Tractability-preserving Transformations of Global Cost Functions

David Allouche, Christian Bessière, Patrice Boizumault, Simon De Givry, Patricia Gutierrez, Samir Loudni, Jean-Philippe Metivier, Thomas Schiex

Tractability-preserving Transformations of Global Cost Functions[☆]

David Allouche[a], Christian Bessiere[b], Patrice Boizumault[e], Simon de Givry[a], Patricia Gutierrez[c], Jimmy H.M. Lee[d,*], Ka Lun Leung[d], Samir Loudni[e], Jean-Philippe Métivier[e], Thomas Schiex[a,*], Yi Wu[d]

[a] *MIAT, UR-875, INRA, F-31320 Castanet Tolosan, France*
[b] *CNRS, University of Montpellier, France*
[c] *IIIA-CSIC, Universitat Autonoma de Barcelona, 08193 Bellaterra, Spain*
[d] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*
[e] *GREYC, Université de Caen Basse-Normandie,6 Boulevard du Maréchal Juin, 14032 Caen cedex 5, France*

Abstract

Graphical model processing is a central problem in artificial intelligence. The optimization of the combined cost of a network of local cost functions federates a variety of famous problems including CSP, SAT and Max-SAT but also optimization in stochastic variants such as Markov Random Fields and Bayesian networks. Exact solving methods for these problems typically include branch and bound and local inference-based bounds. In this paper we are interested in understanding when and how dynamic programming based optimization can be used to efficiently enforce soft local consistencies on Global Cost Functions, defined as parameterized families of cost functions of unbounded arity. Enforcing local consistencies in cost function networks is performed by applying so-called Equivalence Preserving Transformations (EPTs) to the cost functions. These EPTs may transform global cost functions and make them intractable to optimize. We identify as *tractable projection-safe* those global cost functions whose optimization is and remains tractable after applying the EPTs used for enforcing arc consistency. We also provide new classes of cost functions that are tractable projection-safe thanks to dynamic programming. We show that dynamic programming can either be directly used inside filtering algorithms, defining polynomially DAG-filterable cost functions, or emulated by arc consistency filtering on a Berge-acyclic network of bounded-arity cost functions, defining Berge-acyclic network-decomposable cost functions. We give examples of such cost functions and we provide a systematic way to define decompositions from existing decomposable global constraints. These two approaches to enforcing consistency in global cost functions are then embedded in a solver for extensive experiments that confirm the feasibility and efficiency of our proposal.

[☆] This paper is an extended version of [50] and [4].

[*] Corresponding author

1. Introduction

Cost Function Networks (CFNs) offer a simple and general framework for modeling and solving over-constrained and optimization problems. They capture a variety of problems that range from CSP, SAT and Max-SAT to maximization of likelihood in stochastic variants such as Markov Random Fields or Bayesian networks. They have been applied to a variety of real problems, in resource allocation, bioinformatics or machine learning among others [18, 60, 28, 29, 63, 2, 35].

Besides being equipped with an efficient branch and bound procedure augmented with powerful local consistency techniques, a practical CFN solver should have a good library of global cost functions to model the often complex scenarios in real-life applications.

Enforcing local consistencies requires to apply Equivalence Preserving Transformations (EPTs) such as cost projection and extension [20]. Most local consistencies require to compute minima of the cost function to determine the amount of cost to project/extend. By applying these operations, local consistencies may reduce domains and, more importantly, tighten a global lower bound on the criteria to optimize. This is crucial for branch and bound efficiency. Global cost functions have unbounded arity, but may have a specific semantics that makes available dedicated polynomial-time algorithms for minimization. However, when local consistencies apply EPTs, they modify the cost function and may break the properties that makes it polynomial-time minimizable. We say that a cost function is *tractable* if it can be minimized in polynomial time. The notion of *tractable projection-safety* captures precisely those functions that remain tractable even after EPTs.

In this paper, we prove that any tractable global cost function remains tractable after EPTs to/from the zero-arity cost function ($W_\varnothing$), and cannot remain tractable if arbitrary EPTs to/from r-ary cost functions for $r \geq 2$ are allowed. When $r = 1$, we show that the answer is indefinite. We describe a simple tractable global cost function and show how it becomes intractable after projections/extensions to/from unary cost functions. We also show that flow-based projection-safe cost functions [46] are positive examples of tractable projection-safe cost functions.

For $r = 1$, we introduce *polynomially DAG-filterable* global cost functions, which can be transformed into a filtering Directed Acyclic Graph with a polynomial number of simpler cost functions for (minimum) cost calculation. Computing minima of such cost functions, using a polynomial time dynamic programming algorithm, is tractable and remains tractable after projections/extensions. Thus, polynomially DAG-filterable cost functions are tractable projection-safe. Adding to the existing repertoire of global cost functions, cost function variants of existing global constraints such as Among, Regular, Grammar, and Max/Min, are proved to be polynomially DAG-filterable.

To avoid the need to implement dedicated dynamic programming algorithms, we also consider the possibility of directly using decompositions of global cost functions into polynomial size networks of cost functions with bounded arities, usually ternary cost functions. We show how such *network-decompositions* can be derived from known global constraint decompositions and how Berge-acyclicity allows soft local consistencies to emulate dynamic programming in this case. We prove that Berge-acyclic network-decompositions can also be used to directly build polynomial filtering DAGs.

