair (N_x, G_*), written $N_x \to G_*$, where (1) x is a string, (2) $(N_x, |x|)$ forms a plane non-terminal and (3) $G_* = \langle X, E, F, o \rangle$ is a plane graph system such that $|F| = 2$ and $o = [x^R]$.

Notice that the (unique) inner face of G_* does not have to be $[x]$: there may be pendant edges. A lexical rewrites a non-terminal into a PGS whose outer face is the mirror of the string used in the definition of the non-terminal.

Definition 3.3. (Plane graph production)
A *plane graph production* is a tuple written $N_{x_0}^0 \to N_{x_1}^1 \dots N_{x_k}^k$, $k \geq 2$, where (1) $(N_{x_i}^i, |x_i|)$, $\forall i, 0 \leq i \leq k$, are plane non-terminals, (2) $\forall i \geq 1$, $H_i = \langle X_i, E_i, \{[x_i], [x_i]^R\}, [x_i]^R \rangle$ is a PGS, (3) $\forall i, 1 \leq i < k$, $H_1 \diamond \dots \diamond H_i$ and H_{i+1} are concatenable and (4) $H = H_1 \diamond \dots \diamond H_k = \langle X_H, E_H, \cup_{1 \leq i \leq k} \{[x_i]\} \cup \{[x_0]^R\}, [x_0]^R \rangle$ is a PGS.

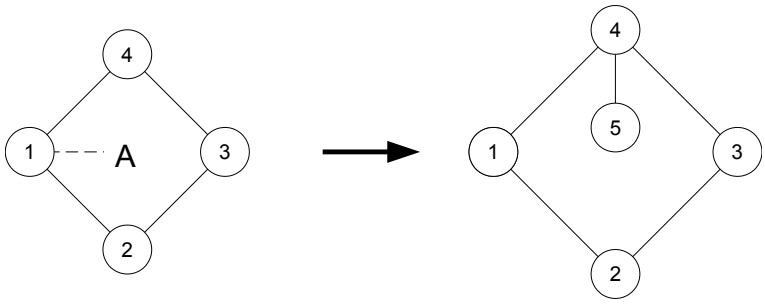

Figure 3. A graphical representation of the plane graph lexical rule $A_{1234} \rightarrow [123454]$. The dashed lines attach the non-terminal to the vertex appearing at the head of the string in the definition of the non-terminal.

A production can be seen as the development of a face f made of $\texttt{rank}(N^0)$ vertices into k adjacent faces whose overall shape is the same as f. Fig. 4 gives a graphical representation of an example of a plane graph production.

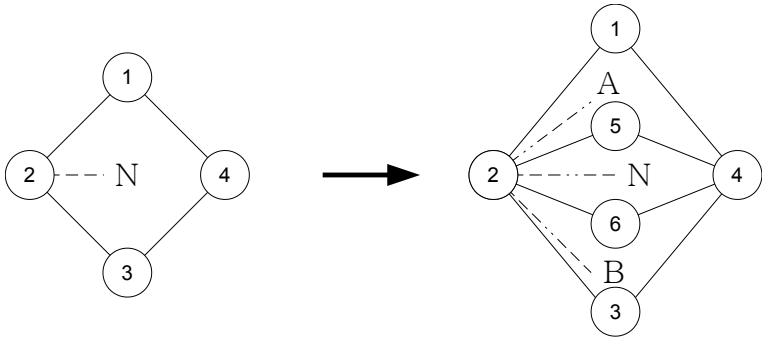

Figure 4. A graphical representation of the (recursive) plane graph production $N_{2341} \rightarrow A_{2541}\ N_{2645}\ B_{2346}$. The dashed lines attach each non-terminal to the vertex appearing at the head of the string corresponding to that face in the rule.

Definition 3.4. (Plane graph grammar)
A *plane graph grammar (PGG)* $\mathcal{G}$ is a tuple $\langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$ such that $\mathcal{N}$ is a set of plane non-terminals, P_L is a set of plane graph lexical rules, P is a set of plane graph productions and $\mathcal{A} \subseteq \mathcal{N}$ is the set of axioms.

Example 3.5. Let $\mathcal{G}_1$ be the grammar $\langle \{(N,4),(A,4),(B,4)\}, P_L, P, \{(N,4)\} \rangle$, with $P_L = \{A_{1234} \rightarrow [123454], B_{1234} \rightarrow [15123634], N_{1234} \rightarrow [1234]\}$ (for sake of simplicity, inner faces are given instead of PGS for right-handsides) and $P = \{N_{2341} \rightarrow A_{2541}\ N_{2645}\ B_{2346}\}$. The unique production of this grammar is the one represented in Figure 4 while the first of the three lexical rules is the one of Figure 3.

In order to describe the derivation process of a PGS in a PGG, we need to introduce the plane graph analogue of string grammar sentential forms. Contrary to the string case, where such forms are just strings containing both terminal and non-terminal symbols, a plane sentential form consists of a PGS together with a labeling function. The role of this function is to attached non-terminals to faces: if a face is not labeled with a non-terminal, then it cannot be rewritten and can thus be considered as a terminal face; on the other hand, a face that is labeled can be seen as a non-terminal face: rules rewriting the labeling non-terminal can be applied to it.

Definition 3.6. (Plane sentential form)
Let $\mathcal{G} = \langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$ be a plane graph grammar. A *plane sentential form* is a couple $\langle G, \mathcal{L} \rangle$ where $G = \langle X, E, F, o \rangle$ is a PGS and $\mathcal{L} : F \to \mathcal{N} \times X$ is a partial function such that $\mathcal{L}(f) = (N, a)$ implies that $|\texttt{nodes}(f)| = \texttt{rank}(N)$ and $a \in \texttt{nodes}(f)$.

Function $\mathcal{L}$ labels some faces with non-terminal symbols. It more precisely attaches the label to one distinguished vertex a of the face. This allows us to introduce some control during the application of a rule. Indeed, this trick is used to avoid all possible rotations of the right hand-side of the rule when this right hand-side is glued in the mother graph. We formally detail this below and a sequence of plane sentential forms is given in Figure 5.

3.1. Applying a lexical rule

A lexical rule $R : N_{a_1 \dots a_m} \to G_*$, with $G_* = \langle X_*, E_*, \{f_*, o_*\}, o_* \rangle$, is applicable to a sentential form $S = \langle G, \mathcal{L} \rangle$, with $G = \langle X, E, F, o \rangle$, if there exists a face $f = [a'_1 \dots a'_m] \in F$ such that $\mathcal{L}(f) = (N, a'_1)$. The resulting sentential form corresponds to S where $\mathcal{L}(f)$ is not defined anymore and f is replaced by the graph G_*, whose vertices are consistently renamed. More formally, *applying R to S following f*, consists in creating the sentential form $S' = \langle G', \mathcal{L}' \rangle$ with $G' = \langle X', E', F', o' \rangle$ such that

- $F' = F \setminus \{f\} \cup \hat{\phi}(f_*)$
- $\forall f' \in F, f' \neq f$: if $\mathcal{L}(f')$ is defined then $\mathcal{L}'(f') = \mathcal{L}(f')$
- $X' = X \cup \hat{\phi}(X_*)$
- $E' = E \cup \hat{\phi}(E_*)$

where ϕ is a injection from X_* to $\mathbb{N}$ defined as follow: if $\exists i, 1 \leq i \leq m, a = a_i$ then $\phi(a) = a'_i$; otherwise $\phi(a) = c \in \mathbb{N} \setminus X$.

Plane graph grammars associate no semantic to nodes label and generate plane graphs whose nodes have all distinct labels. This is explicit when a rule is applied: if a new node needs to be created, its label is picked in the set of natural numbers that are not used by a pre-existing node.

Example 3.7. The lexical rule $A_{1234} \to [123454]$ in $\mathcal{G}_1$ of Example 3.5 is applicable to the sentential S_3 of Figure 5. Once applied, the result is the sentential form S_4 of the same figure.

