A Detachment Algorithm for Inferring a Graph from Path Frequency

Hiroshi Nagamochi*

Dept. of Applied Mathematics and Physics
Graduate School of Informatics
Kyoto University
nag@amp.i.kyoto-u.ac.jp

Abstract: Inferring graphs from path frequency has been studied as an important problem which has a potential application to drug design and elucidation of chemical structures. Given a multiple set g of strings of labels with length at most K, the problem asks to find a vertex-labeled graph G that attains a one-to-one correspondence between g and the occurrences of labels along all paths of length at most K in G. In this paper, we prove that the problem with $K = 1$ can be formulated as a problem of finding a loopless and connected detachment, based on which an efficient algorithm for solving the problem is derived. Our algorithm also solves the problem with an additional constraint such that every vertex in an inferred graph is required to have a specified degree.

Keywords: connectivity, detachment algorithm, graph inference, multigraphs

1 Introduction

Kernel methods have been popular tools for designing classifiers such as support vector machines [7]. In kernel methods, a set of objects (or data) in the target problem are mapped to a space, called a *feature space*, where an object is transformed into a vector with real coordinates, and a kernel function is defined as an inner product of two feature vectors. Recently, a feature space has been used in a new approach in order to design or choose a desired (possibly unknown) object [3, 4]. As in kernel methods, given objects mapped to points in a feature space, this approach searches a point y in the feature space using a suitable objective function, and then maps this point back to an object in the input space, where the object mapped back is called a *pre-image* of the point. Given a mapping ϕ from an input space to a feature space and a point y in the feature space, the pre-image problem asks to find an object x with $y = \phi(x)$ in the input space. The pre-image problem for graphs is very important because it has a potential application to drug design and elucidation of chemical structures from mass/NMR spectra data, and has been studied by several researchers [4, 14].

However, the pre-image problem for graphs has not been studied from a computational point of view until recently Akutsu and Fukagawa [1] started an investigation of the theoretical

*Technical report 2007-008, February 6, 2007. This research was partially supported by the Scientific Grant-in-Aid from Ministry of Education, Culture, Sports, Science and Technology of Japan.

aspect of the problem of inferring graphs from *path frequency*. In this case, a feature vector g is a multiple set of strings of labels with length at most K which represents path frequency (i.e., the numbers of occurrences of vertex-labeled paths of length at most K). Given a feature vector g, they considered the problem of finding a vertex-labeled graph G that attains a one-to-one correspondence between g and the set of sequences of labels along all paths of length at most K in G (where the degrees of vertices in an inferred graph are not specified). They proved that the problem of inferring planar graphs is NP-hard even for $K = 4$ [2]. Recently it was shown that the problem of inferring graphs is NP-hard even for $K = 2$ [13]. For the problem of inferring a tree, Akutsu and Fukagawa [1] gave dynamic programming algorithms that runs in polynomial time in n when K and the number of labels are bounded by constants, where n denotes the size of an output graph. Akutsu and Fukagawa [2] extended their dynamic programming algorithms to the problem of inferring a graph in a restricted class of outerplanar graphs. However, the time complexity of these dynamic programming algorithms is a polynomial of n whose exponent is exponential in K and the number of labels.

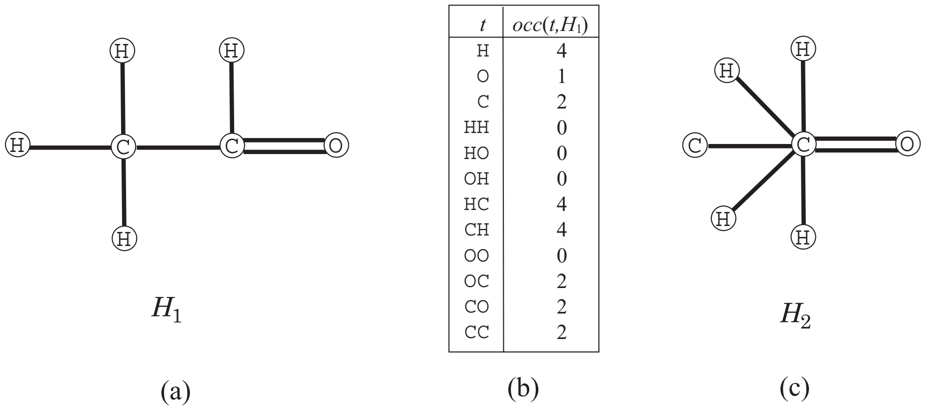
In this paper, we consider the problem of inferring a multigraph from a feature vector g of path frequency with $K = 1$. We show that the problem can be formulated as a problem of finding loopless and connected *detachments* of graphs, and give an efficient algorithm based on matroid intersection in discrete optimization. Our algorithm can test whether there exists a solution to a given vector g or not in $O(\min\{|g|2^{|g|}, n^{3.5} + m\})$ time, where $|g|$ is the number of nonzero entries in an input vector g and n and m are the numbers of vertices and edges of a multigraph to be constructed. For a feasible instance, the algorithm can deliver a solution in $O(n^{3.5} + m)$ time. In particular, for testing the feasibility of g, the running time is constant if the number of labels is bounded by a constant. We next introduce a graph inference problem with an additional constraint such that every vertex is required to have a specified degree, and prove that, for $K = 1$, the graph inference problem with such a degree specification can be solved in $O(\min\{n + |g|2^{|g|}, n^{3.5} + m\} + mn^2)$ time. We also consider an important variant of this problem, which asks to find a vertex-labeled graph G whose path frequency contains a given g as its subset, and prove that the variant can be solved in the same time complexity.

The paper is organized as follows. Section 2 introduces problems of inferring graphs from path frequency. Section 3 reviews some mathematical notions on graphs and gives efficient algorithms for finding loopless and connected detachments. Section 4 then shows that the above-mentioned graph inference problems with $K = 1$ can be solved efficiently. Section 5 makes some concluding remarks.

2 Graph Inference Problem

This section defines problems of inferring graphs from path frequency.

A graph is called a *multigraph* if it is allowed to have multiple edges and self-loops; otherwise it is called *simple*. A multigraph having no self-loops is called *loopless*. A multigraph G with a vertex set V and an edge set E is denoted by (V, E). The vertex set and edge set of a given multigraph G may be denoted by $V(G)$ and $E(G)$, respectively. An edge e with end vertices u and v is denoted by $\{u, v\}$. An alternating sequence $\pi = (v_0, e_1, v_1, e_2, v_2, \ldots, e_h, v_h)$ of vertices



t	$occ(t,H_1)$
H	4
O	1
C	2
HH	0
HO	0
OH	0
HC	4
CH	4
OO	0
OC	2
CO	2
CC	2

(a) (b) (c)

Figure 1: (a) A (Σ, ρ)-labeled multigraph H_1, where $\Sigma = \{\mathtt{H}, \mathtt{O}, \mathtt{C}\}$, $\rho(x) = 1$ if $\ell(x) = \mathtt{H}$, $\rho(x) = 2$ if $\ell(x) = \mathtt{O}$, and $\rho(x) = 4$ if $\ell(x) = \mathtt{C}$, respectively; (b) $occ(t, H_1) = occ(t, H_2)$ for all sequences $t \in \Sigma^{\leq 2}$; (c) A Σ-labeled multigraph H_2.

and edges in G is called a *walk* if, for each $i = 1, \ldots, h$, edge e_i joins vertices v_{i-1} and v_i, and its length is defined by h (note that the same vertex and the same edge may appear more than once in a walk).

Let $\mathbf{Z}_+$ denote the set of nonnegative integers. Let Σ be a set of labels, Σ^k be the set of all sequences of k labels in Σ, and $\Sigma^{\leq k} = \cup_{1 \leq j \leq k} \Sigma^j$. Let $\mathcal{F}_k(\Sigma)$ denote the set of all nonnegative integer vectors g whose coordinate are indexed by $t \in \Sigma^{\leq k+1}$ (i.e., g is a mapping from $\Sigma^{\leq k+1}$ to $\mathbf{Z}_+$). A vector $g \in \mathcal{F}_k(\Sigma)$ may be called a *feature vector*. Let $g(t)$ denote the entry of $g \in \mathcal{F}_k(\Sigma)$ indexed by $t \in \Sigma^{\leq k+1}$.