To demonstrate the feasibility of these approaches, we implement and embed various global cost functions using filtering DAG and network-decompositions in `toulbar2`, an

open source cost function networks solver. We conduct experiments using different benchmarks to evaluate and to compare the performance of the DAG-based and network-based decomposition approaches.

The rest of the paper is organized as follows. Section 2 contains the necessary background to understand our contributions. Section 3 analyses the tractability of enforcing local consistencies on global cost functions and characterizes the conditions for preserving tractability after applying EPTs. In Section 4 we define DAG-filtering and in Section 5 we give an example of a polynomial DAG-filterable global cost function. Sections 6 and 7 present network-decomposability and the conditions for preserving the level of local consistency. Section 8 shows the relation between network-decompositions and DAG-filtering. Section 9 provides an experimental analysis of the two approaches on several classes of problems. Section 10 concludes the paper.

2. Background

We give preliminaries on cost function networks and global cost functions.

2.1. Cost Function Networks

A cost function network (CFN) is a special case of the valued constraint satisfaction problem [62] with a specific cost structure $([0, \dots, \top], \oplus, \leq)$. We give the formal definitions of the cost structure and CFN as follows.

Definition 1 (Cost Structure [62])**.** *The* cost structure $([0, \dots, \top], \oplus, \leq)$ *is a tuple defined as:*

- $[0, \dots, \top]$ *is the interval of integers from* 0 *to* $\top$ *ordered by the standard ordering* $\leq$*, where* $\top$ *is either a positive integer or* $+\infty$*.*
- $\oplus$ *is the addition operation defined as* $a \oplus b = min(\top, a + b)$*. We also define the subtraction* $\ominus$ *operator for any* a *and* b*, where* $a \geq b$*, as:*

$$a \ominus b = \begin{cases} a - b, & \text{if } a \neq \top; \\ \top, & \text{otherwise} \end{cases}$$

Note that more general additive cost structures have also been used. Specifically, VAC and OSAC [19] local consistencies are defined using a structure using non-negative rational instead of non-negative integer numbers. For ease of understanding, our discussion assumes integral costs. However, it can easily be generalized to rational costs.

Definition 2 (Cost Function Network [61])**.** *A* Cost Function Network *(CFN) is a tuple* $(\mathcal{X}, \mathcal{W}, \top)$*, where:*

- $\mathcal{X}$ *is an ordered set of discrete domain variables* $\{x_1, x_2, \dots, x_n\}$*. The domain of* $x_i \in \mathcal{X}$ *being denoted as* $D(x_i)$*;*
- $\mathcal{W}$ *is a set of cost functions* W_S *each with a scope* $S = \{x_{s_1}, \dots, x_{s_r}\} \subseteq \mathcal{X}$ *that maps tuples* $\ell \in D^S$*, where* $D^S = D(x_{s_1}) \times \cdots \times D(x_{s_r})$*, to* $[0, \dots, \top]$*.*

When the context is clear, we abuse notation by denoting an assignment of a set of variables $S \subseteq \mathcal{X}$ as a tuple $\ell = (v_{s_1}, \ldots, v_{s_r}) \in D^S$. The notation $\ell[x_{s_i}]$ denotes the value v_{s_i} assigned to x_{s_i} in ℓ, and $\ell[S']$ denotes the tuple formed by projecting ℓ onto $S' \subseteq S$. Without loss of generality, we assume $\mathcal{W} = \{W_\varnothing\} \cup \{W_i \mid x_i \in \mathcal{X}\} \cup \mathcal{W}^+$. $W_\varnothing$ is a constant zero-arity cost function. W_i is a unary cost function associated with each $x_i \in \mathcal{X}$. $\mathcal{W}^+$ is a set of cost functions W_S with scope S and $|S| \geq 2$. If $W_\varnothing$ and $\{W_i\}$ are not defined, we assume $W_i(v) = 0$ for all $v \in D(x_i)$ and $W_\varnothing = 0$. To simplify notation, we also denote by $W_{s_1,s_2,\ldots,s_r}$ the cost function on variables $\{x_{s_1}, x_{s_2}, \ldots, x_{s_r}\}$ when the context is clear.

Definition 3. *Given a CFN $(\mathcal{X}, \mathcal{W}, \top)$, the* cost *of a tuple $\ell \in D^\mathcal{X}$ is defined as $cost(\ell) = \bigoplus_{W_S \in \mathcal{W}} W_S(\ell[S])$. A tuple $\ell \in D^\mathcal{X}$ is* feasible *if $cost(\ell) < \top$, and it is an* optimal solution *of the CFN if $cost(\ell)$ is minimum among all tuples in $D^\mathcal{X}$.*

We observe that a classical *Constraint Network* is merely a CFN where all cost functions $W_S \in \mathcal{W}$ are such that $\forall \ell \in D^S, W_S(\ell) \in \{0, \top\}$. The problem of the existence of a solution in a constraint network, called *Constraint Satisfaction Problem (CSP)*, is NP-complete. Finding an optimal solution to a CFN is thus above NP. Restrictions to Boolean variables and binary constraints are known to be APX-hard [53]. In the terminology of stochastic graphical models, this problem is also equivalent to the Maximum A Posteriori (MAP/MRF) problem or the Maximum Probability Explanation (MPE) in Bayesian networks [35]. CFNs can be solved exactly with depth-first branch-and-bound search using $W_\varnothing$ as a lower bound. Search efficiency is enhanced by maintaining local consistencies that increase the lower bound by redistributing costs among W_S, pushing costs into $W_\varnothing$ and W_i, and pruning values while preserving the equivalence of the problem (*i.e.*, the cost of each tuple $\ell \in D^\mathcal{X}$ is unchanged).