3.2. Applying a production rule

A production rule $R : N^0_{a^0_1 \dots a^0_m} \to N^1_{a^1_1 \dots a^1_{n_1}} \dots N^k_{a^k_1 \dots a^k_{n_k}}$ is applicable to a sentential form $S = \langle G, \mathcal{L} \rangle$, where $G = \langle X, E, F, o \rangle$, if there exists a face $f = [a'_1 \dots a'_m] \in F$ such that $\mathcal{L}(f) = (N, a'_1)$. *Applying R to S following f*, consists in creating the sentential form $S' = \langle G', \mathcal{L}' \rangle$ with $G' = \langle X', E', F', o' \rangle$ such that

- $F' = (F \setminus \{f\}) \bigcup_{1 \le i \le k} \{\hat{\phi}([a^i_1 \dots a^i_{n_i}])\}$
- $\forall f' \in F, f' \neq f$: if $\mathcal{L}(f')$ is defined then $\mathcal{L}'(f') = \mathcal{L}(f')$
- $\forall i, 1 \le i \le k$, $\mathcal{L}'(\hat{\phi}([a^i_1 \dots a^i_{n_i}])) = (N^1, \hat{\phi}(a^i_1))$
- $X' = X \cup_{1 \le i \le k} \texttt{nodes}(\hat{\phi}([a^i_1 \dots a^i_{n_i}]))$
- $E' = E \cup_{1 \le i \le k} \text{edges}(\hat{\phi}([a^i_1 \dots a^i_{n_i}]))$

where ϕ is a injection from $\{a^j_i : 1 \le j \le k, 1 \le i \le n_j\}$ to $\mathbb{N}$ defined as follow: if $\exists i, 1 \le i \le m, a = a^0_i$ then $\phi(a) = a'_i$; otherwise $\phi(a) = c \in \mathbb{N} \setminus X$.

Example 3.8. The production $N_{2341} \to A_{2541}\ N_{2645}\ B_{2346}$ in $\mathcal{G}_1$ of Example 3.5 is applicable to the sentential S_1 of Figure 5 (it is also applicable to S_2). When applying it, it creates the sentential form S_2 of the same figure.

Notice that applying a lexical rule or a production is replacing a face by a PGS whose outer face is the previous face, while the rest of the graph is unchanged. Renaming of the nodes (resp. the edges, faces) by the function $\hat{\phi}$ ensures consistency: no two nodes (resp. edges, faces) have the same label.

3.3. Representable languages

Given a plane graph grammar $\mathcal{G} = \langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$, we say that a plane graph $G = \langle X, E, F, o \rangle$ is generated by $\mathcal{G}$, or that $\mathcal{G}$ derives G, if there exists a sequence of sentential forms $S_1, \dots, S_n$ such that

- $S_1 = \langle G_1, \mathcal{L}_1 \rangle$ is an *initial sentential form*, that is, $G_1 = \langle X_1, E_1, F_1, o_1 \rangle$ with $F_1 = \{o^R, o\}$ and $o_1 = o$ and $\mathcal{L}_1$ is only defined for o^R: $\mathcal{L}_1(o^R) = (N, a)$, with $N \in \mathcal{A}$ and $a \in \text{nodes}(o^R)$,
- $\forall i, 1 \le i < n$: S_{i+1} is obtained from S_i by applying a rule of $\mathcal{G}$,
- $S_n = \langle G_n, \mathcal{L}_n \rangle$ with $G_n = G$ and $\forall f, \mathcal{L}_n(f)$ is not defined.

The *length of the derivation* is $n - 1$. An example of a derivation of length 4 is given in Fig. 5.

We will write $N \Rightarrow^*_{\mathcal{G}} G$, or simply $N \Rightarrow^* G$ if $\mathcal{G}$ is obvious from the context, when $G = \langle X, E, F, o \rangle$ is derivable with $\mathcal{G}$ using an initial sentential form S_1 where $\mathcal{L}_1$ is defined only on o^R and $\mathcal{L}_1(o^R) = (N, a)$, for some $a \in \texttt{nodes}(o^R)$. If the length of the derivation is n, we will write in the usual way $N \Rightarrow^n_{\mathcal{G}} G$.

The language represented by $\mathcal{G}$ is $L(\mathcal{G}) = \{G : \exists G', \exists N \in \mathcal{A}\ s.t.\ N \Rightarrow^*_{\mathcal{G}} G' \wedge G' \cong_p G\}$.

Note that any PGS can be represented by a PGG. Indeed, it suffices to have one production that from the outer face of the PGS generates the concatenation of the faces of PGS (without pendant edges) and

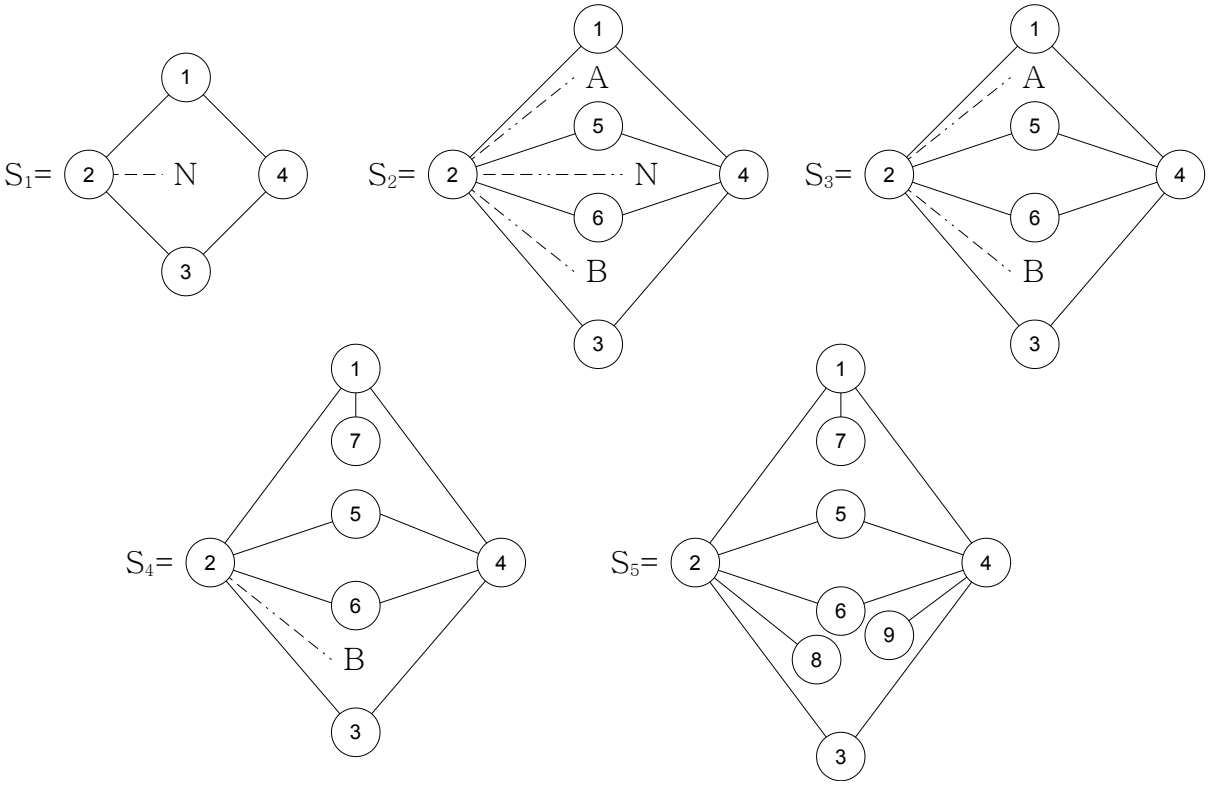

Figure 5. A graphical representation of an example of a derivation in the grammar $\mathcal{G}_1$ in Example 3.5. This derivation corresponds to the sequence of sentential forms S_1, S_2, S_3, S_4, S_5. The functions $\mathcal{L}_i$, $1 \leq i \leq 5$, are represented via dash lines.

labeled each face with a different non-terminal. Then all that is needed is a lexical rule for each non-terminal that generates the corresponding face. However, as it is discussed later on, not all plane graph languages can be represented by a PGG.

3.4. Plane Graph Grammars and Related Formalism's

Two main types of graph grammars have been investigated in the literature. The first one is the family of node replacement grammars [28] and relies on a mechanism that replaces one given node by a subgraph using gluing conditions; many gluing conditions were studied and yield several subfamilies of node-replacement grammars. This is clearly different from how the generative process occurs with plane graph grammars and thus a comparison between these two formalisms is difficult. However, a dual graph can be built from each planar graph [33], where each node corresponds to a face in the original graph, edges in one graph being edges in the other. From this standpoint, replacing a face in the primal graph (*i.e.* the original PGS) by a pattern corresponds to substituting a node of the dual graph by the corresponding dual subgraph. Embeddings[2] in node replacement grammars differ from embeddings in

[2]An embedding is the information about how to glue the new subgraph within the rest of the graph

plane graph grammars in that it relies on node label semantics.

Hyperedge replacement grammars [8] are another type of graph grammar formalisms that seems closer to the one introduced in this paper. Indeed, in these grammars, a hyperedge, *i.e.* a labeled entity that links up several nodes together, is replaced by a subgraph. One can imagine that plane graph grammars are a special kind of hyperedge replacement grammars, seeing a face labeled by a non-terminal as a hyperedge with the same label. However, these two elements are of different nature. For instance, there is no order on the outer nodes of a hyperedge and it is not possible to define a unique one. This implies that if the same non-terminal can be derived following two different ordering of the external nodes, then it needs to corresponds to two different hyperedges.