A multigraph H is called Σ*-labeled* if each vertex $v \in V(H)$ is labeled by a label $\ell(v) \in \Sigma$. Let H be a loopless Σ-labeled multigraph. For a walk $\pi = (v_0, e_1, v_1, e_2, v_2, \ldots, e_h, v_h)$ in H, let $\ell(\pi)$ denote the sequence of the vertex labels in π, i.e., $\ell(\pi) = \ell(v_0)\ell(v_1)\ldots\ell(v_h)$. For a label sequence t over Σ, let $occ(t, H)$ denote the number of walks π such that $\ell(\pi) = t$. The *feature vector* $f_K(H)$ *of level* K in H is a vector $g \in \mathcal{F}_K(\Sigma)$ such that $g(t) = occ(t, H)$ for all $t \in \Sigma^{\leq K+1}$, i.e.,

$$f_K(H) = (occ(t, H))_{t \in \Sigma^{\leq K+1}}.$$

For example, Figure 1(a) shows a loopless Σ-labeled multigraph H_1, where $\Sigma = \{\mathtt{H}, \mathtt{O}, \mathtt{C}\}$, Figure 1(b) gives $occ(t, H_1)$ for all $t \in \Sigma^{\leq 2}$, and we have $f_1(H_1) = (4, 1, 2, 0, 0, 0, 4, 4, 0, 2, 2, 2)$. Figure 1(c) shows a different loopless Σ-labeled multigraph H_2 such that $f_1(H_2) = f_1(H_1)$.

For a given feature vector $g \in \mathcal{F}_K(\Sigma)$, there may be no Σ-labeled multigraph H with $f_K(H) = g$. Different Σ-labeled graphs H and H' may have the same feature vector $f_K(H) = f_K(H') = g$, as observed in Fig. 1.

Akutsu and Fukagawa [1] formulated the following important problem.

Graph Inference from Path Frequency (GIPF) Given a feature vector $g \in \mathcal{F}_K(\Sigma)$, output a loopless and connected Σ-labeled multigraph H with $f_K(H) = g$. If there does not exist such H, then output "no solution."

The following complexity results on GIPF are known. Akutsu and Fukagawa [1] proved that GIPF with $K = O(\log n)$ is NP-hard even if both $|\Sigma|$ and the maximum degree Δ of inferred graphs are bounded from above by constants. Akutsu and Fukagawa [2] also showed that GIPF with $K = 3$ is NP-hard even if inferred graphs are restricted to be tree, and that GIPF with $K = 4$ is NP-hard even if inferred graphs are restricted to be planar, where $|\Sigma|$ and Δ are unbounded. Recently it was shown [13] that GIPF with $K = 2$ is NP-hard if $|\Sigma|$ and Δ are unbounded. Akutsu and Fukagawa [2] proposed a dynamic programming algorithm that solves GIPF for trees in polynomial time in the size of an output graph H when K and $|\Sigma|$ are constant. They also extended their dynamic programming algorithm to the problem of inferring an outerplanar graph such that both the degrees and the length of facial cycles are bounded by constants. However, the time complexity of these dynamic programming algorithms is exponential in $|\Sigma|$ even if $K = 1$.

It is important to consider a problem of inferring a graph that meets some degree constraint in some applications such as chemical graphs where every vertex with the same label has a specified valence. In this paper, we define a degree-constrained graph inference problem as follows. A *valence-sequence* ρ is a function $\rho : V(H) \to \mathbf{Z}_+$. A Σ-labeled multigraph H is called (Σ, ρ)*-labeled* if

$$deg(x; H) = \rho(x) \text{ for each } x \in V(H).$$

(Note that ρ specifies the degree of each vertex, and possibly $\rho(x) \neq \rho(y)$ even if $\ell(x) = \ell(y)$.)

Figure 1(a) shows a (Σ, ρ)-labeled multigraph H_1 for valence-sequence ρ such that $\rho(x) = 1$ if $\ell(x) = \mathtt{H}$, $\rho(x) = 2$ if $\ell(x) = \mathtt{O}$, and $\rho(x) = 4$ if $\ell(x) = \mathtt{C}$, respectively. Note that multigraph H_2 in Fig. 1(c) is not (Σ, ρ)-labeled.

Graph Inference from Path Frequency and Label Valence (GIFV) Given a feature vector $g \in \mathcal{F}_K(\Sigma)$ and a valence-sequence ρ, output a loopless and connected (Σ, ρ)-labeled multigraph H with $f_K(H) = g$. If there does not exist such H, then output "no solution."

No complexity result on GIFV with $K \geq 1$ is known. Note that GIFV with $K = 0$ is a trivial problem which asks whether a given set $V(H)$ of labeled vertices has an enough number of degrees to form a connected graph, and can be easily solved by checking if the sum of degrees of the labeled vertices is not less than the number of the labeled vertices minus 1. However, the problem of enumerating all solutions to an instance of GIFV with $K = 0$ contains *isomer enumeration*, which is one of the important research issue in chemical graph theory, and Pólya [18] gave the most powerful enumeration method to the problem.

In this paper, we also consider the situation where a given feature vector g represents only partial information on graphs that we want to infer. To handle this case, we modify the problem setting of GIFV as follows.

Graph Inference from Partial Path Frequency and Label Valence (GIPPFV) Given a feature vector $g \in \mathcal{F}_K(\Sigma)$ and a valence-sequence ρ, output a loopless and connected (Σ, ρ)-labeled multigraph H with $(occ(t, H))_{t \in \Sigma} = (g(t))_{t \in \Sigma}$ and $f_K(H) \geq g$. If there does not exist such H, then output "no solution."

In this paper, we show that, for $K = 1$, all of GIPF, GIFV and GIPPFV can be solved efficiently. Before deriving these results, we will prepare some mathematical tools in the next section.

3 Detachments in Multigraphs

This section shows some results on graph algorithms and combinatorial optimization.

3.1 Multigraphs and Matroids

A singleton set $\{x\}$ may be simply written as x. Let $G = (V, E)$ be a multigraph which may have self-loops. For two subsets $X, Y \subset V$ (not necessarily disjoint), $E(X, Y; G)$ denotes the set of edges e joining a vertex in X and a vertex in Y (i.e., $e = \{u, v\}$ satisfies $u \in X$ and $v \in Y$), and $d(X, Y; G)$ denotes $|E(X, Y; G)|$. Note that $E(X, Y; G)$ includes all self-loops $\{u, u\}$ with $u \in X \cap Y$ if any. We may write $E(X, V-X; G)$ and $d(X, V-X; G)$ as $E(X; G)$ and $d(X; G)$, respectively. Note that $d(u, v; G)$ is the number of multiple edges with end vertices u and v in G. Then the number of edges incident to a vertex v is given by $deg(v; G) = d(v; G) + d(v, v; G)$. The *degree* of a vertex v is defined to be $deg(v; G) = d(v; G) + 2d(v, v; G)$. A vertex v is called *isolated* if $deg(v; G) = 0$. For a multigraph $G = (V, E)$ and a subset $X \subseteq E$ (resp., $X \subseteq V$), let $G - X$ denotes the multigraph obtained by removing the edges in X (resp., the vertices in X together with the incident edges) from G.