2.2. Soft local consistencies and EPTs

Different consistency notions have been defined. Examples include NC* [42], (G)AC* [61, 42, 20, 46, 48], FD(G)AC* [42, 41, 46, 48], (weak) ED(G)AC* [32, 47, 48], VAC and OSAC [19]. Enforcing such local consistencies requires applying equivalence preserving transformations (EPTs) that shift costs between different scopes. The main EPT is defined below and described as Algorithm 1. This is a compact version of the projection and extension defined in [22].

Definition 4 (EPTs [22])**.** *Given two cost functions W_{S_1} and W_{S_2}, $S_2 \subset S_1$, the EPT* `Project` *(S_1, S_2, ℓ, α) shifts an amount of cost α between a tuple $\ell \in D^{S_2}$ of W_{S_2} and the cost function W_{S_1}. The direction of the shift is given by the sign of α. The precondition guarantees that costs remain non negative after the EPT has been applied.*

Denoting by $r = |S_2|$, the EPT is called an r-EPT. It is an r-projection when $\alpha \geq 0$ and an r-extension when $\alpha < 0$.

It is now possible to introduce local consistency enforcing algorithms.

Definition 5 (Node Consistency [42])**.** *A variable x_i is* star node consistent *(NC*) if each value $v \in D(x_i)$ satisfies $W_i(v) \oplus W_\varnothing < \top$ and there exists a value $v' \in D(x_i)$ such that $W_i(v') = 0$. A CFN is NC* iff all variables are NC*.*

Precondition: $-W_{S_2}(\ell) \leq \alpha \leq \min_{\ell' \in D^{S_1}, \ell'[S_2]=\ell} W_{S_1}(\ell')$;

Procedure `Project`(S_1, S_2, ℓ, α)

$W_{S_2}(\ell) \leftarrow W_{S_2}(\ell) \oplus \alpha$;

foreach $(\ell' \in D^{S_1}$ such that $\ell'[S_2] = \ell)$ **do**

$W_{S_1}(\ell') \leftarrow W_{S_1}(\ell') \ominus \alpha$;

Algorithm 1: A cost shifting EPT used to enforce soft local consistencies. The $\oplus, \ominus$ operations are extended here to handle possibly negative costs as follows: for non-negative costs α, β, we have $\alpha \ominus (-\beta) = \alpha \oplus \beta$ and for $\beta \leq \alpha$, $\alpha \oplus (-\beta) = \alpha \ominus \beta$.

Procedure `enforceNC*`()

1 **foreach** $x_i \in \mathcal{X}$ **do** `unaryProject`(x_i);

2 **foreach** $x_i \in \mathcal{X}$ **do** `pruneVar`(x_i);

Procedure `unaryProject`(x_i)

3 $\alpha := \min\{W_i(v) \mid v \in D(x_i)\}$;

4 `Project`$(\{x_i\}, \varnothing, (), \alpha)$;

Procedure `pruneVar`(x_i)

5 **foreach** $v \in D(x_i)$ s.t. $W_i(v) \oplus W_\varnothing = \top$ **do**

6 $D(x_i) := D(x_i) \setminus \{v\}$;

Algorithm 2: Enforce NC*

Procedure `enforceNC*`() in Algorithm 2 enforces NC*, where `unaryProject`() applies EPTs that move unary costs towards $W_\varnothing$ while keeping the solution unchanged, and `pruneVar`(x_i) removes infeasible values.

Definition 6 ((Generalized) Arc Consistency [20, 46, 48])**.** *Given a CFN $P = (\mathcal{X}, \mathcal{W}, \top)$, a cost function $W_S \in \mathcal{W}^+$ and a variable $x_i \in S$.*

- *A tuple $\ell \in D^S$ is a* simple support *for $v \in D(x_i)$ with respect to W_S with $x_i \in S$ iff $\ell[x_i] = v$ and $W_S(\ell) = 0$.*
- *A variable $x_i \in S$ is* star generalized arc consistent (GAC*) *with respect to W_S iff*
  - *x_i is NC*;*
  - *each value $v_i \in D(x_i)$ has a simple support ℓ with respect to W_S.*
- *A CFN is* GAC* *iff all variables are GAC* with respect to all related non-unary cost functions.*

To avoid exponential space complexity, the GAC* definition and the algorithm is slightly different from the one given by Cooper and Schiex [20], which also requires for every tuple $\ell \in D^S$, $W_S(\ell) = \top$ if $W_\varnothing \oplus \bigoplus_{x_i \in S} W_i(\ell[x_i]) \oplus W_S(\ell) = \top$.

The procedure `enforceGAC*`() in Algorithm 3, enforces GAC* on a single variable $x_i \in \mathcal{X}$ with respect to a cost function $W_S \in \mathcal{W}^+$, where $x_i \in S$ in a CFN $(\mathcal{X}, \mathcal{W}, \top)$. The procedure first computes the minimum when $x_i = v$ for each $v \in D(x_i)$ at line 2, then performs a 1-projection from W_S to W_i at line 3. Lines 4 and 5 enforce NC* on x_i.