The main difference between these formalisms and plane graph grammars is the semantics they attach to the node label. Indeed, in both cases, the labels of the nodes of a glued subgraph are the ones the right hand-side of the rule, and if a rule is used several times, it generates the same node labels each time. As a consequence, the set of labels that can be found in a graph generated by such a grammar is bounded. Plane graph grammars are of different nature as each node has a unique label, which allows the closure of graph languages under isomorphism. The embedding mechanism (described in Section 3) does thus not rely on label value. Together with the polynomial testability of sub-isomorphism, this is a remarkable property from a learning standpoint: extracting and comparing patterns from a set of graphs is easy and informative. We shall see in detail how this is usefull in Section 5. If one want the nodes to carry semantic information, the formalism is easily modifiable with the adding of a labeling function, without modifying the core of its generative mechanism.

4. Properties of Plane Graph Grammars

4.1. Context-freeness property

A class of grammars that has the context-freeness property corresponds to a formalism where parts of a derivation that start from different non-terminals of a sentential form do not interfere with each other [7]. Intuitively, this is not the case of Plane Graph Grammars as the name of a node created during the derivation is linked with the name of pre-existing nodes. Hence, the distinct parts of a derivation cannot be treated in any order since the resulting PGS will not have the same node names.

However, this only affects the names of the nodes, and if one is interested in the structure of the generated PGS then each parts of the derivation can be done independently. In other words, the order in which the rules are applied will generate different PGS, but they will all be plane-isomorphic. We thus will be able to describe a derivation tree and a parsing procedure based on the CYK algorithm for string grammars.

The following theorem states that when a PGS is derived from a non-terminal, either it is given directly by a lexical rule, or it is the concatenation of plane graphs obtained from non-terminals that appear together on the right handside of a production whose left handside is the given non-terminal.

Theorem 4.1. (Context-freeness)
Let $\mathcal{G} = \langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$ be a PGG. Let $(N, r) \in \mathcal{N}$ and H a PGS. $N \Rightarrow^* H$ if and only if

- $N \to H$ in P_L,
- Or

1. $\exists N_x \to N^1_{x_1} \dots N^m_{x_m}$ in P and
2. there exist m PGS $H_1, \dots H_m$ such that
   - $\forall i, 1 \leq i \leq m$, $N^i \Rightarrow^* H_i$
   - $H \cong_p H_1 \diamond \dots \diamond H_m$

Proof:
[Sketch] $\Longrightarrow$
Let $N \Rightarrow^k H$. There thus exists a sequence of sentential forms $(G_1, \mathcal{L}_1), \dots, (G_{k+1}, \mathcal{L}_{k+1})$ such that $\forall j \leq k$, G_{j+1} is obtained from G_j by applying a rule of the grammar, $(G_1, \mathcal{L}_1)$ is an initial sentential form with $\mathcal{L}_1(o^R_{G_1}) = (N, a)$, for some $a \in \text{nodes}(o^R_{G_1})$, and $G_{k+1} = H$. If G_2 is obtained from G_1 using a lexical rule, then $k = 1$ and $N \to H$ is in P_L. Otherwise, the first rule is a production $N_x \to N^1_{x_1} \dots N^m_{x_z}$ and $G_2 = \langle X_{G_2}, E_{G_2}, \cup_{1 \leq i \leq m} \{[x_i]\} \cup [x^R]\}, [x^R]\rangle$, $\mathcal{L}_2([x_i]) = (N^i, \texttt{first}(x_i))$, $\forall i$. As a rule does not modify the pre-existing nodes and as it applies only to a specific face, the other steps of the derivation replace one of the faces $[x_i]$ by PGS: no step replacing $[x_i]$, or the set of faces previously derived from $[x_i]$, interferes with the steps that rewrite $[x_j]$, $1 \leq j, i \leq m$, $i \neq j$. Hence, for all $i \leq m$, the sequence of steps that recursively rewrites $[x_i]$ generates a PGS H_i and we have $H = H_1 \diamond \dots \diamond H_m$ (which is correctly defined since the outer face of H_i is $[x_i]^R$, $\forall i$). As each face $[x_i]$ of G_2 is labeled by the non-terminal N^i, we also have $N^i \Rightarrow^* H_i$.
$\Longleftarrow$
Suppose $\exists N_x \to N^1_{x_1} \dots N^m_{x_m}$ in P and there exist m PGS $H_1, \dots, H_m$ such that $N^i \Rightarrow^* H_i$, for all $1 \leq i \leq m$, and $H \cong_p H_1 \diamond \dots \diamond H_m$. Then the initial sentential form $(G_1, \mathcal{L}_1)$, with $G_1 = \langle \texttt{nodes}([x]), \text{edges}[x]), \{[x], [x^R]\}, [x^R]\rangle$ and $\mathcal{L}_1([x]) = (N, \texttt{first}(x))$, can be rewritten into the sentential form $(G_2, \mathcal{L}_2)$, with $G_2 = \langle X_{G_2}, E_{G_2}, \cup_{1 \leq i \leq m} \{[x_i]\} \cup \{[x^R]\}, [x^R]\rangle$ and $\mathcal{L}_2([x_i]) = (N^i, \texttt{first}(x_i))$, $\forall i$ (using the rule). There then exists a sequence of sentential forms $(G_2, \mathcal{L}_2)$, $\dots$, $(G_k, \mathcal{L}_k)$ such that $G_k = \langle X_{G_k}, E_{G_k}, F_{H_1} \cup_{2 \leq i \leq m} \{[x_i]\} \cup \{[x^R]\}, [x^R]\rangle$ and $\mathcal{L}_k$ is defined only on $[x_i]$, with $\mathcal{L}_m([x_i]) = (N^i, \texttt{first}(x_i))$, $\forall i, 2 \leq i \leq m$. As $N^2 \Rightarrow^* H_2$, there exist a sequence of sentential forms $(G_k, \mathcal{L}_k), \dots, (G_l, \mathcal{L}_l)$ that rewrite the face $[x_2]$ of G_2 into a subgraph H'_2 such that $H'_2 \cong_p H_2$ (the label of some internal nodes of H_2 can already exists in G_k and thus another label has to be chosen). Repeating the same reasoning for the other non-terminals of the rule, we obtain that for all i, $2 \leq i \leq m$ it exists a PGS $H'_i \cong_p H_i$ such that $N \Rightarrow^* H' = H_1 \diamond H'_2 \diamond \dots H'_m$. We have $H_1 \diamond H'_2 \cong_p H_1 \diamond H_2$, and for all $i < m$ $H_1 \diamond \dots \diamond H_i \diamond H'_{i+1} \cong_p H_1 \diamond \dots \diamond H_i \diamond H_{i+1}$. Notice that $H_1 \diamond \dots \diamond H_i$ and H'_{i+1} are concatenable folowing the identity function since $\texttt{outer}(H'_{i+1}) = \texttt{outer}(H_{i+1})$ by construction. This implies $H' \cong_p H$. □

One of the consequences of this result is that we can define a normal form for plane graph grammars, where the number of non-terminals on the right hand-side is exactly two. Indeed, given any production rule $N_x \to N^1_{x_1} \dots N^m_{x_m}$ with $m > 2$, one can replace it by two rules $N_x \to N'_{x'} N^m_{x_m}$ and $N'_{x'} \to N^1_{x_1} \dots N^{m-1}_{x_{m-1}}$, with $x' = \texttt{outer}(\{[x_i] : 1 \leq i \leq m-1\})$, which are both correctly defined rules. The set of graphs that N can derive is unchanged and the process can be recursively reproduced until only 2 non-terminals appear in each right hand-sides.

We will consider in the rest of this paper that all the PGG are in such a normal form.

4.2. A Parsing Algorithm

Theorem 4.1 provides a straightforward parsing algorithm whose pseudo-code is given in Algorithm 1.

Algorithm 1: Plane graph grammar parsing algorithm

Input: A PGG $\mathcal{G} = \langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$ in normal form and a PGS G
Output: TRUE if $G \in L(\mathcal{G})$, FALSE otherwise

```
foreach N ∈ A do
    if DERIVE(G, N, G) then
        return TRUE
return FALSE
```

Algorithm 2: DERIVE Procedure

Input: A PGG $\mathcal{G} = \langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$ in normal form, a non-terminal $N \in \mathcal{N}$ and a PGS G
Output: TRUE if $N \Rightarrow_{\mathcal{G}}^{*} G$, FALSE otherwise

```
if N → H ∈ P_L and H ≅_p G then
    return TRUE
foreach N_x → N_y^1 N_z^2 ∈ P do
    if ∃H_1, H_2 such that DERIVE(N^1, H_1) and DERIVE(N^2, H_2) and G ≅_p H_1 ⋄ H_2 then
        return TRUE
return FALSE
```

Proposition 4.2. For all PGS G, Algorithm 1 terminates and yields TRUE iff $G \in L(\mathcal{G})$.