Let $c(G)$ denote the number of components in a multigraph G. Removing k edges from G increases the number of components at most by k. Hence we have:

Lemma 1 *For a multigraph $G = (V, E)$ and a subset $E' \subseteq E$, $c(G - E') \leq c(G) + |E'|$.* ▮

We here review the definition and some important property of matroids (see [6, 15] for more on matroid theory). For a finite set S, let $\mathcal{I}$ be a family of subsets of S. System $(S, \mathcal{I})$ is called a *matroid* if it satisfies three conditions (i) $\emptyset \in \mathcal{I}$, (ii) If $I \in \mathcal{I}$, then any subset I' of I also belongs to $\mathcal{I}$, and (iii) For any $I_1, I_2 \in \mathcal{I}$ with $|I_1| < |I_2|$, there is an element $e \in I_2 - I_1$ such that $I_1 \cup \{e\} \in \mathcal{I}$. For a set $I \in \mathcal{I}$ of a matroid $\mathcal{M} = (S, \mathcal{I})$ and an element $e \in S - I$ with $I \cup \{e\} \notin \mathcal{I}$, the set of elements $e' \in I \cup \{e\}$ such that $I \cup \{e\} - e' \in \mathcal{I}$ is called a *circuit* and is denoted by $C(I, e)$. The *rank function* η of a matroid $\mathcal{M} = (S, \mathcal{I})$ is defined as a function $\eta : 2^S \to \mathbf{Z}_+$ such that $\eta(S')$ is the maximum cardinality $|I|$ of a member $I \in \mathcal{I}$ with $I \subseteq S'$. We here review two examples of matroids. For a partition of S into k disjoint subsets $S_1, S_2, \ldots, S_k$ and k nonnegative integers $b_1, b_2, \ldots, b_k$, family $\mathcal{I} = \{I \subseteq S \mid |I \cap S_i| \leq b_i \ , i = 1, 2, \ldots, k\}$ gives a matroid, called a *partition matroid.* Another example is a *graphic matroid* $(S, \mathcal{I})$, which is defined from a graph G with $E(G) = S$ so that $\mathcal{I} = \{I \subseteq E(G) \mid \ I \text{ contains no cycle in } G\}$.

Given two matroids $\mathcal{M}_1 = (S, \mathcal{I}_1)$ and $\mathcal{M}_1 = (S, \mathcal{I}_2)$ on the same set S, finding a maximum common member $I^* \in \mathcal{I}_1 \cap \mathcal{I}_2$ is known as the matroid intersection problem. It is not difficult to observe that $|I| \leq \eta_1(S') + \eta_2(S - S')$ holds for every $I \in \mathcal{I}_1 \cap \mathcal{I}_2$ and $S' \subseteq S$, where η_i is the rank function of $\mathcal{M}_i$, $i = 1, 2$. Edmonds has proven the following min-max theory.

Theorem 2 [9] *For two matroids $\mathcal{M}_i = (S, \mathcal{I}_i)$ with rank function η_i, $i = 1, 2$, it holds*

$$\max\{|I| \mid I \in \mathcal{I}_1 \cap \mathcal{I}_2\} = \min\{\eta_1(S') + \eta_2(S - S') \mid S' \subseteq S\}.$$

∎

It is known that, for two matroids on the same set S, a maximum common member $I^* \in \mathcal{I}_1 \cap \mathcal{I}_2$ can be found by an $O(|I^*|^{1.5}|S|)$ oracle time algorithm [8].

3.2 Detachments

Let G be a multigraph which may have self-loops. A *detachment* H of G is a multigraph with $E(H) = E(G)$ such that $V(H)$ can be partitioned into $|V(G)|$ subsets W_v, $v \in V(G)$ in such a way that G is obtained from H by contracting each subset W_v into a single vertex v.

Given a function $r : V(G) \to \mathbf{Z}_+$, a detachment $H = (\cup_{v \in V(G)} W_v, E(G))$ of G is called an *r-detachment* of G if $|W_v| = r(v)$, $v \in V(G)$, where we denote $W_v = \{v^1, v^2, \ldots, v^{r(v)}\}$. In other words, H is obtained from G by splitting each vertex $v \in V(G)$ into $r(v)$ copies of v, where each edge $\{u, v\} \in E(G)$ joins some vertices $u^i \in W_u$ and $v^j \in W_v$. Hence an r-detachment H of G is not unique in general. A self-loop $\{u, u\}$ in G may be mapped to a self-loop $\{u^i, u^i\}$ or a non-loop edge $\{u^i, u^j\}$ in a detachment H of G. Note that $d(W_u, W_v; H) = d(u, v; G)$ holds for all $u, v \in V(G)$.

For example, an r-detachment of graph G_g in Fig. 2(a) is shown in Fig. 2(c), where $r(\mathtt{H}) = 4$, $r(\mathtt{O}) = 1$ and $r(\mathtt{C}) = 2$.

For a function $r : V(G) \to \mathbf{Z}_+$, an *$r$-degree specification* is a set ρ of vectors $\rho(v) = (\rho^v_1, \rho^v_2, \ldots, \rho^v_{r(v)})$, $v \in V(G)$ such that

$$\sum_{1 \le i \le r(v)} \rho^v_i = deg(v; G).$$

An r-detachment H of G is called a *ρ-detachment* if each $v \in V$ satisfies

$$deg(v^i; H) = \rho^v_i \text{ for all } v^i \in W_v = \{v^1, v^2, \ldots, v^{r(v)}\}.$$

For a subset $X \subseteq V(G)$, $r(X)$ denotes $\sum_{v \in X} r(v)$.

Nash-Williams [17] obtained the following characterization of connected r-detachments of G which are allowed to have self-loops.

Theorem 3 [17] *Let $G = (V, E)$ be a multigraph and $r : V \to \mathbf{Z}_+$. Then there exists a connected r-detachment H of G if and only if*

$$r(X) + c(G - X) - d(X, V; G) \le 1 \text{ for every nonempty subset } X \subseteq V. \quad (1)$$

Furthermore, if G has a connected r-detachment then there exists a connected ρ-detachment H_ρ of G for every r-degree specification ρ. ∎

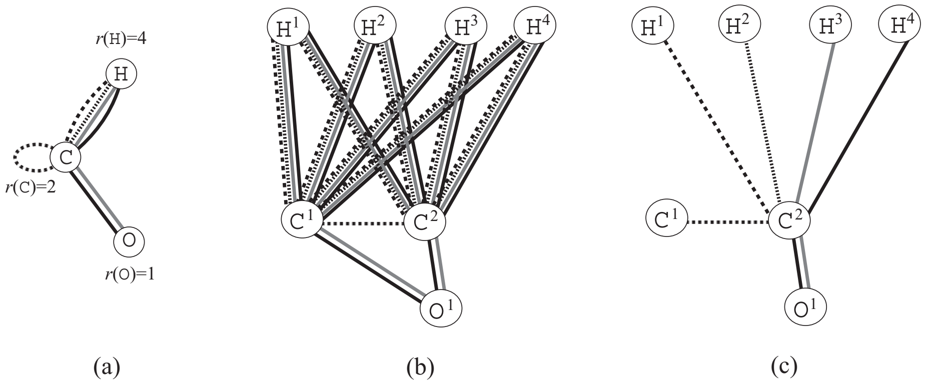


Figure 2: (a) A multigraph G_g obtained from the vector $g \in \mathcal{F}_1(\{\mathtt{H}, \mathtt{O}, \mathtt{C}\})$ in Fig. 1(c); (b) An r-expansion $\hat{H}(G_g)$ of G_g, where $r(\mathtt{H}) = 4$, $r(\mathtt{O}) = 1$ and $r(\mathtt{C}) = 2$; (c) An r-detachment of G_g.

The theorem does not characterize the necessary and sufficient condition for a given multigraph G to have a *loopless* connected r-detachment or ρ-detachment H. Note that G may not have a loopless and connected ρ-detachment even if it has a loopless and connected r-detachment. For example, consider multigraph $G = (\{u, v\}, E = \{e_1 = \{u, v\}, e_2 = \{v, v\}, e_3 = \{v, v\})$, which has a loopless and connected r-detachment H for $r(u) = 1$ and $r(v) = 2$, but cannot have a loopless and connected ρ-detachment H_ρ for $\rho(u) = (1)$ and $\rho(v) = (4, 1)$.

3.3 Loopless Detachments

In this subsection, we give an efficient algorithm for computing a loopless and connected ρ-detachment of a given multigraph. For this, we derive the necessary and sufficient conditions for a given multigraph to have *loopless* connected r- and ρ-detachments as follows.