Procedure `enforceGAC*`(W_S, x_i)

1 **foreach** $v \in D(x_i)$ **do**
2 $\quad \alpha := \min\{W_S(\ell) \mid \ell \in D^S \wedge \ell[x_i] = v\}$;
3 $\quad$ `Project`$(S, \{x_i\}, (v), \alpha)$
4 `unaryProject`(x_i);
5 `pruneVar`(x_i);

Algorithm 3: Enforcing GAC* for x_i with respect to W_S

Local consistency enforcement involves two types of operations: (1) finding the minimum cost returned by the cost functions among all (or part of the) tuples; (2) applying EPTs that shift costs to and from smaller-arity cost functions.

Minimum cost computation corresponds to line 3 in Algorithm 2, and line 2 in Algorithm 3. For simplicity, we write $min\{W_S(\ell) \mid \ell \in D^S\}$ as $min\{W_S\}$.

In practice, projections and extensions can be performed in constant time using the Δ data-structure introduced in Cooper and Schiex [20]. For example, when we perform 1-projections or 1-extensions, instead of modifying the costs of all tuples, we store the projected and extended costs in $\Delta^-_{x_i,v}$ and $\Delta^+_{x_i,v}$ respectively. Whenever we compute the value of the cost function W_S for a tuple ℓ with $\ell[x_i] = v$, we return $W_S(\ell) \ominus \Delta^-_{x_i,v} \oplus \Delta^+_{x_i,v}$. The time complexity of enforcing one of the previous consistencies is thus entirely defined by the time complexity of computing the minimum of a cost function during the enforcing.

Proposition 1. *The procedure* `enforceGAC*`() *in Algorithm 3 requires* $O(d \cdot f_{min})$ *time, where* d *is the maximum domain size and* f_{min} *is the time complexity of minimizing* W_S.

Proof. Line 2 requires $O(f_{min})$ time. We can replace the domain of x_i by $\{v\}$, and run the minimum computation to get the minimum cost. Projection at line 3 can be performed in constant time. Thus, each iteration requires $O(f_{min})$. Since the procedure iterates d times, and the procedures `unaryProject` and `pruneVar` requires $O(d)$, the overall complexity is $O(d \cdot f_{min} + d) = O(d \cdot f_{min})$. □

In the general case, f_{min} is in $O(d^r)$ where r is the size of the scope and d the maximum domain size. However, a *global cost function* may have specialized algorithms which make the operation of finding minimum, and thus consistency enforcement, tractable.

2.3. Global Constraints, Soft Global Constraints and Global Cost Functions

Definition 7 (Global Constraint [9, 59])**.** *A* global constraint*, denoted by* $\mathrm{GC}(S, A_1, \ldots, A_t)$*, is a family of hard constraints parameterized by a scope* S*, and possibly extra parameters* $A_1, \ldots, A_t$.

Examples of global constraints are AllDifferent [43], GCC [58], Same [11], Among [10], Regular [55], Grammar [37], and Maximum/Minimum constraints [7]. Because of their unbounded scope, global constraints cannot be efficiently propagated by generic local consistency algorithms, which are exponential in the arity of the constraint. Specific propagation algorithms are designed to achieve polynomial time complexity in the size of the input, *i.e.* the scope, the domains and extra parameters.

To capture the idea of costs assigned to constraint violations, the notion of *soft global constraint* has been introduced. This is a traditional global constraint with one extra variable representing the cost of the assignment w.r.t. to an existing global constraint. The cost is given by a violation measure function.

Definition 8 (Soft Global Constraint [56]). *A* soft global constraint*, denoted by* $\text{Soft\_GC}^{\mu}(S \cup \{z\}, A_1, \ldots, A_t)$*, is a family of hard constraints parameterized by a violation measure* μ*, a scope* S*, a cost variable* z*, and possibly extra parameters* $A_1, \ldots, A_t$*. The constraint is satisfied if and only if* $z = \mu(S, A_1, \ldots, A_t)$*.*

Soft global constraints are used to introduce costs in the CSP framework, and therefore inside constraint programming solvers [57]. It requires the introduction of extra cost variables and does not exploit the stronger propagation offered by some of the soft local consistencies. A possible alternative, when a sum of costs needs to be optimized, lies in the use of *global cost functions*.

Definition 9 (Global Cost Function [65, 48]). *A* global cost function*, denoted as* $\text{W\_GCF}(S, A_1, \ldots, A_t)$*, is a family of cost functions parameterized by a scope* S *and possibly extra parameters* $A_1, \ldots, A_t$*.*

For example, if S is a set of variables with non-negative integer domains, it is easy to define the Global Cost Function $\text{W\_Sum}(S) \equiv \bigoplus_{x_i \in S} \min(\top, x_i)$.

It is possible to derive a global cost function from an existing soft global constraint $\text{Soft\_GC}^{\mu}(S \cup \{z\}, A_1, \ldots, A_t)$. In this case, we denote the corresponding global cost function as W_GCF^{μ}. Its value for a tuple $\ell \in D^S$ is equal to $\min(\top, \mu(\ell))$.