Proof:
Algorithm 1 is nothing else than an algorithmic version of the context-freeness lemma. Since the size of the patterns H_1 and H_2 is smaller than the size of H, the algorithm must eventually terminates. □

Since there is only a finite set of lexical rules and of productions in $\mathcal{G}$, there are only polynomially many possibilities to consider in steps (1) and (3) of the DERIVE procedure. The plane-isomorphism can also be tested in polynomial time. Hence, there is a single point that may cause an exponential running time of the algorithm: the number of candidates H_1 and H_2 to test in step (4). Therefore we aim at finding a condition to impose on $L(\mathcal{G})$ that implies a polynomial upper bound on the number of such candidates. The following restriction is inspired by the k-separability, defined for hyperedge replacement grammars [26].

Definition 4.3. (Rank)
For k in $\mathbb{N}$, the k-rank of a PGS G is the number of patterns of G whose outer face contains k nodes. For every language L of PGS, $rank_L : \mathbb{N} \to \mathbb{L}$ is defined by

$$\text{rank}_L(n) = max_{1 \leq k \leq order(L,n)}\{k\text{-rank}(G) : G \in L \text{ and } |G| \leq n\}$$

where $order(L, n) = max\{|\texttt{outer}(S)| : S \text{ is a pattern of } G \in L \text{ and } |G| \leq n\}$.

Note that the order of a language L and an integer n is the maximal number of nodes of the outer face of a pattern of a plane graph in L whose size is at most n. We obviously have $order(L, n) \leq |G| \leq n$.

The idea behind the rank of a plane graph language is to link the size of the PGS of the languages with the number of their subgraphs that have an outer face of a given size. The aim is to tackle the combinatorial explosion that can occur when one is checking whether a rule can be applied. Indeed, to test if a rule $N_x \to N_y^1 N_z^2$ can be applied to derived a PGS G, Algorithm 2 needs to be recursively called on all decompositions of G into two patterns whose outer faces contain $|y|$ and $|z|$ nodes. In general, the number of such decompositions is exponential in the size of the grammar: this is the case for example of the PGS that correspond to a checkerboard.

Proposition 4.4. Let $\mathcal{G}$ be a PGG. If $\text{rank}_{L(\mathcal{G})}(n)$ is polynomial in n then Algorithm 1 can be implemented in time polynomial in the size of its input.

Proof:
Given a production $N_x \to N_y^1 N_z^2$ only patterns of $|y|$ (resp. $|z|$) nodes on the outer face can be derived from N^1 (resp. N^2). As $\text{rank}_{L(\mathcal{G})}(n)$ is polynomial, there are a polynomial number of patterns with $|y|$ (resp. $|z|$) external nodes on outer face. The number of candidates H_1, H_2 is thus polynomial if $G \in L(\mathcal{G})$. □

5. Learning substitutable plane graph languages

This section shows how plane graph grammars are good candidates for grammatical inference: their nice properties allow to extend works on distributional learning of string grammars. The simplest string class that has been proven learnable in this approach, is the one of substitutable context-free languages [4].

An earlier (and messier) version of this section has been published in the proceedings of the 11th International Conference on Grammatical Inference [12].

5.1. Substitutable plane graph languages

The core of the learning algorithm for this class is to observe the distribution of substrings into contexts and then to use the simple properties of substitutable languages to infer a correct grammar. And hence we first need a notion of context to transpose this work to plane graph languages.

Definition 5.1. (Plane context)
A plane context is a tuple $C = \langle X, E, F, h, o\rangle$ such that (1) $\langle X, E, F, o\rangle$ is a plane graph system and (2) $h \in F \setminus \{o\}$ is a distinguished face called the *hole* of context C and (3) h has no pendant edge.

The plane-isomorphism relation is extended to contexts in the obvious way: two contexts $C = \langle X, E, F, h, o\rangle$ and $C' = \langle X', E', F', h', o'\rangle$ are plane-isomorphic if $\langle X, E, F, o\rangle \cong_p \langle X', E', F', o'\rangle$ and the image of h by the bijection on the faces is h', i.e. $\xi(h) = h'$.

Let $S = \langle X, E, F, o\rangle$ and $S' = \langle X', E', F', o'\rangle$ be two PGS such that $X \cap X' = \emptyset$. Let $f \in F$ and $f' \in F'$ be two faces. Every 1-to-1 mapping $\phi : \texttt{nodes}(f) \to \texttt{nodes}(f')$ can be extended to the set of all vertices X as follows: $\hat{\phi} : X \to \texttt{nodes}(f') \cup X$ such that $\hat{\phi}(a) = \phi(a)$ if $a \in \texttt{nodes}(f)$ and $\hat{\phi}(a) = a$ otherwise. It can then be extended in the usual way to sets of nodes, faces, sets of faces, to PGS, and to plane contexts.

We can now define the *gluing* or *wrapping* operation.

Definition 5.2. (Gluing)
Let $C = \langle X, E, F, h, o\rangle$ be a context and $S = \langle X', E', F', o'\rangle$ be a PGS such that $X \cap X' = \emptyset$. Let ϕ be a bijective function from $\texttt{nodes}(o')$ to $\texttt{nodes}(h)$. The *gluing* of S into C following gluing function ϕ, denoted $C \odot_\phi S$, is the PGS $G = \langle X_G, E_G, F_G, o_G\rangle$ such that (1) $X_G = X \cup X' \setminus \texttt{nodes}(o')$, (2) $E_G = E \cup \hat{\phi}(E')$, (3) $F_G = (F \setminus \{h\}) \cup \hat{\phi}(F' \setminus \{o'\})$ and (4) $o_G = o$. Notice that S is a pattern of G.

We now need to define the notion of substitutability. Informally, two patterns of a given language are substitutable if the fact that they appear in the same context once, implies they always appear in the same context, glued in a similar way.

Definition 5.3. (Substitutability)
Two PGS $G = \langle X, E, F, o\rangle$ and $G' = \langle X', E', F', o'\rangle$ are *substitutable w.r.t.* a plane graph language L if whenever there exist two contexts C and C', $C \cong_p C'$, and two gluing functions ϕ and ϕ' such that $C \odot_\phi G$ is in L and $C' \odot_{\phi'} G'$ is in L, then for all contexts C'',

$$\big(\exists \phi_1 :\ C'' \odot_{\phi_1} G \in L\big) \iff \big(\exists \phi_2 :\ C'' \odot_{\phi_2} G' \in L\big)$$

where ϕ_1 and ϕ_2 are gluing functions such that, for all $a \in \texttt{nodes}(o)$, for all $b \in \texttt{nodes}(o')$, if $\phi(a) = \phi'(b)$ then $\phi_1(a) = \phi_2(b)$.

If G and G' are substitutable *w.r.t.* a language L, we will note $G \equiv_S^L G'$, or $G \equiv_S G'$ when there is no ambiguity.

The following lemma states that substitutability is not affected by plane-isomorphism.

Lemma 5.4. Let G, G' and G'' be PGS. If $G \cong_p G'$ and $G' \equiv_S G''$ then $G \equiv_S G''$.

Proof:
[Hint] Let χ be the 1-to-1 function that maps the vertices of G onto those of G' (as in the definition of plane isomorphism). Let C be a context such that there exists a gluing functions ϕ such that $C \odot_\phi G''$ in L. As $G' \equiv_S G''$ there exists ϕ' such that $C \odot_{\phi'} G'$ in L. By construction we have $C \odot_{\phi' o \chi} G \cong_p C \odot_{\phi'} G'$ and thus $C \odot_{\phi' o \chi} G$ is in L. □

We can then define substitutable plane graph languages:

Definition 5.5. (Substitutable languages)
A plane graph language L is substitutable iff all pairs of patterns that share a context are substitutable *w.r.t.* L.