Theorem 4 *Let $G = (V, E)$ be a multigraph and $r : V \to \mathbf{Z}_+$. Then:*

(i) *There exists a loopless and connected r-detachment H of G if and only if* (1) *holds and $r(v) \geq 2$ for each self-loop $\{v, v\} \in E$.*

(ii) *Whether* (1) *holds or not can be tested in $O(\min\{r(V)^{3.5} + |E|, \ r(V)^{1.5}|E|r_{max}^2\})$ time, and a multigraph H in* (i) *if any can be constructed in $O(\min\{r(V)^{3.5} + |E|, \ r(V)^{1.5}|E|r_{max}^2\})$ time, where $r_{max} = \max_{v \in V} r(v)$.* ∎

Theorem 5 *Let $G = (V, E)$ be a multigraph, $r : V \to \mathbf{Z}_+$, and ρ be an r-degree specification. Then:*

(i) *G has a loopless and connected ρ-detachment H_ρ if and only if it hold* (1) *and*

$$1 \leq \rho_i^v \leq d(v; G) + d(v, v; G) \text{ for all } v^i \in W_v \text{ and } v \in V. \tag{2}$$

(ii) *Assume that* (2) *holds. Given a loopless and connected r-detachment H of G, a loopless and connected ρ-detachment H_ρ can be constructed in $O(|E| \ \min\{r(V)^2, \ |E|r_{max}^2\})$ time.* ∎

3.3.1 Proof of Theorem 4

We first prove Theorem 4. First consider the necessity of Theorem 4(i). If $r(v) = 1$ for some self-loop $\{v, v\} \in E$, then clearly G cannot have a loopless r-detachment. Assume that there is a connected r-detachment $H = (\cup_{v\in V} W_v, E(G))$ of $G = (V, E)$. Let X be an arbitrary nonempty subset of V. For $X_H = \cup_{v\in X} W_v$ and $E' = E(X_H, V(H); H)$, each vertex in X_H has no incident edge in graph $H - E'$, and $c(H - E') = |X_H| + c(H - X_H) \geq r(X) + c(G - X)$ holds. Since $1 = c(H) \geq c(H-E') - |E'|$ holds by Lemma 1 and $|E'| = d(X_H, V(H); H) = d(X, V; G)$ holds, we have $1 \geq r(X) + c(G - X) - d(X, V; G)$, which implies the necessity of Theorem 4(i).

We now show the sufficiency of Theorem 4(i). Given a multigraph $G = (V, E)$ and a function r in Theorem 4, we define an *r-expansion* as a multigraph $\hat{H}(G) = (W = \cup_{v\in V} W_v, F)$ such that its vertex set W is the union of $|V|$ disjoint vertex subsets $W_v = \{v^1, v^2, \ldots, v^{r(v)}\}$, $v \in V$ and its edge set F is the union of $|E|$ disjoint edge subsets F_e, $e \in E$ defined by

$$\begin{aligned} F_e &= \{\{u^i, v^j\} \mid u^i \in W_u,\ v^j \in W_v\} \quad \text{if } e = \{u, v\} \in E\ (u \neq v),\\ F_e &= \{\{u^i, u^j\} \mid u^i, u^j \in W_u,\ i \neq j\} \quad \text{if } e = \{u, u\} \in E. \end{aligned}$$

Note that $|W| = r(V)$ and $|F| = \sum_{\{u,v\}\in E: u\neq v} r(u)r(v) + \sum_{\{u,u\}\in E} r(u)(r(u)-1)/2 = O(|E|r_{max}^2)$ hold, and that the resulting multigraph (W, F) is loopless since $|W_u| = r(u) \geq 2$ holds for any self-loop $e = \{u, u\} \in E$ by the assumption on r. Any subset $F' \subseteq F$ such that $|F' \cap F_e| = 1$, $e \in F$ can be viewed as a loopless r-detachment (W, F') of G.

We here introduce a partition matroid $\mathcal{M}_1 = (F, \mathcal{I}_1)$ with

$$\mathcal{I}_1 = \{I \subseteq F \mid |I \cap F_e| \leq 1\ \forall e \in E\}$$

and the graphic matroid $\mathcal{M}_2 = (F, \mathcal{I}_2)$ of $\hat{H}(G)$, i.e.,

$$\mathcal{I}_2 = \{I \subseteq F \mid I \text{ contains no cycle in } \hat{H}(G)\}.$$

Observe that, for any loopless r-detachment (W, F') of G, its maximal forest $F'' \subseteq F'$ (i.e., a maximal subset of F' having no cycle) satisfies

$$c((W, F'')) = |W| - |F''| \text{ and } F'' \in \mathcal{I}_1 \cap \mathcal{I}_2.$$

In particular, $c((W, F'')) = |W| - |F''| = 1$ if (W, F') is connected. Therefore, it suffices to show that $\mathcal{I}_1 \cap \mathcal{I}_2$ contains a subset I^* with $|I^*| = |W| - 1$ if (1) holds, since this implies that $c((W, I^*)) = |W| - |I^*| = 1$ and that a loopless and connected r-detachment (W, F') is obtained from I^* by adding $|E| - |I^*|$ more edges choosing an arbitrary edge $e' \in F_e$ for each $e \in E$ with $I^* \cap F_e = \emptyset$ so that $|F' \cap F_e| = 1$ holds for all $e \in E$.

By Theorem 2, the maximum cardinality $|I|$ of a member $I \in \mathcal{I}_1 \cap \mathcal{I}_2$ is equal to $\min\{\eta_1(F') + \eta_2(F - F') \mid F' \subseteq F\}$, where η_i is the rank function of $\mathcal{M}_i$, $i = 1, 2$. We prove the next property, which is a variant of a result in [17] to prove Theorem 3.

Lemma 6 *For rank function η_i of matroid $\mathcal{M}_i$, $i = 1, 2$, if* (1) *holds, then*

$$\eta_1(F') + \eta_2(F - F') \geq r(V) - 1 \textit{ for every subset } F' \subseteq F. \tag{3}$$

PROOF: Let F' be an arbitrary subset of F. Let

$$E' = \{e \in E \mid F' \cap F_e \neq \emptyset\}.$$

Then we have $\eta_1(F') = |E'|$. To show $\eta_2(F - F') \geq r(V) - 1 - |E'|$, we consider the graph $G - E'$. Let X be the set of isolated vertices in $G - E'$, and $G_1, G_2, \ldots, G_p$ be the remaining components of $G - E'$, where each G_i contains at least one edge. Then $E' \subseteq E(X, V; G)$ holds, and $|E''| = |E'| - d(X, V; G)$ holds for $E'' = E' - E(X, V; G)$. Note that

$$p = c((G - X) - E'') \text{ holds.}$$

By Lemma 1, we have $c((G - X) - E'') \leq c(G - X) + |E''|$. Hence $p \leq c(G - X) + |E''| = c(G - X) + |E'| - d(X, V; G)$. Since $r(X) + c(G - X) - d(X, V; G) \leq 1$ by (1), we have

$$p \leq |E'| + 1 - r(X). \tag{4}$$

Now consider the r-expansion $\hat{H}(G_i)$ of each G_i. The union $\hat{H}(G_1) \cup \hat{H}(G_2) \cup \cdots \cup \hat{H}(G_p)$ of these graphs is a subgraph of $\hat{H}(G)$ satisfying

$$E(\hat{H}(G_1) \cup \hat{H}(G_2) \cup \cdots \cup \hat{H}(G_p)) \subseteq F - F',$$

which implies

$$\begin{aligned}
\eta_2(F - F') &\geq \eta_2(E(\hat{H}(G_1) \cup \hat{H}(G_2) \cup \cdots \cup \hat{H}(G_p))) \\
&= (|V(\hat{H}(G_1))|-1) + (|V(\hat{H}(G_2))|-1) + \cdots + (|V(\hat{H}(G_p))|-1) \\
&= r(V(G_1)) + \cdots + r(V(G_p)) - p = r(V) - r(X) - p \\
&\geq r(V) - 1 - |E'| \text{ (by (4))},
\end{aligned}$$

as required. ∎

Given a multigraph G and a function r, we compute a member $I^* \in \mathcal{I}_1 \cap \mathcal{I}_2$ with the maximum cardinality $|I^*|$. If (1) holds, then $|I^*| = \min\{\eta_1(F') + \eta_2(F - F') \mid F' \subseteq F\} \geq r(V) - 1$ must hold by this lemma and Theorem 2, and G admits a loopless and connected r-detachment. This shows the sufficiency of (1), proving Theorem 4(i).