For example, global cost functions $\text{W\_AllDifferent}^{var}$/$\text{W\_AllDifferent}^{dec}$ [46, 48] can be derived from two different violation measures of AllDifferent, namely variable-based and decomposition-based [56, 34], respectively. Other examples include W_GCC^{var} and W_GCC^{val} [46, 48], W_Same^{var} [46, 48], $\text{W\_SlidingSum}^{var}$ [51], W_Regular^{var} and W_Regular^{edit} [4, 46, 48], W_EGCC^{var} [51], $\text{W\_Disjunctive}^{val}$ and $\text{W\_Cumulative}^{val}$ [51, 49].

3. Tractable Projection-Safety

All soft local consistencies are based on the use of EPTs, shifting costs between two scopes. The size of the smallest scope used in a EPT is called the *order* (r) of the EPT. Such a EPT is called an r-EPT. It is directly related to the level of local consistency enforced: node consistency uses EPTs onto the empty scope ($r = 0$), arc consistencies use unary scopes ($r = 1$) whereas higher-order consistencies use larger scopes ($r \geq 2$) [22]. In this section, we show that the order of the EPTs directly impacts the tractability of global cost function minimization.

To be able to analyze complexities in global cost functions, we first define the decision problem associated with the optimization problem $\min\{W\_GCF(S, A_1, \ldots, A_t)\}$.

$\text{IsBetterThan}(\text{W\_GCF}(S, A_1, \ldots, A_t), m)$

Instance. A global cost function W_GCF, a scope S with domains for the variables in S, values for the parameters $A_1, \ldots, A_t$, and a fixed integer m.

Question. Does there exist a tuple $\ell \in D^S$ such that $\text{W\_GCF}(S, A_1, \ldots, A_t)(\ell) < m$?

We can then define the tractability of a global cost function.

Definition 10. *A global cost function* W_GCF$(S, A_1, \ldots, A_t)$ *is said to be* tractable *iff the problem* IsBetterThan*(*W_GCF$(S, A_1, \ldots, A_t), m)$ *is in* P.

For a tractable global cost function $W_S =$W_GCF$(S, A_1, \ldots, A_t)$, the time complexity of computing $\min\{W_S\}$ is bounded above by a polynomial function in the size of the input, including the scope, the corresponding domains, the other parameters of the global cost function, and $\log(m)$.

We introduce *tractable r-projection-safety* global cost functions, which remain *tractable* after applying r-EPTs.

Definition 11. *We say that a global cost function W_GCF$(S, A_1, \ldots, A_t)$ is* tractable r-projection-safe *iff:*

- *it is tractable and;*
- *any global cost functions that can be derived from W_GCF$(S, A_1, \ldots, A_t)$ by a series of r-EPTs is also tractable.*

The tractability after r-EPTs depends on r. We divide the discussion of tractable r-projection-safety into three cases: $r = 0$, $r \geq 2$ and $r = 1$. In the following, given a tractable global cost function W_S, we denote by $\nabla_r(W_S)$ the global cost function resulting from the application of an arbitrary finite sequence of r-EPTs on W_S.

3.1. Tractability and 0-EPTs

When $r = 0$, EPTs are performed to/from $W_\varnothing$. This kind of EPTs is used when enforcing Node Consistency (NC*) [42] but also in $\varnothing$-inverse consistency [65], and strong $\varnothing$-inverse consistency [46, 48].

We show that if a global cost function is tractable, it remains tractable after applying such EPTs.

Theorem 1. *Every tractable global cost function is tractable 0-projection-safe.*

Proof. Consider a tractable global cost function $W_S = W\_GCF(S, A_1, \ldots, A_t)$. Clearly, W_S and $\nabla_0(W_S)$ only differ by a constant, *i.e.* there exists α^- and α^+, where $\alpha^-, \alpha^+ \in \{0, \ldots, \top\}$, such that:

$$\nabla_0(W_S)(\ell) = W_S(\ell) \oplus \alpha^+ \ominus \alpha^-, \text{ for all } \ell \in D^S$$

If $W_S(\ell) = \min\{W_S\}$ for some $\ell \in D^S$, then $\nabla_0(W_S)(\ell) = \min\{\nabla_0(W_S)\}$. If W_S is tractable, so is $\nabla_0(W_S)$. □

3.2. Tractability and EPTs of order greater than 2

When $r \geq 2$, EPTs are performed to/from r-arity cost functions. This is required for enforcing higher order consistencies and is used in practice in ternary cost functions processing [60] and complete k-consistency [22].

If arbitrary sequences of r-EPTs are allowed, we show that tractable global cost functions always become intractable after some sequence of r-EPT applications, where $r \geq 2$.

Theorem 2. *Any tractable global cost function* W_GCF$(S, A_1, \ldots, A_t)$ *returning finite costs is not tractable r-projection-safe for $r \geq 2$, unless $P = NP$.*

Proof. Let us first define the binary constraint satisfaction problem AritytwoCSP as follows.

AritytwoCSP$(\mathcal{X}, \mathcal{W}^h)$

Instance. A CSP instance $(\mathcal{X}, \mathcal{W}^h)$, where every constraint $C_S^h \in \mathcal{W}^h$ involves two variables, *i.e.* $|S| = 2$.

Question. Is the CSP $(\mathcal{X}, \mathcal{W}^h)$ satisfiable?