The following technical lemma states that substitutability in a substitutable language is a congruence with respect to concatenation:

Lemma 5.6. Let L be a substitutable plane graph language and G_1, G_2, G_1', G_2' be 4 PGS s.t. both G_1 and G_2, and G_1' and G_2' are concatenable following the identity function. If $G_1 \equiv_S^L G_1'$ and $G_2' \equiv_S^L G_2'$ then $G_1 \diamond G_2 \equiv_S^L G_1' \diamond G_2'$

Proof:
[Sketch] We index each element of the definition of a PGS by its name. For instance X_{G_1} are the vertices of G_1, $F_{G_1 \diamond G_2}$ are the faces of $G_1 \diamond G_2$, $o_{G'_1 \diamond G'_2}$ is the outer face of $G'_1 \diamond G'_2$. Let C be a plane context such that there exists a gluing function ϕ: $C \odot_\phi G_1 \diamond G_2$ is in L. We suppose *w.l.o.g.* that the set of nodes of C is distinct of the set of nodes of the 4 PGS under consideration.

Let $C_{C \diamond \hat{\phi}(G_1)} = \langle X_C \cup \hat{\phi}(X_{G_1}) \cup \texttt{nodes}(\hat{\phi}(o_{G_2})), E_C \cup \hat{\phi}(E_{G_1}) \cup \texttt{edges}(\hat{\phi}(o_{G_2})), (F_C \setminus h) \cup \hat{\phi}(F_{G_1} \setminus o_{G_1}) \cup \hat{\phi}(o^R_{G_2}), \hat{\phi}(o^R_{G_2}), o_C \rangle$. Less formally, $C_{C \diamond \hat{\phi}(G_1)}$ is the context C with G_1 glued in it following ϕ (its hole is thus the face $\hat{\phi}(o^R_{G_2})$). Notice that $C_{C \diamond \hat{\phi}(G_1)}$ is correctly defined as its faces are contiguous since, by definition of the concatenation, there exists $e \in E_{G_1}$, $e \in \texttt{edges}(o_{G_1 \diamond G_2})$ and $\phi : \texttt{nodes}(o_{G_1 \diamond G_2}) \to \texttt{nodes}(h)$. It is easy to verify by using the definitions that $C_{C \diamond \hat{\phi}(G_1)} \odot_\phi G_2 = C \odot_\phi G_1 \diamond G_2$.

As $G_2 \equiv_S G'_2$ there exists ϕ' such that $C_{C \diamond \hat{\phi}(G_1)} \odot_{\phi'} G'_2 \in L$. But $C_{C \diamond \hat{\phi}(G_1)} \odot_{\phi'} G'_2$ is equal by construction to $C \odot_{id} \hat{\phi}(G_1) \diamond \hat{\phi}'(G'_2)$ where id is the identity function. Using the same kind of construction, we can construct the context $C_{C \diamond \hat{\phi}'(G'_2)}$ such that $C_{C \diamond \hat{\phi}'(G'_2)} \odot_{id} \hat{\phi}(G_1) = C \odot_{id} \hat{\phi}(G_1) \diamond \hat{\phi}'(G'_2) \in L$. As $G_1 \equiv_S G'_1$ and $\hat{\phi}(G_1) \cong_p G_1$, there exists ϕ'' such that $C_{C \diamond \hat{\phi}'(G'_2)} \odot_{\phi''} G'_1 \in L$. Again, this is equivalent to write $C \odot_{id} \hat{\phi''}(G'_1) \diamond \hat{\phi}'(G'_2)$, and as $\hat{\phi''}(G'_1) \diamond \hat{\phi}'(G'_2) \cong_p G'_1 \diamond G'_2$ then by Lemma 5.4 and the closure under plane-isomorphism of L, there exists χ: $C \odot_\chi G'_1 \diamond G'_2 \in L$. □

Finally, we can define a notion of congruence classes:

Definition 5.7. (Congruence classes)
Given a plane graph language L and a pattern G of a graph of the language, the congruence class of G in L, denoted $\lceil G \rceil_L$ (or simply $\lceil G \rceil$), is the set of all patterns substitutable with G in L: $\lceil G \rceil_L = \{G' : G' \equiv_L G\}$.

In a substitutable plane graph language, if two patterns appear once in the same context, they belong to the same congruence class.

5.2. The Learner

Our learning algorithm is described in Algorithm 3.

As we are interested in the class of substitutable plane graph languages, the learning algorithm has to deal with the distribution of contexts between subgraphs. To do so, it computes, from a finite set S of PGS of the language, the *observable congruence classes*: two PGS G and G' are in the same observable congruence class if there exist $k \geq 2$ and $G_1, \ldots, G_k$ such that $G = G_1$, $G' = G_k$, and $\forall i < k$, G_i and G_{i+1} appear at least once in the same context in the sample S. As in the case of strings, this computation can be done using a substitution graph [4] or by hashing from subgraphs to list of contexts (more complex and efficient structure can of course be designed to be able to decide whether two subgraphs appear in the same context). Notice that two PGS that are observed in the same congruence class are substitutable but that the converse is not true in general: two PGS that are substitutable may not be observed in the same congruence class in a given sample.

Given a set of contiguous inner faces F, we define $split(F)$ to be the set of couples (F_1, F_2) such that $F_1 \cup F_2 = F$ and G_{F_1} and G_{F_2} are concatenable following the identity, where for $i \in \{1, 2\}$, $G_{F_i} = \langle \texttt{nodes}(F_i), \texttt{edges}(F_i), F_i \cup \{\texttt{outer}(F_i)\}, \texttt{outer}(F_i) \rangle$. The function $number\_nodes(C)$ returns the number of nodes of the outer face of the graphs in the observable congruence class C.

Algorithm 3: Learning algorithm for substitutable plane graph languages

Input: A learning set of plane graph systems $LS = \{G_i\}_{i=1}^{n}$
Output: A plane graph grammar $\langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$

$CC \leftarrow compute\_observable\_congruence\_classes(LS)$;
$\mathcal{N} \leftarrow \emptyset$; $P_L \leftarrow \emptyset$; $P \leftarrow \emptyset$; $\mathcal{A} \leftarrow \emptyset$;
foreach *Observable congruence class* C_i *of* CC **do**
  $\mathcal{N} \leftarrow \mathcal{N} \cup \{(N^i, number\_nodes(C_i))\}$;
  foreach G *in* C_i **do**
    $\mathcal{N}^t(G) \leftarrow N^i$;
    if $G \in LS$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \cup \{N^i\}$
foreach $G = \langle X_G, E_G, F_G, [(a_1 \dots a_n)^R] \rangle$ *in* V **do**
  if $|F_G| = 2$ **then**
    $P_L \leftarrow P_L \cup \{\mathcal{N}^t(G)_{a_1 \dots a_n} \rightarrow G\}$;
  else
    foreach $(F_1, F_2) \in split(F_G \setminus \{o_G\})$ **do**
      $P \leftarrow P \cup \{\mathcal{N}^t(G_*)_{a_1 \dots a_m} \rightarrow \mathcal{N}^t(G_1)_{b_1 \dots b_n}\, \mathcal{N}^t(G_2)_{c_1 \dots c_p}\}$ where
      $[b_1 \dots b_n] = \mathtt{outer}(F_1)$, $[c_1 \dots c_p] = \mathtt{outer}(F_2)$, $b_1 = b$, $c_1 = c$ and
      $\forall i : G_i = \langle X_i, E_i, F_i \cup \{\mathtt{outer}(F_i)\}, \mathtt{outer}(F_i) \rangle$;
return $\langle N, P_L, P, \mathcal{A} \rangle$

5.3. Identification in the limit

We now define our learning criterion. This is set-driven identification from positive text, with a polynomial bound on computation:

Definition 5.8. (Set-driven identification in polynomial time)
A representation class $\mathbb{R}$ is set-driven identifiable from positive data in polynomial time iff there exist a polynomial p and an algorithm $\mathcal{A}$ such that:

1. Given a positive sample LS of size m, $\mathcal{A}$ returns a representation $R \in \mathbb{R}$ in time $p(m)$;
2. For each representation $R \in \mathbb{R}$ there exists a characteristic set CS such that if $CS \subseteq LS$, $\mathcal{A}$ returns a representation R' such that $L(R) = L(R')$.

Note that the size of a set of plane graphs LS is defined as $|LS| = \sum_{G \in LS} |G|$.

The initial definition of this learning paradigm requires the size of the characteristic sample to be polynomial in the size of the target representation [19]. However, this definition, initially designed for the learning of regular string languages, is already unsuitable as a model for context free string grammars [34], so one cannot expect this requirements to be fulfilled in the case of graph grammars. As it is beyond the scope of this paper to attempt to resolve this difficulty, we shall thus adopt this approach in this paper: for a complete discussion, the Reader is referred to this book chapter [11].