To test whether (1) holds or not, we compute a maximum common member $I^* \in \mathcal{I}_1 \cap \mathcal{I}_2$. Before resorting the matroid intersection algorithm to find such I^*, we reduce the size of graph $\hat{H}(G)$ as follows. Since any member in $\mathcal{I}_2$ contains no cycle, we do not need to have multiple edges in an r-expansion $\hat{H}(G) = (W, F)$ of G. Let $\tilde{H} = (W, \tilde{F})$ be the simple graph obtained from $\hat{H}(G)$ by replacing multiple edges between two vertices with a single edge. Formally, the edge set $\tilde{F}$ is given by $\cup_{u,v \in V} F_{uv}$ such that

$$\begin{aligned}
&F_{uv} = F_{uv} = \{\{u^i, v^j\} \mid u^i \in W_u,\ v^j \in W_v\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if } E \text{ contains an edge } \{u, v\}\ (u \neq v), \\
&F_{uu} = \{\{u^i, u^j\} \mid u^i, u^j \in W_u,\ i \neq j\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if } E \text{ contains a self-loop } \{u, u\}, \\
&F_{uv} = F_{vu} = \emptyset \qquad\qquad\qquad\quad \text{otherwise.}
\end{aligned}$$

Then graph $\tilde{H} = (W, \tilde{F})$ is simple, and contains $|\tilde{F}| = O(\min\{|W|^2, |F|\}) = O(\min\{r(V)^2, |E|r_{max}^2\})$ edges. In $\tilde{H}$, we define a partition matroid

$$\tilde{\mathcal{M}}_1 = (\tilde{F}, \tilde{\mathcal{I}}_1 = \{I \subseteq \tilde{F} \mid |I \cap F_{uv}| \leq d(u, v; G) \ \forall u, v \in V\})$$

and a graphic matroid

$$\tilde{\mathcal{M}}_2 = (\tilde{F}, \tilde{\mathcal{I}}_2 = \{I \subseteq \tilde{F} \mid I \text{ contains no cycle in } \tilde{H}\}).$$

By definition, we easily see that finding a maximum member in $\tilde{\mathcal{I}}_1 \cap \tilde{\mathcal{I}}_2$ is equivalent to that in $\mathcal{I}_1 \cap \mathcal{I}_2$.

For the current common subset $I \in \tilde{\mathcal{I}}_1 \cap \tilde{\mathcal{I}}_2$, the matroid intersection algorithm in [8] constructs a bipartite digraph $B_I = (I \cup \{s_1, s_2\}, \tilde{F} - I, \mathcal{E})$ as follows. It has two vertex sets $I \cup \{s_1, s_2\}$ and $\tilde{F} - I$, where $s_1, s_2 \notin \tilde{F}$ are new elements, and an edge set $\mathcal{E}$ which consists of four types of sets of directed edges (i) $\{(s_1, y) \mid y \in \tilde{F} - I, \ I \cup \{y\} \in \mathcal{I}_2\}$, (ii) $\{(y, s_2) \mid y \in \tilde{F} - I, \ I \cup \{y\} \in \mathcal{I}_1\}$, (iii) $\{(x, y) \mid x \in I, \ y \in F - I, \ x \in C_2(I, y)\}$, and (iv) $\{(y, x) \mid x \in I, \ y \in F - I, \ x \in C_1(I, y)\}$, where $C_i(I, e)$ denotes the circuit in $I \cup \{e\}$ for matroid $\tilde{\mathcal{M}}_i$. It is known that the current I is a maximum common set if and only if B_I has no directed path from s_1 to s_2 (see [6, 10, 15, 16]). Moreover, starting from $I = \emptyset$, the algorithm [8] finds a maximum common set I^* in time $O(|I^*|^{0.5} \max_{I \in \mathcal{I}_1 \cap \mathcal{I}_2} |B_I|)$ time, where $|B_I|$ denotes the number of vertices and edges in B_I. The number of all edges $(y, x) \in \mathcal{E}$ is $O(|\tilde{F}|)$ and all such edges can be identified in $O(|\tilde{F}|)$ time. For each $y \in F - I$, $|C_2(I, y)| = O(|I|)$ holds and all x in $C_2(I, y)$ can be identified in $O(|I|)$ time. Thus the total number of directed edges $(x, y) \in \mathcal{E}$ is $O(|I||\tilde{F}|)$ and all such edges can be identified in $O(|I||\tilde{F}|)$ time.

Therefore, $|B_I| = O(|W||\tilde{F}|)$ holds, and the entire time complexity for computing a maximum common subset I^* is $O(|I^*|^{0.5}|W||\tilde{F}|) = O(r(V)^{1.5}|\tilde{F}|) = O(\min\{r(V)^{3.5}, r(V)^{1.5}|E|r_{max}^2\})$.

If $|I^*| < r(V) - 1$, then we can conclude that the given G and r do not satisfy (1). Otherwise, if $|I^*| = r(V) - 1$, then we can construct a loopless and connected r-detachment by choosing $|E| - |I^*|$ more edges. This can be executed in $O(r(u) + r(v) + d(u, v; G))$ time for each pair of $u, v \in V$ and in $O(|V|r(V) + |E|)$ time in total. This proves Theorem 4(ii).

3.3.2 Proof of Theorem 5

Next we prove Theorem 5. We first consider the necessity of Theorem 5(i). Assume that G has a loopless and connected ρ-detachment H_ρ. It is easy to see that $1 \leq deg(v^i) = \rho_i^v$ holds for all $v^i \in W_v$ and $v \in V$. If $\rho_i^v = deg(v^i) > d(v; G) + d(v, v; G)$ holds, then at least one self-loop in $E(v, v; G)$ must be incident to v^i. Hence $\rho_i^v \leq d(v; G) + d(v, v; G)$ necessarily holds.

To show the sufficiency of Theorem 5(i), we again consider an r-detachment H of G as a spanning subgraph $H = (W, F')$ of the r-expansion $\hat{H}(G) = (W, F)$ such that $|F' \cap F_e| = 1$ for every $e \in E$. Given an r-degree specification ρ in Theorem 5, we show that F' can be modified into a ρ-detachment of G. Let $D(H)$ denote the sum $\sum_{v \in V} \sum_{1 \leq i \leq r(v)} |deg(v^i; H) - \rho_i^v|$ of difference of degrees.

Lemma 7 *Let $H = (W, F')$ be a connected spanning subgraph of the r-expansion $\hat{H}(G)$ such that $|F' \cap F_e| = 1$ for every $e \in E$. If $D(H) > 0$, then one of the following* (i) *and* (ii) *holds:*

(i) *There are edges $e_a \in F' \cap F_e$ and $e_b \in F_e - F'$ for some edge $e \in E$ such that $H' = (W, (F' - e_a) \cup \{e_b\})$ remains connected and $D(H') = D(H) - 2$ holds.*

(ii) *There are edges $e_a \in F' \cap F_e$, $e_b \in F_e - F'$ $e'_a \in F' \cap F_{e'}$ and $e'_b \in F_{e'} - F'$ for some edges $e, e' \in E$ such that $H' = (W, (F' - \{e_a, e'_a\}) \cup \{e_b, e'_b\})$ remains connected and $D(H') = D(H) - 2$ holds.*

Proof: Recall that $\hat{H}(G) = (W, F)$ is a loopless multigraph. If $D(H) > 0$, then, for some $v \in V$, there are vertices $v^j, v^h \in W_v$ such that $deg(v^j; H) < \rho^v_j$ and $deg(v^h; H) > \rho^v_h$ since $\sum_{1 \le i \le r(v)} \rho^v_i = deg(v; G)$ holds. Let $W_v = \{v^1, v^2, \ldots, v^{r(v)}\}$ and assume $deg(v^1; H) < \rho^v_1$ without loss of generality. We first claim that there is a vertex $v^\ell \in W_v$ such that $E(v^\ell, W - v^1; H) \neq \emptyset$. (hence $deg(v^\ell; H) \le \rho^v_\ell$ must hold). Otherwise if $E(v^i, W - v^1; H) = \emptyset$ holds for all $v^i \in W_v - v^1$, i.e., all edges in F' incident to a vertex in W_v are incident to v_1, then we would have $\rho^v_1 > deg(v^1; H) \ge d(v; G) + d(v, v; G)$, contradicting the assumption of ρ. This proves the claim.