AritytwoCSP is NP-hard as graph coloring can be solved through a direct modeling into AritytwoCSP. We reduce the problem AritytwoCSP$(\mathcal{X}, \mathcal{W}^h)$ to the problem IsBetterThan$(\nabla_2(W_\mathcal{X}), \top)$, where $W_\mathcal{X} =$ W_GC$(\mathcal{X}, A_1, \ldots, A_t)$ is an arbitrary global cost function using only finite costs. We first construct a CFN $(\mathcal{X}, \mathcal{W} \cup \{W_\mathcal{X}\}, \top)$. The upper bound $\top$ is a sufficiently large integer such that $\top > W_\mathcal{X}(\ell)$ for every $\ell \in D^S$, which is always possible given that $W_\mathcal{X}$ remains finite. This technical restriction is not significant: if a global cost function W_S maps some tuples to infinity, we can transform it to another cost function W'_S such that the infinity costs are replaced by a sufficiently large integer p such that $p \gg \max\{W_S(\ell) \mid \ell \in D^S \wedge W_S(\ell) \neq +\infty\}$.

The cost functions $W_S \in \mathcal{W} \setminus \{W_\mathcal{X}\}$ are defined as follows:

$$W_S(\ell) = \begin{cases} 0, & \text{if } \ell \text{ is accepted by } C_S^h \in \mathcal{W}^h; \\ \top, & \text{otherwise} \end{cases}$$

From the CFN, ∇_2 can be defined as follows: for each forbidden tuple $\ell[S]$ in each $C_S^h \in \mathcal{W}^h$, we add an extension of $\top$ from W_S to $W_\mathcal{X}$ with respect to $\ell[S]$ into ∇_2. Under this construction, $\nabla_2(W_\mathcal{X})(\ell)$ can be represented as:

$$\nabla_2(W_\mathcal{X})(\ell) = W_\mathcal{X}(\ell) \oplus \bigoplus_{W_S \in \mathcal{W}} W_S(\ell[S])$$

For a tuple $\ell \in D^\mathcal{X}$, $\nabla_2(W_\mathcal{X})(\ell) = \top$ iff ℓ is forbidden by some C_S^h in $\mathcal{W}^h$. As a result IsBetterThan$(\nabla_2(W_\mathcal{X}), \top)$ is satisfiable iff AritytwoCSP$(\mathcal{X}, \mathcal{W}^h)$ is satisfiable. As AritytwoCSP is NP-hard, IsBetterThan$(\nabla_2(W_GC), \top)$ is not polynomial, unless $P = NP$. Hence, $\nabla_2(W\_GC)$ is not tractable, and then, W_GC is not tractable 2-projection-safe, unless $P = NP$. □

3.3. Tractability and 1-EPTs

When $r = 1$, 1-EPTs cover 1-projections and 1-extensions, which are the backbone of the consistency algorithms of (G)AC* [42, 46, 48], FD(G)AC* [41, 46, 48], (weak) ED(G)AC* [32, 47, 48], VAC, and OSAC [19]. In these cases, tractable cost functions are tractable 1-projection-safe only under special conditions. For example, Lee and Leung define *flow-based projection-safety* based on a flow-based global cost function.

Definition 12 (Flow-based [46, 48])**.** *A global cost function $W\_GCF(S, A_1, \ldots, A_t)$ is* flow-based *iff it can be represented as a flow network G such that the minimum cost among all maximum flows between a fixed source and a fixed destination is equal to* $\min\{W\_GCF(S, A_1, \ldots, A_t)\}$.

Definition 13 (Flow-based projection safe [46, 48])**.** *A global cost function* $W\_GCF(S, A_1, \ldots, A_t)$ *is flow-based projection-safe iff it is is flow-based, and is still flow-based following any sequence of* 1*-projections and* 1*-extensions.*

Lee and Leung [46, 48] further propose sufficient conditions for tractable cost functions to be flow-based projection-safe. Flow-based projection-safety implies tractable 1-projection-safety. We state the result in the following theorem.

Theorem 3. *Any flow-based projection-safe global cost function is tractable* 1*-projection-safe.*

Proof. Follows directly from the tractability of the minimum cost flow algorithm. □

However, tractable cost functions are not necessarily tractable 1-projection-safe. One example is W_2SAT, which is a global cost function derived from an instance of the polynomial 2SAT problem.

Definition 14. *Given a set of Boolean variables* S*, a set of binary clauses* F*, and a positive integer* c*, the global cost function* W_2SAT(S, F, c) *is defined as:*

$$W\_2SAT(S, F, c)(\ell) = \begin{cases} 0, & \text{if } \ell \text{ satisfies } F \\ c, & \text{otherwise} \end{cases}$$

W_2SAT is tractable, because the 2SAT problem is tractable [39]. However, it is not tractable 1-projection-safe.

Theorem 4. W_2SAT *is not tractable* 1*-projection-safe, unless* $P = NP$*.*

Proof. Let us first define the WSAT-2-CNF problem.

> WSAT-2-CNF
> **Instance**. A 2-CNF formula F (a set of binary clauses) and a fixed integer k.
> **Question**. Is there an assignment that satisfies all clauses in F with at most k variables set to *true* ?

WSAT-2-CNF was shown NP-hard in [31, page 69]. We reduce it to the problem IsBetterThan$(\nabla_1(W_2SAT), \top)$.