5.3.1. Time complexity

The number of patterns (and thus of contexts) that can be generated from a given PGS can be exponential in the size of that PGS (it is the case for instance of the plane graph corresponding to a grid, like a chest board). So the size of observable congruence classes is in general exponential in the size of the learning sample. This is a well-known problem while using graph grammar formalisms as it is related to the one of having an efficient parsing algorithm. However, the requirement of having a language of polynomial rank, needed for efficient parsing (see Section 4.2) implies that the number of patterns to considerate is polynomial in the size of the learning sample. Therefore the number of elements that have to be taken into account to compute the congruence classes is polynomial.

To compute these observable congruence classes, we also need to compare all pairs of contexts to decide if they are plane isomorphic. This can be done in polynomial time in the size of the contexts [20]. For the same reason, testing if two PGS are plane isomorphic can be done in polynomial time and thus so is the construction of the congruence classes.

All other steps of Algorithm 3 are polynomial in the size of the observable congruence classes.

5.3.2. Proof the hypothesis is not too large

The following lemma states that patterns in the sample can be generated by the output grammar.

Lemma 5.9. If $G = \langle X, E, F, o\rangle$ is a subgraph of a sample LS, then there exists a plane graph G' such that $\mathcal{N}^t(G) \Rightarrow^* G'$ and $G \cong_p G'$.

Proof:
[Sketch] The proof can be done by induction on the number of faces of the graph. if $|F| = 2$, then by the construction of the grammar there is a lexical rule $\mathcal{N}^t(G)_{a_n \dots a_1} \to G$ with $[a_1 \dots a_n] = o$. Suppose the property holds for graphs with $|F| = k \geq 2$ faces. Let F_1 and F_2 be two sets of contiguous faces such that $F_1 \cap F_2 = \emptyset$ and $F_1 \cup F_2 = F_G \setminus \{o_G\}$. Let G_1 (resp. G_2) be the PGS whose inner faces are F_1 (resp. F_2). We have $G_1 \diamond G_2 = G$. G_1 and G_2 are also subgraphs of LS by definition and, by construction of the grammar, there exists a rule $\mathcal{N}^t(G)_{a_1 \dots a_m} \to \mathcal{N}^t(G_1)_{b_1 \dots b_n} \mathcal{N}^t(G_2)_{c_1 \dots c_n}$ with $[a_1 \dots a_m] = o$, $[b_1 \dots b_n] = \texttt{outer}(F_1) (= o_{G_1})$ and $[c_1 \dots c_p] = \texttt{outer}(F_2) (= o_{G_2})$.

This rule can be applied to the sentential form $\langle G', \mathcal{L}'\rangle$, with $G' = \langle X', E', \{o^R, o\}, o\rangle$, $\mathcal{L}'(o^R) = (\mathcal{N}^t(G), a_1)$. It gives the sentential form $\langle G'', \mathcal{L}''\rangle$, with $G'' = \langle X'', E'', \{\hat{\phi}(\texttt{outer}(F_1)), \hat{\phi}(\texttt{outer}(F_2)), o\}, o\rangle$, $\mathcal{L}''(\texttt{outer}(F_1)) = (\mathcal{N}^t(G_1), b_1)$ and $\mathcal{L}''(\texttt{outer}(F_2)) = (\mathcal{N}^t(G_2), c_1)$ By the inductive hypothesis there exist G'_1 and G'_2 such that $\mathcal{N}^t(G_1) \Rightarrow^* G'_1$, $\mathcal{N}^t(G_2) \Rightarrow^* G'_2$, $G_1 \cong_p G'_1$ and $G_2 \cong_p G'_2$. G'_1 and G'_2 might not be concatenable, as they can have inner nodes that share the same label. However, w.l.o.g. one can change the labels of one of the graph, for instance G'_1, in order to obtain a PGS G''_1 that is plane-isomorphic to G_1 and concatenable to G'_2. Thus we have $\mathcal{N}^t(G) \Rightarrow^* G''_1 \diamond G'_2$ and, by Lemma 2.10, $G \cong_p G'_1 \diamond G'_2$. □

Lemma 5.10. For all subgraphs G of a learning sample LS, for all PGS G', if $\mathcal{N}^t(G) \Rightarrow^* G'$ then G and G' are substitutable.

Proof:
[Hint] Let $G = \langle X, E, F, o\rangle$. As the lemma holds for $G' \cong_p G$, we restrict ourselves to the case

$G' \ncong_p G$. By induction on the length of the derivation k. If $k = 1$, then it means that a lexical production $\mathcal{N}^t(G)_{a_1 \dots a_n} \to G''$ is applied and that $G'' \cong_p G'$. By the construction of the lexical rules, it means that G'' is a subgraph of LS that appears in the congruence class than G and thus G and G'' are substitutable. Lemma 5.4 implies that $G' \equiv_S G$.

Suppose this is true for all derivations of length strictly less than k and let G' be a PGS obtained from $\mathcal{N}^t(G)$ using k derivation steps. It means that there exists a sequence of sentential form $S_1, \dots, S_k$, such that that $\forall i$, S_i is derived from S_{i-1}, $S_i = \langle G_i, \mathcal{L}_i \rangle$ with $G_1 = \langle X_1, E_1, \{o^R, o\}, o \rangle$, $\mathcal{L}_1(o^R) = (\mathcal{N}^t(G), a)$ for some $a \in \texttt{nodes}(o)$, and $G_k = G'$, $\mathcal{L}_k$ being undefined for all faces of G_k. S_2 is obtained from S_1 applying a rule $\mathcal{N}^t(G)_{a_1 \dots a_m} \to \mathcal{N}^t(G_{F_1})_{b_1 \dots b_n} \mathcal{N}^t(G_{F_2})_{c_1 \dots b_p}$, where $\texttt{outer}(F_1) = [b_1 \dots b_n]$ and $\texttt{outer}(F_2) = [c_1 \dots c_p]$, $G_{F_i} = \langle X_{F_i}, E_{F_i}, F_i \cup \texttt{outer}(F_i), \texttt{outer}(F_i) \rangle$, for $i \in \{1, 2\}$. By construction, there exists G_* in the same observable congruence class of G such that $G_* = G_{F_1} \diamond G_{F_2}$ and thus $G_{F_1} \diamond G_{F_2} \equiv_S G$. There exist G'_{F_1} and G'_{F_2} such that $\mathcal{N}^t(G_{F_1}) \Rightarrow^* G'_{F_1}$, $\mathcal{N}^t(G_{F_2}) \Rightarrow^* G'_{F_2}$ and $G_k = G'_{F_1} \diamond G'_{F_2}$. By recursion, $G'_{F_1} \equiv_S G_{F_1}$ and $G'_{F_2} \equiv_S G_{F_2}$. As Lemma 5.6 holds, we have $G'_{F_1} \diamond G'_{F_2} \equiv_S G_{F_1} \diamond G_{F_2}$ and thus $G_k \equiv_S G$. □

Theorem 5.11. For all samples of a language L, the output $\mathcal{G}$ of Algorithm 3 is such that $L(\mathcal{G}) \subseteq L$.

Proof:
Let $G \in L(\mathcal{G})$. Then there exists a plane graph G' in the learning sample and a plane graph G'' such that $\mathcal{N}^t(G') \in \mathcal{A}$, $\mathcal{N}^t(G') \Rightarrow^* G''$ and $G'' \cong_p G$. Lemma 5.10 states that G'' and G' are substitutable and thus $G \equiv_S^L G'$. As G' is an element of L, $G \in L$. □

5.3.3. Proof the hypothesis is large enough

To prove that the hypothesis is large enough, we need to define a characteristic set, *i.e.* a subset of the target language L_* which ensures that the output $\mathcal{G}$ of the algorithm is such that $L(\mathcal{G}) = L_*$.

Construction of a characteristic sample. Let $\mathcal{G}_* = \langle \mathcal{N}_*, P_{L*}, P_*, \mathcal{A}_* \rangle$ be a target grammar. We will assume without loss of generality, that $\mathcal{G}_*$ is reduced, that is to say for every non-terminal N, (1) there exists a derivation that starts from an axiom and labels at least one face with N, and (2) a PGS without any non-terminal labeling a face can be derived from a sentential form where one face is labeled by N. We are going to define a set $CS(\mathcal{G}_*)$ of plane graphs of L_*, such that Algorithm 3 will identify L_* from any superset of $CS(\mathcal{G}_*)$.