Case-1: There is a vertex $v^k \in W_v$ such that $deg(v^k; H) > \rho^v_k$ and $E(v^k, W - v^1; H) \neq \emptyset$. We claim that v^1 and v^k remain connected in $H - e_a$ for some edge $e_a = \{v^k, u^q\} \in E(v^k, W - v^1; H)$, where $v^1 \neq u^q \in W_u$ (possibly $u = v$). If $E(v^k, v^1; H) \neq \emptyset$, then any edge $e_a \in E(v^k, W - v^1; H)$ will do. Assume that $E(v^k, v^1; H) = \emptyset$. Then by $deg(v^k; H) > \rho^v_k \ge 1$, we now have two edges $e_1, e_2 \in E(v^k, W - v^1; H)$. We see that v^k and v^1 remain connected in $H - e_1$ or $H - e_2$ (since v^1 and v^k become disconnected in $(W, F' - e_2)$ only when e_2 is on every path between these vertices, but in this case e_1 is not on a path containing e_2).

Therefore, v^1 and v^k are connected in $H - e_a$ for some edge $e_a = \{v^k, u^q\} \in E(v^k, W - v^1; H)$, where e_a belongs to F_e for some $e = \{v, u\} \in E$. Then by letting $e_a = \{v^k, u^q\} \in F' \cap F_e$ and $e_b = \{v^1, u^q\} \in F_e - F'$, we see that $H' = (W, (F' - e_a) \cup \{e_b\})$ remains connected and $D(H') = D(H) - 2$.

Case-2: There is no vertex $v^k \in W_v$ in Case-1. By the above claim, there are vertices $v^k, v^\ell \in W_v$ such that $d(v^k, v^1; H) = deg(v^k; H) > \rho^v_k \ge 1$ and $E(v^\ell, W - v^1; H) \neq \emptyset$. Choose edges $e_a \in E(v^k, v^1; H)$ and $e_b \in E(v^\ell, W - v^1; H)$, where $e_a = \{v^k, v^1\} \in F' \cap F_e$ and $e'_a = \{v^k, u^q\} \in F' \cap F_{e'}$ for some edges $e, e' \in E$. Note that v^k and v^1 remain connected in $H - e_a$ since $d(v^k, v^1; H) = deg(v^k; H) > \rho^v_k \ge 1$. Let $e'_a = \{v^k, u^q\} \in F' \cap F_{e'}$ and $e'_b = \{v^1, u^q\} \in F_{e'} - F'$. Therefore $H' = (W, (F' - \{e_a, e'_a\}) \cup \{e_b, e'_b\})$ remains connected and $D(H') = D(H) - 2$ holds. ∎

After modifying F' into $(F' - e_a) \cup \{e_b\}$ by edges e_a and e_b in (i) of this lemma, the resulting $H = (W, F')$ remains connected and satisfies $|F' \cap F_e| = 1$ for every $e \in E$, and the difference $D(H)$ reduces by 2. We have an analogous observation for the modification by (ii) of the lemma. Therefore by repeating these procedures until $D(H)$ becomes zero, we obtain a loopless and connected ρ-detachment $H = (W, F')$ of G. This proves Theorem 5(i). Since $D(H) \le 2|E|$, and the modification is applied $O(|E|)$ times. We represent a multigraph $H = (W, F')$ as an edge-weighted subgraph of the above simple graph $\tilde{H} = (W, \tilde{F})$, where the weight of an edge $\{u, v\}$ in the subgraph is given by $d(u, v; H)$. Then the connectivity of two vertices in H can be tested in $O(|\tilde{F}|) = O(\min\{|W|^2, |F|\}) = O(\min\{r(V)^2, |E|r_{max}^2\})$ time, and we can obtain a loopless and connected r-detachment of G in $O(|E|\min\{r(V)^2, |E|r_{max}^2\})$ time, proving Theorem 5(ii).

4 Inferring Multigraphs

We are ready to prove our results on graph inference. Given a feature vector $g \in \mathcal{F}_K(\Sigma)$, let g_k denote the vector which consists of entries $g(t)$, $t \in \Sigma^k$, $|g_k|$ denote the number of nonzero entries in g_k, and let $V_k = \{t \in \Sigma^k \mid g(t) \geq 1\}$, where $|g_k| = |V_k|$ holds. We assume that a given feature vector g is represented only by its positive entries, since otherwise it would require unnecessarily large space complexity to store many zero entries. Let $|g|$ denote the number of nonzero entries in g, and let $n = \sum_{t \in \Sigma^1} g(t)$ and $p = \max_{t \in \Sigma^1} g(t)$. Thus, $g \in \mathcal{F}_1(\Sigma)$ is given by $O(|g| \log n)$ space.

4.1 Algorithms for GIPF and GIFV

This subsection gives an algorithm for GIPPFV. A feature vector $g \in \mathcal{F}_1(\Sigma)$ is called *valid* with respect to Σ if, for the label sets $V_1 = \{t \in \Sigma^1 \mid g(t) \geq 1\}$ and $V_2 = \{t \in \Sigma^2 \mid g(t) \geq 1\}$,

$$V_2 \subseteq V_1 \times V_1, \quad g(uv) = g(vu) \text{ for all } uv \in V_2,$$

$$g(uu) \text{ is an even integer and } g(u) \geq 2 \text{ for all } uu \in V_2.$$

For a given valence-sequence $\rho : \cup_{u \in V_1} W_u \to \mathbf{Z}_+$, we call g *valid* with respect to ρ if

$$|W_u| = g(u) \text{ for all } u \in V_1,$$

$$\sum_{x \in W_u} \rho(x) = g(uu) + \sum_{v \in V_1 - \{u\}} g(uv) \text{ for all } u \in V_1.$$

$$1 \leq \rho(x) \leq g(uu)/2 + \sum_{v \in V_1 - \{u\}} g(uv) \text{ for all } x \in W_u \text{ and } u \in V_1.$$

Let $m = \sum_{t \in \Sigma^2} g(t)$.

We easily observe the following properties.

Lemma 8 *For any Σ-labeled loopless multigraph H, its feature vector $f_1(H) \in \mathcal{F}_1(\Sigma)$ of level 1 is valid with respect to Σ.* ∎

Lemma 9 *For any (Σ, ρ)-labeled loopless multigraph H, its feature vector $f_1(H) \in \mathcal{F}_1(\Sigma)$ of level 1 is valid with respect to Σ and ρ.* ∎

We derive the following results from Theorem 4.

Theorem 10 *Given an instance $I = g \in \mathcal{F}_1(\Sigma)$ of GIPF, the feasibility of I can be tested in $O(\min\{|g|2^{|g_1|}, n^{3.5} + m, n^{1.5}mp^2\})$ time, and a solution of I (if any) can be constructed in $O(\min\{n^{3.5} + m, n^{1.5}mp^2\})$ time.*

PROOF: Given a feature vector $g \in \mathcal{F}_1(\Sigma)$, we can check whether or not g is valid with respect to Σ in $O(|g_1| + |g_2|) = O(|g|)$ time. If g is not valid, then there is no loopless Σ-labeled multigraph H with $f_1(H) = g$ by Lemma 8. Consider the case where g is valid. By regarding $V_1 = \{t \in \Sigma^1 \mid g(t) \geq 1\}$ and $V_2 = \{t \in \Sigma^2 \mid g(t) \geq 1\}$ as a vertex set and an edge set, we construct a multigraph $G_g = (V = V_1, E = V_2)$ such that $d(u, v; G_g) = g(uv)(= g(vu))$ for all $u, v \in V$ with $u \neq v$ and $d(u, u; G_g) = g(uu)/2$ for all $u \in V$, where a set of edges $E(u, v; G_g)$ is stored as a single edge weighted by integer $d(u, v; G_g)$. Let $r(v) := g(v)$, $v \in V$. Since g is valid, such a multigraph G_g exists and $r(v) \geq 2$ holds for each self-loop $\{v, v\} \in E$. We see that any loopless and connected Σ-labeled multigraph H with $f_1(H) = g$ is a loopless and connected r-detachment of G_g. We test whether there exists an r-detachment H of G_g or not, and find such a solution H to I if any. This can be done in $O(\min\{r(V)^{3.5} + |E|, r(V)^{1.5}|E|r_{max}^2\}) = O(\min\{n^{3.5} + m, n^{1.5}mp^2\})$ time by Theorem 4(ii). Note that the feasiblity of I can also be tested by checking (1) for all possible subsets X of V. This takes $O(|g|2^{|g_1|})$ time since $c(G_g - X)$ can be computed in $O(|g|)$ time. ∎

For example, given feature vector $g \in \mathcal{F}_1(\{\mathtt{H}, \mathtt{O}, \mathtt{C}\})$ with $g(t) = \mathit{occ}(t, H_1)$ in Fig. 1(b), multigraph $G_g = (V, E)$ in this proof is given as in Fig. 2(a). An r-expansion $\hat{H}(G_g)$ is given in Fig 2(b), from which a loopless and connected r-detachment is obtained in Fig. 2(c), which is equivalent to graph H_2 in Fig. 1(c).