We construct a particular sequence of 1-projections and/or 1-extensions ∇_1 such that the WSAT-2-CNF instance can be solved using $W_{\mathcal{X}} = W\_2SAT(\mathcal{X}, F, k+1)$ from the Boolean CFN $N = (\mathcal{X}, \mathcal{W} \cup \{W_{\mathcal{X}}\}, k+1)$. $\mathcal{W}$ only contains unary cost functions W_i, which are defined as follows:

$$W_i(v) = \begin{cases} 1, & \text{if } v = true; \\ 0, & \text{otherwise} \end{cases}$$

Based on N, we construct ∇_1 as follows: for each variable $x_i \in \mathcal{X}$, we add an extension of 1 from W_i to $W_{\mathcal{X}}$ with respect to the value *true* into ∇_1. As a result, a tuple ℓ with $\nabla_1(W_{\mathcal{X}})(\ell) = k' \leq k$ contains exactly k' variables set to *true* (because every $x_i = true$ incurs a cost of 1) and also satisfies F (or it would have cost $k+1 = \top$). Thus, the WSAT-2-CNF instance with threshold k is satisfiable iff IsBetterThan$(\nabla_1(W_{\mathcal{X}}), k+1)$ is satisfiable. As WSAT-2-CNF is NP-hard, IsBetterThan$(\nabla_1(W_2SAT), k+1)$ is not polynomial, unless $P = NP$. Hence, $\nabla_1(W\_2SAT)$ is not tractable, and then, W_2SAT is not tractable 1-projection-safe, unless $P = NP$. □

When the context is clear, we use tractable projection-safety, projection and extension to refer to tractable 1-projection-safety, 1-projection and 1-extension respectively hereafter.

4. Polynomial DAG-Filtering

Beyond flow-based global cost functions [46, 48], we introduce now an additional class of tractable projection-safe cost functions based on dynamic programming algorithms. As mentioned by Dasgupta *et al.* [25], every dynamic programming algorithm has an underlying DAG structure.

Definition 15 (DAG)**.** *A* directed acylic graph (DAG) $T = (V, E)$*, where* V *is a set of vertices (or nodes) and* $E \subseteq V \times V$ *is a set of directed edges, is a directed graph with no directed cycles, and:*

- *An edge* $(u, v) \in E$ *points from* u *to* v*, where* u *is the* parent *of* v*, and* v *is the* child *of* u*;*
- *A* root *of a DAG is a vertex with zero in-degree;*
- *A* leaf *of a DAG is a vertex with zero out-degree;*
- *An* internal vertex *of a DAG is any vertex which is not a leaf;*

We now introduce the *DAG filterability* of a global cost function.

Definition 16 (DAG-filter)**.** *A* DAG-filter *for a cost function* W_S *is a DAG* $T = (V, E)$ *such that:*

- T *is connected;*
- $V = \{\omega_{S_i}\}_i$ *is a set of cost function vertices each with a scope* S_i*, among which vertex* W_S *is the root of* T*;*
- *Each internal vertex* ω_{S_i} *in* V *is associated with an aggregation function* f_i *that maps a multiset of costs* $\{\alpha_j \mid \alpha_j \in [0 \dots \top]\}$ *to* $[0 \dots \top]$ *and is based on an associative and commutative binary operator;*
- *For every internal* $\omega_{S_i} \in V$*,*
  - *the scope of* ω_{S_i} *is composed from its children's scopes:*

$$S_i = \bigcup_{(\omega_{S_i}, \omega_{S_j}) \in E} S_j$$

  - ω_{S_i} *is the aggregation of its children:*

$$\omega_{S_i}(\ell) = f_i(\{\omega_{S_j}(\ell[S_j]) \mid (\omega_{S_i}, \omega_{S_j}) \in E\});$$

  - min *is distributive over* f_i*:*

$$\min\{\omega_{S_i}\} = f_i(\{\min\{\omega_{S_j}\} \mid (\omega_{S_i}, \omega_{S_j}) \in E\}).$$

When a cost function W_S has a DAG-filter T, we say that W_S is DAG-filterable by T. Note that any cost function W_S has a trivial DAG filter which is composed of a single vertex that defines W_S as a cost table (with size exponential in the arity $|S|$).

In the general case, a DAG-filter (recursively) transforms a cost function into cost functions with smaller scopes until it reaches the ones at the leaves of a DAG, which may be trivial to solve. The (minimum) costs can then be aggregated using the f_i functions at each internal vertex to get the resultant (minimum) cost, through dynamic programming. However, further properties on DAG-filters are required to allow for projections and extensions to operate on the DAG structure.

Definition 17 (Safe DAG-filter)**.** *A DAG-filter $T = (V, E)$ for a cost function W_S is* safe *iff:*

- *projection and extension are distributive over f_i, i.e. for a variable $x \in S$, a cost α and a tuple $\ell \in D^S$,*
  - $\omega_{S_i}(\ell[S_i]) \oplus \nu_{x,S_i}(\alpha) = f_i(\{\omega_{S_k}(\ell[S_k]) \oplus \nu_{x,S_k}(\alpha) \mid (\omega_{S_i}, \omega_{S_k}) \in E\})$*, and;*
  - $\omega_{S_i}(\ell[S_i]) \ominus \nu_{x,S_i}(\alpha) = f_i(\{\omega_{S_k}(\ell[S_k]) \ominus \nu_{x,S_k}(\alpha) \mid (\omega_{S_i}, \omega_{S_k}) \in E\})$*,*

  where the function ν is defined as:

$$\nu_{x,S_j}(\alpha) = \begin{cases} \alpha, & \text{if } x \in S_j, \\ 0, & \text{otherwise.} \end{cases}$$

The requirement of a distributive f_i with respect to projection and extension at each vertex in T implies that the structure of the DAG is unchanged after projections and extensions. Both operations can be distributed down to the leaves. We formally state this as the following theorem. Given a variable x, with a value $a \in D(x)$, and a cost function W_S, we denote as W'_S the cost function obtained by the application of `Project`$(S, \{x\}, (v), \alpha)$ on W_S if $x \in S$ or W_S otherwise.