Given a non-terminal N^k, we define $C(N^k)$ to be one of the smallest context $\langle X_{G_k}, E_{G_k}, F_{G_k}, h_k, o_{G_k} \rangle$ such that there exists a sequence of sentential forms $\langle G_1, \mathcal{L}_1 \rangle, \dots, \langle G_k, \mathcal{L}_k \rangle$ with $\langle G_1, \mathcal{L}_1 \rangle$ being an initial sentential form such that $F_{G_1} = [o_{G_k}, o^R_{G_k}]$, $\mathcal{L}_1(o^R_{G_k}) = (N^i, a_1)$, $N^i \in \mathcal{A}_*$, and $\forall i, 1 \leq i < k$, $\langle G_{i+1}, \mathcal{L}_{i+1} \rangle$ is obtained from $\langle G_i, \mathcal{L}_i \rangle$ by applying a rule of $\mathcal{G}_*$, $\mathcal{L}_k(h_k) = (N^k, a_k)$ for some $a_k \in \texttt{nodes}(h_k)$, $\mathcal{L}_k$ is undefined on other faces.

We also define $G(N^k)$ to be one of the smallest PGS such that $N^k \Rightarrow^*_{\mathcal{G}_*} G(N^k)$.

We can now define the characteristic set $CS(\mathcal{G}_*)$. For each production $N^i_x \to N^j_y N^k_z$ in P_*, we add to $CS(\mathcal{G}_*)$ the PGS $C \odot_\phi \hat{\chi}(G_1) \diamond \hat{\chi}(G_2)$ where $\phi : \texttt{nodes}([x]) \to \texttt{nodes}(h)$ is a bijective function, $C = C(N^i), G_1 = G(N^j), G_2 = G(N^k)$ and $\chi : \texttt{nodes}(o_{G_1}) \cup \texttt{nodes}(o_{G_2}) \to \texttt{nodes}([y]) \cup \texttt{nodes}([z])$ is a bijective function such that $\hat{\chi}(o_{G_1}) = [y]$ and $\hat{\chi}(o_{G_2}) = [z]$. For each lexical rule $N^i_x \to G$ in P_{L*}

we add to $CS(\mathcal{G}_*)$ the PGS $C \odot_\phi G$ where $\phi : \texttt{nodes}([x]) \to \texttt{nodes}(h)$ is a bijective function and $C = C(N^i)$.

The cardinality of this set is at most $|P_*| + |P_{L*}|$ which is clearly polynomially bounded. In general the cardinality of the set will not polynomially bound the size of the sample, as it is already the case for string context-free grammars (see [4] for a detailed discussion). However, notice that if there exists a polynomial-sized structurally complete sample – that is to say a sample where for each production rule there is at least one plane graph that can be generated by using it [9] – then the size of our characteristic set is polynomial. In addition, one can show that the size of this characteristic set is polynomial in the size of the target grammar and of its *thickness*, following the refinement of the learning paradigm suggested by Ryo Yoshinaka [34].

Convergence. We must show that for any substitutable plane graph grammar $\mathcal{G}_*$, if the sample LS contains the characteristic sample $CS(\mathcal{G}_*)$, then $L(\mathcal{G}) = L(\mathcal{G}_*)$ where $\mathcal{G} = \langle \mathcal{N}, P_L, P, \mathcal{A} \rangle$ is the inferred grammar.

Lemma 5.12. If $N \Rightarrow^*_{\mathcal{G}_*} G$ then there exists a subgraph G' of the learning sample and a plane graph G'' such that $N \Rightarrow^*_{\mathcal{G}_*} G', \mathcal{N}^t(G') \Rightarrow^*_{\mathcal{G}} G''$ and $G'' \cong_p G$.

Proof:
[Sketch] By recursion on the number of derivation steps k in $\mathcal{G}_*$. If $k = 1$ then there exists $N \to G'$ in P_{L*}, $G' \cong_p G$. By construction of the characteristic sample, G' is a subgraph of LS and thus $\mathcal{N}^t(G') \to G'$ is in P_L.

Suppose it is true for all derivations of size less than $k > 1$. There exists a sequence of sentential forms $\langle G_1, \mathcal{L}_1 \rangle, \ldots, \langle G_k, \mathcal{L}_k \rangle$ such that $\langle G_1, \mathcal{L}_1 \rangle$ is an initial sentential form with $\mathcal{L}(f_1) = (N, a)$, S_{i+1} is obtained from S_i by using a rule of $\mathcal{G}_*$, $G_k = G$ and $\mathcal{L}_k$ is not defined for any face. Let $N_x \to N^i_y N^j_z$ be the rule applied to S_1 to obtain S_2. By construction, there exist G_1 and G_2, $N^i \Rightarrow^*_{\mathcal{G}_*} G_1$, $N^j \Rightarrow^*_{\mathcal{G}_*} G_2$, and $G_1 \diamond G_2 = G$.

By recursion, there exist two subgraphs of LS, G'_1 and G'_2, and two PGS G''_1 and G''_2 such that $N^i \Rightarrow^*_{\mathcal{G}_*} G'_1$, $N^j \Rightarrow^*_{\mathcal{G}_*} G'_2$, $\mathcal{N}^t(G'_1) \Rightarrow^*_{\mathcal{G}} G''_1$, $\mathcal{N}^t(G'_2) \Rightarrow^*_{\mathcal{G}} G''_2$ and $G''_1 \cong_p G_1$, $G''_2 \cong_p G_2$. Notice that this implies there exists a renaming function ϕ on the vertices of the external faces of G''_1 and G''_2 such that $\hat{\phi}(G''_1)$ and $\hat{\phi}(G''_2)$ are concatenable and $\hat{\phi}(G''_1) \diamond \hat{\phi}(G''_2) \cong_p G$ (Lemma 2.10).

By construction of the characteristic sample, there exist two subgraphs G'''_1 and G'''_2 of LS such that $G'''_1 \diamond G'''_2$ is a subgraph of LS, $G'''_1 \cong_p G(N^i)$ and $G'''_2 \cong_p G(N^j)$. As $L(\mathcal{G}_*)$ is a substitutable language, we have $G'''_1 \equiv_S G'_1$ and $G'''_2 \equiv_S G'_2$. Thus G'_1 and G'''_1 appear in the same component and thus correspond to the same non-terminal (and similarly for G'''_2 and G'_2). As there is a rule $\mathcal{N}^t(G'''_1 \diamond G'''_2)_x \to \mathcal{N}^t(G'''_1)_y \mathcal{N}^t(G'''_2)_z$ in P, we have $\mathcal{N}^t(G'''_1 \diamond G'''_2) \Rightarrow^*_{\mathcal{G}} \hat{\phi}(G''_1) \diamond \hat{\phi}(G''_2)$. □

Theorem 5.13. Let $\mathcal{G}_*$ be the target plane graph grammar corresponding to a substitutable plane graph language. Algorithm 3 returns a grammar $\mathcal{G}$ from any sample containing $CS(\mathcal{G}_*)$ such that $L(\mathcal{G}) = L(\mathcal{G}_*)$.

Proof:
If $G \in L(\mathcal{G}_*)$ then there exists $N \in \mathcal{A}_*$ such that $N \Rightarrow^*_{\mathcal{G}_*} G$. By Lemma 5.12, it implies that there exists a subgraph G' of the learning sample and a plane graph G'' such that $G' \in L(\mathcal{G}_*)$, $\mathcal{N}^t(G') \Rightarrow^*_{\mathcal{G}} G''$ and $G'' \cong_p G$. By construction of the grammar, $\mathcal{N}^t(G') \in \mathcal{A}$ and thus $G \in L(\mathcal{G})$. Therefore $L(\mathcal{G}_*) \subseteq L(\mathcal{G})$. Due to Theorem 5.11, we have $L(\mathcal{G}) = L(\mathcal{G}_*)$. □

6. Discussion

In addition to substitutability, other restrictions on the learned class have been done, explicitly or not. First, the grammar formalism implies that the number of nodes of the outer face of any generated PGS has to be bounded: otherwise an infinite number of axioms would be needed. Then, the requirement of having a polynomial rank, that is used both for efficient parsing and for the polynomial computation time of the learning algorithm, is clearly restrictive.

Despite all these issues, this paper describes, to our knowledge, one of the first positive formal learning result for a non-trivial class of graph grammars. The work on substitutable string languages [4] has been the starting point of several positive learning results on more complex classes, and similar developments are likely to be tractable for plane graph languages. It seems to be the case for instance of the extension to contexts with several holes [35] using multiple context-free grammars (that are a context-sensitive formalism with a polynomial time parsing algorithm). It might also be possible to adapt the learning algorithm in a way that allows a learning result in the PAC paradigm [31]. Finally, due to the interest of planar graphs in image processing [30], it is likely that the learning of plane graph grammars, and more generally grammatical inference techniques, could be used to tackle image classification tasks.

A. Proof of Theorem 2.3

In this section, we prove that if a PGS $S = \langle X, E, F, o, \mathcal{D}\rangle$ is valid, then it denotes a plane graph. Let us remark that $\langle X, E\rangle$ is a connected planar simple graph by assumption. Moreover, Condition (1) of Def. 2.2 ensures that the boundary of every face is made of well-defined edges. Typically, no face can be described with a boundary like $[xx\ldots]$, and if a boundary like $[xy\ldots]$ appears, then $\{x, y\}$ is an edge. Notice that sets X and E are redundant in a valid PGS, since we can deduce them from the boundaries of the faces.