The case where inferred graphs are restricted to trees can be solved by Theorem 10.

Corollary 11 *Given an instance $I = g \in \mathcal{F}_1(\Sigma)$ of GIPF for trees, the feasibility of I can be tested in $O(\min\{|g|2^{|g_1|}, n^{3.5}, n^{2.5}p^2\})$ time, and a solution of I (if any) can be constructed in $O(\min\{n^{3.5}, n^{2.5}p^2\} + \min\{mn^2, m^2p^2\})$ time.*

PROOF: We can test if a given g satisfies $m = n-1$ or not in $O(\min\{|g|, n\})$ time. If $m \neq n-1$, then no Σ-labeled tree T with $f_2(T) = g$ exists. Otherwise (if $m = n-1$) we apply Theorem 10 to obtain a connected Σ-labeled multigraph H with $f_1(H) = g$ if any, which must be a tree since $|V(H)| = n$ and $|E(H)| = m = n-1$. ∎

Theorem 12 *Given an instance $I = (g \in \mathcal{F}_1(\Sigma), \rho)$ of GIFV, the feasibility of I can be tested in $O(\min\{n + |g|2^{|g_1|}, n^{3.5} + m, n^{1.5}mp^2\})$ time, and a solution of I (if any) can be constructed in $O(\min\{n^{3.5} + m, n^{1.5}mp^2\} + \min\{mn^2, m^2p^2\})$ time.*

PROOF: We can check if a given g is valid with respect to Σ and ρ in $O(n + |g_2|)$ time. If g is not valid, then there is no loopless (Σ, ρ)-labeled multigraph H with $f_1(H) = g$ by Lemma 9. Consider the case where g is valid. We construct a multigraph $G_g = (V = V_1, E = V_2)$, as in the proof of Theorem 10. For each $v \in V$, let $r(v) := g(v)$ and $\rho_i^v := \rho(v^i) \geq 1$ $(1 \leq i \leq r(v))$. Since g is valid, such a multigraph G_g exists. Hence ρ satisfies the necessary condition in (i) of Theorem 5. Now any loopless and connected (Σ, ρ)-labeled multigraph H with $f_1(H) = g$ is a loopless and connected r-detachment of G_g. We test whether there exists a ρ-detachment H of G_g or not and find such a solution H to I if any. This can be done in $O(\min\{r(V)^{3.5} + |E|, r(V)^{1.5}|E|r_{max}^2\}) + O(|E| \min\{r(V)^2, |E|r_{max}^2\}) = O(\min\{n^{3.5} +

$m, n^{1.5}mp^2\} + \min\{mn^2, m^2p^2\})$ time by Theorems 4(ii) and 5(i)-(ii). We can also test the feasibility of I by checking (1) for all subsets X of V, taking $O(|g|2^{|g_1|})$ time. ∎

Analogously with Corollary 11, we have the next result.

Corollary 13 *Given an instance $I = (g \in \mathcal{F}_1(\Sigma), \rho)$ of GIFV for trees, the feasibility of I can be tested in $O(\min\{n+|g|2^{|g_1|}, n^{3.5}, n^{2.5}p^2\})$ time, and a solution of I (if any) can be constructed in $O(\min\{n^{3.5}, n^{2.5}p^2\})$ time.* ∎

4.2 Algorithm for GIPPFV

This subsection gives an algorithm for GIPPFV with $K = 1$. Let $g \in \mathcal{F}_1(\Sigma)$ be a valid feature vector with respect to Σ, define V_1 and V_2 as in the previous subsection, and let $m = \sum_{u \in V_1} \sum_{x \in W_u} \rho(x)$. For each $u \in V_1$, we define its *deficit* by

$$\mathrm{dfc}(u) := \sum_{x \in W_u} \rho(x) - g(uu) - \sum_{v \in V_1 - \{u\}} g(uv).$$

For a given valence-sequence $\rho : \cup_{u \in V_1} W_u \to \mathbf{Z}_+$, we call g *weakly valid* with respect to ρ if

$$|W_u| = g(u) \text{ for all } u \in V_1,$$

$$\mathrm{dfc}(u) \geq 0 \text{ for all } u \in V_1,$$

$$1 \leq \rho(x) \leq g(uu)/2 + \sum_{v \in V_1 - \{u\}} g(uv) \text{ for all } x \in W_u \text{ and } u \in V_1.$$

We easily see that g needs to be weakly valid respect to ρ to admit a feasible solution to GIPPFV.

We now show how to reduce GIPPFV to GIFV. Suppose that there exists a loopless and connected (Σ, ρ)-labeled multigraph $H = (\cup_{u \in V_1} W_u, F_1 \cup F_2)$ with $(occ(t, H))_{t \in \Sigma} = (g(t))_{t \in \Sigma}$ and $(occ(t, H))_{t \in \Sigma^2} \geq (g(t))_{t \in \Sigma^2}$, where F_1 denotes the set of edges $(g(t))_{t \in \Sigma^2}$ and F_2 denotes the rest of edges in H. Observe that it holds

$$|F_2| = (1/2) \sum_{u \in V_1} \mathrm{dfc}(u).$$

For simplicity, we first consider the case where there is no label $\hat{u} \in V_1$ such that

$$\mathrm{dfc}(\hat{u}) > \sum_{u \in V_1 - \{\hat{u}\}} \mathrm{dfc}(u). \tag{5}$$

To reduce GIPPFV to GIFV, we introduce a new label $\mathtt{e}$, and convert H into a $\Sigma \cup \{\mathtt{e}\}$-labeled multigraph H' by subdividing each edge $e = (v, v') \in F_2$ with a new vertex v_e labeled with $\mathtt{e}$ (i.e., replacing e with two edges (v, v_e) and (v_e, v')). Note that H' remains loopless. We then contract each set of vertices with the same label into a single vertex to obtain graph $G' = (V_1 \cup \{\mathtt{e}\}, E')$, where $\mathtt{e} \in V(G')$ stands for the vertex obtained from contracting the set of vertices labeled with $\mathtt{e}$. Thus H' can be regarded as a loopless and connected detachment of G'. Finally we encode G' into a vector $g' \in \mathcal{F}_1(\Sigma \cup \{\mathtt{e}\})$ and a valence-sequence ρ' such that

$$g'(u) = g(u) \text{ for all } u \in V_1,$$

$$g'(\mathtt{e}) = |F_2| \text{ for } \mathtt{e},$$

$$g'(uv) = g(uv) \text{ for all } u, v \in V_1,$$

$$g'(u\mathtt{e}) = g'(\mathtt{e}u) = d(\mathtt{e}, u; G') \text{ for all } u \in V_1,$$

$$\rho'(x) = \rho(x) \text{ for all } x \in W_u,\ u \in V_1,$$

$$\rho'(x) = 2 \text{ for all } x \in W_{\mathtt{e}}.$$

Note that it holds

$$g'(u\mathtt{e}) = \mathrm{dfc}(u) \quad \text{for all } u \in V_1,$$

$$g'(\mathtt{e}) = (1/2) \sum_{u \in V_1} g'(u\mathtt{e}),$$

implying that the above vector g' can be determined uniquely from given g and ρ. Therefore, if GIPPFV for the given g and ρ has a solution H, then GIFV for the above vector $g' \in \mathcal{F}_1(\Sigma \cup \{\mathtt{e}\})$ and valence-sequence ρ' has a solution H'.

We now consider the case where there may exist a label $\hat{u} \in V_1$ satisfying (5), where such a label is unique. This implies that at least $(1/2)(\mathrm{dfc}(\hat{u}) - \sum_{u \in V_1 - \{\hat{u}\}} \mathrm{dfc}(u))$ edges in F_2 must join vertices labeled with $\hat{u}$ in any solution to GIPPFV. Therefore, in this case, we apply the following modification to the given vector g before converting it into the above vector g': Let

$$g(\hat{u}\hat{u}) := g(\hat{u}\hat{u}) + \mathrm{dfc}(\hat{u}) - \sum_{u \in V_1 - \{\hat{u}\}} \mathrm{dfc}(u),$$

while keeping the other entries in g unchanged (note that the modified vector g has no longer a label satisfying (5)). We see that GIFV with the resulting g' and ρ' has a solution if so does GIPPFV with given g and ρ.

We finally show that the converse is also true.

Lemma 14 *Given a weakly valid vector $g \in \mathcal{F}_1(\Sigma)$ with respect to a valence-sequence $\rho : \cup_{u \in V_1} W_u \to \mathbf{Z}_+$, define vector $g' \in \mathcal{F}_1(\Sigma \cup \{\mathtt{e}\})$ and valence-sequence ρ' in the above. Then if GIFV with g' and ρ' has a solution H', then a solution H to GIPPFV with g and ρ can be constructed from H' in $O(mn)$ time.*

Proof: Let H' be a solution to GIFV with g' and ρ'. A pair $\{(u, u_e), (u_e, u')\}$ of edges incident to a vertex u_e labeled with $\mathtt{e}$ is called an $\mathtt{e}$-*pair*. To obtain a soluton to GIPPFV from H', we eliminate all $\mathtt{e}$-pairs by apply the following transformations (a) and (b):

(a) If there is an $\mathtt{e}$-pair $\{(v, v_e), (v_e, v)\}$ of multiple edges, then find an $\mathtt{e}$-pair $\{(u, u_e), (u_e, u')\}$ such that $\{v, v_e\} \cap \{u, u', u_e\} = \emptyset$, and replace two edges (v, v_e) and (u, u_e) with two new edges (v, u_e) and (u, v_e). Repeat this until no $\mathtt{e}$-pair of multiple edges exists.

(b) Replace each e-pair $\{(v, v_e), (v_e, v')\}$ with a single edge (v, v'). Let H be the resulting Σ-labeled multigraph.

We see that an iteration in transformation (a) reduces the number of e-pairs of multiple edges by one without losing the connectivitiy and looplessness of the solution. Hence it suffices to show that a desired e-pair $\{(u, u_e), (u_e, u')\}$ can always be chosen in (a). To lead a contradiction, assume that, for an e-pair $\{(v, v_e), (v_e, v)\}$ of multple edges, there is no e-pair $\{(u, u_e), (u_e, u')\}$ such that $\{v, v_e\} \cap \{u, u', u_e\} = \emptyset$ in (a). Then all e-pairs $\{(u, u_e), (u_e, u')\}$ satisfy $v \in \{u, u'\}$, and thereby $\mathrm{dfc}(\hat{v}) > \sum_{z \in V_1 - \{\hat{v}\}} \mathrm{dfc}(z)$ holds for the label $\hat{v}$ of vertex v. This, however, contradicts that any label satisfying (5) in a given g has been eliminated by modifying g. Therefore we can execute (a) and (b) to obtain H, which is a solution to GIPPFV with g and ρ.

The number of iterations in (a) is $O(m)$, and an iteration in (a) can be executed in $O(n)$ time. Hence H can be obtained in $O(mn)$ time. ▮

By this lemma, we see that GIPPFV with g and ρ has a solution if and only if so does GIFV with g' and ρ'. By noting that G' has $O(|V(G')|^2) = O(|g_1|^2)$ weighted edges, we can derive the following results from Theorem 12, Corollary 16 and Lemma 14.

Theorem 15 *Given an instance $I = (g \in \mathcal{F}_1(\Sigma), \rho)$ of GIPPFV, the feasibility of I can be tested in $O(\min\{n + |g_1|^2 2^{|g_1|}, n^{3.5} + m, n^{1.5}mp^2\})$ time, and a solution of I (if any) can be constructed in $O(\min\{n^{3.5} + m, n^{1.5}mp^2\} + \min\{mn^2, m^2p^2\})$ time.* ▮

Corollary 16 *Given an instance $I = (g \in \mathcal{F}_1(\Sigma), \rho)$ of GIPPFV for trees, the feasibility of I can be tested in $O(\min\{n + |g|2^{|g_1|}, n^{3.5}, n^{2.5}p^2\})$ time, and a solution of I (if any) can be constructed in $O(\min\{n^{3.5}, n^{2.5}p^2\})$ time.* ▮

5 Concluding Remarks

In this paper, we proved that the problem of inferring a multigraph from frequency of paths of length at most $K = 1$ can be solved efficiently by formulating it as a problem of finding a connected detachment. Our algorithm can handle the case where each vertex is required to have a specified degree. Our new approach can be applied to problems of inferring multigraphs/digraphs with a higher connectivity since the characterizations of k-edge-connected detachments of multigraphs/digraphs have already been obtained [5, 11, 12, 17].

References

[1] T. Akutsu and D. Fukagawa, Inferring a graph from path frequency, Proc. 16th Annual Symposium on Combinatorial Pattern Matching Lecture Notes in Computer Science, 3537 (2005) 371–382.

[2] T. Akutsu and D. Fukagawa, On inference of a chemical structure from path frequency, Proc. 2005 International Joint Conference of InCoB, AASBi and KSBI, (2005) 96–100.

[3] G. H. Bakir, J. Weston, and B. Schölkopf, Learning to find pre-images, Adavnces in Neural Information Processing Systems, 16 (2004) 449–456.

[4] G. H. Bakir, A. Zien, and K. Tsuda, Learning to find graph pre-images, In Proc. the 26th DAGM Symposium, Lecture Notes in Computer Science, 3175 (2004) 253–261.

[5] A. R. Berg, B. Jackson, and T. Jordán, Highly edge-connected detachments of graphs and digraphs, J. Graph Theory, 43 (2003) 67–77.

[6] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, Combinatorial Optimization, A Wiley-Interscience Publication, John Wiley & Sons, Inc., New York, 1998.

[7] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[8] W. H. Cunningham, Improved bounds for matroid partition and intersection algorithms, SIAM J. Computing, 15 (1986) 948–957.

[9] J. Edmonds, Matroids, submodular functions, and certain polyhedra, in: Combinatorial Structures and Their Applications (R.K. Guy, H. Hanani, N. Sauer, and J. Schönheim, eds), Gordon and Breach, New York, 69–87, 1970.

[10] A. Frank, A weighted matroid intersection algorithm, J. Algorithms, 2 (1981) 328–336.

[11] T. Fukunaga and H. Nagamochi, Some theorems on detachments preserving local-edge-connectivity, Fifth CRACOW Conference on Graph Theory Electronic Notes in Discrete Mathematics, 24, (2006) 173–180.

[12] B. Jackson and T. Jordán, Non-separable detachments of graphs, J. Combin. Theory (B), 87 (2003) 17–37.

[13] J. Jansson and K. Sadakane, Private Communication.

[14] H. Kashima, K. Tsuda, and A. Inokuchi, Marginalized kernels between labeled graphs, Proc. of the 20th International Conference on Machine Learning (2003) 321–328.

[15] B. Korte and J. Vygen, Combinatorial Optimization: Theory and Algorithms, Springer-Verlag, Berlin, Heidelberg, New York, 2000.

[16] E. L. Lawler, Matroid intersection algorithms, Mathematical Programming, 9 (1975) 31–56.

[17] St. J. A. Nash-Williams, Connected detachments of graphs and generalised Euler trails, J. Lond Math Soc, 31 (1985) 17–29.

[18] G. Pólya, Kombinatorische Anzahlbestimmungen fur Gruppen, Graphen und chemische Verbindungen, Acta Math., 68 (1937) 145–254.