Now, in order to prove Theorem 2.3, we are going to show that a valid PGS actually defines a 2D-combinatorial map $M = \langle D, \alpha, \beta\rangle$. Then, thanks to Condition (4) of Def. 2.2, we know that the genus of this map is null, so it can be embedded on a sphere with no crossing edges. Finally, the fact that we distinguish face o as the external face allows us to continuously deform the sphere into a plane and finally get a plane graph (remember that a plane is isomorphic to a sphere minus a point).

Let us first define the set of darts: $D = \{\vec{xy} : \exists f \in F, \exists u \in X^*, \mathcal{D}(f) = [xyu]\}$. We now define $\beta : D \to D$ as follows: $\beta(\vec{xy}) = \vec{yx}$. We claim that β is well-defined and obviously an involution over D. Indeed, if $\vec{xy} \in D$, then there exists a face $f \in F$ such that $\mathcal{D}(f) = [xyu]$ for some $u \in X^*$. By Condition (1) of Def.2.2, we deduce that pair $\{x, y\}$ is an edge. So by Condition (2), we deduce that there exists a face $f' \in F$ such that $\mathcal{D}(f') = [yxv]$ for some $v \in X^*$, thus $\vec{yx} \in D$.

The definition of $\alpha : D \to D$ is a bit more intricate. Consider a dart $\vec{xy} \in D$. Then there exists an unique face $f \in F$ such that $\mathcal{D}(f) = [xyu]$ for some $u \in X^*$. We can assume without loss of generality that $|u| > 0$. Otherwise, face f is bounded by a single edge, and as the graph is connected, it is actually reduced to a single edge. Therefore, there exists $z \in X$ such that $\mathcal{D}(f) = [xyzu]$ for some $u \in X^*$, and we set $\alpha\vec{xy} = \vec{yz}$.

FINIR : montrer que α est une permutation

1. For all $f \in F$ and $x, y \in X$ and $u \in X^*$, if $\mathcal{D}(f) = [xyu]$ then $\{x, y\} \in E$;

2. For all $e = \{x, y\} \in E$, there exist a unique face $f \in F$ such that $\mathcal{D}(f) = [xyu]$, and an unique face $f' \in F$ such that $\mathcal{D}(f') = [yxu]$, for some $u, u' \in X^*$;

3. For all $f \in F$ and $x, y, z, z' \in X$ and $u, u' \in X^*$, if $\mathcal{D}(f) = [xyzuxyz'u']$ or $\mathcal{D}(f) = [zxyuz'xyu']$ then $z = z'$;

4. Euler's formula holds, that is, $|X| - |E| + |F| = 2$.

References

[1] Bailly, R., Denis, F., Rabusseau, G.: Recognizable Series on Hypergraphs, *Language and Automata Theory and Applications - 9th International Conference, LATA 2015, Nice, France, March 2-6, 2015, Proceedings*, LNCS 8977, 2015.

[2] Brijder, R., Blockeel, H.: On the inference of non-confluent NLC graph grammars, *Journal of Logic and Computation*, 2011.

[3] Clark, A.: Towards General Algorithms for Grammatical Inference, in: *Proc. of ALT*, Springer, 2010, 11–30, Invited Paper.

[4] Clark, A., Eyraud, R.: Polynomial Identification in the Limit of Substitutable Context-free Languages, *J. Machine Learning Research*, **8**, 2007, 1725–1745.

[5] Cook, D. J., Holder, L. B.: Graph-Based Data Mining, *IEEE Intelligent Systems*, **15**(2), 2000, 32–41, ISSN 1541-1672.

[6] CostaFlorencio, C.: Identification of Hyperedge-Replacement graph grammars, *MLG'09, Poster Session*, Leuven, Belgium, 2009.

[7] Courcelle, B.: An axiomatic definition of context-free rewriting and its application to NLC graph grammars, *TCS*, **55**, 1987, 141–181.

[8] Drewes, F., Kreowski, H.-J., Habel, A.: Hyperedge Replacement Graph Grammars, in: *Handbook of Graph Grammars*, [29], 1997.

[9] Dupont, P., Miclet, L., Vidal, E.: What Is the Search Space of the Regular Inference?, *Proc. of ICGI*, 1994.

[10] Engelfriet, J.: Tree automata and tree grammars, 1975.

[11] Eyraud, R., Heinz, J., Yoshinaka, R.: Efficiency in the Identification in the Limit Learning Paradigm, in: *Topics in Grammatical Inference* (J. Heinz, J. M. Sempere, Eds.), Springer-Verlag, 2016, In press.

[12] Eyraud, R., Janodet, J.-C., Oates, T.: Learning Substitutable Binary Plane Graph Grammars, *Proc. ICGI'12*, JMLR W&CP 21, 2012.

[13] Fáry, I.: On straight line representation of planar graphs, *Acta Univ Szeged. Sect. Sci. Math*, **11**, 1948, 229–233.

[14] Fusy, E.: *Combinatoire des graphes planaires et applications algorithmiques (in english)*, Ph.D. Thesis, Ecole Polytechnique - ParisTech, 2007.

[15] Garey, M. R., Johnson, D. S.: *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, 1979.

[16] Gibbons, A.: *Algorithmic graph theory*, Cambridge University Press, 1985.

[17] Gold, E.: Language Identification in the Limit, *Information and Control*, **10**(5), 1967, 447–474.

[18] Harchaoui, Z., Bach, F. R.: Image Classification with Segmentation Graph Kernels, *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[19] de la Higuera, C.: Characteristic Sets for Polynomial Grammatical Inference, *Machine Learning Journal*, **27**, 1997, 125–138.

[20] de la Higuera, C., Janodet, J.-C., Samuel, E., Damiand, G., Solnon, C.: Polynomial Algorithms for Open Plane Graph and Subgraph Isomorphisms, *Theoretical Computer Science*, **498**, 2013, 76–99.

[21] Huan, J., Wang, W., Prins, J.: Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism, *ICDM*, 2003.

[22] Jeltsch, E., Kreowski, H.-K.: Grammatical Inference Based on Hyperedge Replacement, *Proc. Graph Grammars and their Application to Computer Science*, 1991.

[23] Jonyer, I., Holder, L. B., Cook, D. J.: MDL-Based Context-Free Graph Grammar Induction, *International Journal of Artificial Intelligence Tools*, **13**, 2003, 65–79.

[24] Kasprzik, A., Yoshinaka, R.: Distributional Learning of Simple Context-Free Tree Grammars, *ALT*, 2011.

[25] Kukluk, J., Holder, L., Cook, D.: Inference of Edge Replacement Graph Grammars, *International Journal on Artificial Intelligence Tools*, **17**(3), 2008, 539–554.

[26] Lautemann, C.: The Complexity of Graph Languages Generated by Hyperedge Replacement, *Acta Inf.*, **27**(5), 1989, 399–421.

[27] Matsuda, T., Motoda, H., Washio, T.: Graph-based induction and its applications, *Advanced Engineering Informatics*, 2002, 135–143.

[28] Nagl, M.: Formal languages of labelled graphs, *Computing*, **16**, 1976, 113–137.

[29] Rozenberg, G., Ehrig, H.: *Handbook of Graph Grammars and Computing by Graph Transformation*, vol. 1–3, World Scientific, 1997.

[30] Samuel, E., de la Higuera, C., Janodet, J.-C.: Extracting Plane Graphs from Images, *Proc. SSPR/SPR'10*, LNCS 6218, 2010.

[31] Shibata, C., Yoshinaka, R.: PAC Learning of Some Subclasses of Context-Free Grammars with Basic Distributional Properties, *Proceedings of 24th International Conference on Algorithmic Learning Theory*, 2013.

[32] Vishwanathan, S., Schraudolph, N., Kondor, R. I., Borgwardt, K.: Graph Kernels, *Journal of Machine Learning Research*, **11**, 2010, 1201–1242.

[33] Whitney, H.: Non-separable and planar graphs, *Proc. Nat. Acad. Sci. U.S.A.*, **17**(2), 1931, 125–127, ISSN 0027-8424, JFM:57.0727.05.

[34] Yoshinaka, R.: Identification in the Limit of (k,l)-Substitutable Context-Free Languages, *ICGI*, 2008.

[35] Yoshinaka, R.: Efficient learning of multiple context-free languages with multidimensional substitutability from positive data., *Theor. Comput. Sci.*, **412**(19), 2011, 1821–1831.