Theorem 5. *For a cost function W_S with a safe DAG-filter $T = (V, E)$, W'_S has a safe DAG-filter $T' = (V', E')$, where each $\omega_{S_i} \in V'$ is defined as:*

$$\omega'_{S_i} = \begin{cases} \omega_{S_i} \ominus \nu_{x,S_k}(\alpha), & \text{if } \omega_{S_i} \text{ is a leaf of } T, \\ \omega_{S_i}, & \text{otherwise.} \end{cases}$$

and $(\omega'_{S_i}, \omega'_{S_k}) \in E'$ iff $(\omega_{S_i}, \omega_{S_k}) \in E$, i.e. T' is isomorphic to T. Moreover, both $\omega'_{S_i} \in V'$ and $\omega_{S_i} \in V$ are associated with the same aggregation function f_i.

Proof. Follows directly from Definition 17. □

Two common choices for f_i are $\oplus$ and min, with which distributivity depends on how scopes intersect. In the following, we show that the global cost function is safely DAG-filterable if the internal vertices that are associated with $\oplus$ have children with non-overlapping scopes, and those associated with min have children with identical scopes.

Proposition 2. *Any DAG-filter $T = (V, E)$ for a cost function W_S such that*

- *each $\omega_{S_i} \in V$ is associated with the aggregation function $f_i = \bigoplus$;*

- *for any distinct* $\omega_{S_j}, \omega_{S_k} \in V$*, which are children of* ω_{S_i}*,* $S_j \cap S_k = \varnothing$*.*

is safe.

Proof. We need to show that min, projection and extension are distributive over $\oplus$. Since the scopes of the cost functions do not overlap, min is distributive over $\oplus$. We further show the distributivity with respect to projection ($\ominus$), while extension ($\oplus$) is similar. We consider an internal vertex $\omega_{S_i} \in V$. Given a variable $x \in S_i$, a cost α, and a tuple $\ell \in D^S$, since the scopes of the cost functions $\{\omega_{S_k} \mid (\omega_{S_i}, \omega_{S_k}) \in E\}$ are disjoint, there must exist exactly one cost function ω_{S_j} such that $x \in S_j$, *i.e.*:

$$\begin{aligned}\omega_{S_i}(\ell) \ominus \alpha &= (\omega_{S_j}(\ell[S_j]) \ominus \alpha) \oplus \bigoplus_{k \neq j \wedge (\omega_{S_i}, \omega_{S_k}) \in E} \omega_{S_k}(\ell[S_k]) \\ &= \bigoplus_{(\omega_{S_i}, \omega_{S_k}) \in E} (\omega_{S_k}(\ell[S_k]) \ominus \nu_{x,S_k}(\alpha))\end{aligned}$$

The result follows. □

Proposition 3. *Any DAG-filter* $T = (V, E)$ *for a cost function* W_S *such that*

- *each* $\omega_{S_i} \in V$ *is associated with the aggregation function* $f_i = \min$*;*
- $\forall \omega_{S_j} \in V$*, which are children of* ω_{S_i}*,* $S_j = S_i$*.*

is safe.

Proof. Since the scopes are completely overlapping,

$$\begin{aligned}\min\{\omega_{S_i}\} &= \min_{\ell \in D^{S_i}} \{ \min_{(\omega_{S_i}, \omega_{S_k}) \in E} \{\omega_{S_k}(\ell)\}\} \\ &= \min_{(\omega_{S_i}, \omega_{S_k}) \in E} \{ \min_{\ell \in D^{S_k}} \{\omega_{S_k}(\ell)\}\} \\ &= f_i(\{\min\{\omega_{S_k}\} \mid (\omega_{S_i}, \omega_{S_k}) \in E\})\end{aligned}$$

It is trivial to see that projection and extension are distributive over f_i. The result follows. □

We are now ready to define polynomial DAG-filterability of global cost functions. As safe DAG-filters can be exponential in size, we need to restrict to safe DAG-filters of polynomial size by restricting the size of the DAG to be polynomial and by bounding the arity of the cost functions at the leaves of the DAG.

Definition 18 (Polynomial DAG-filterability)**.** *A global cost function* $W\_GCF(S, A_1, \ldots, A_t)$ *is* polynomially DAG-filterable *iff*

1. *any instance* W_S *of* $W\_GCF(S, A_1, \ldots, A_t)$ *has a safe DAG-filter* $T = (V, E)$
2. *where* $|V|$ *is polynomial in the size of the input parameters of* $W\_GCF(S, A_1, \ldots, A_t)$*;*
3. *each leaf in* V *is a unary cost function, and*
4. *each aggregation function* f_i *associated with each internal vertex is polynomial-time computable.*

Dynamic programming can compute the minimum of a polynomially DAG-filterable cost function in a tractable way. Projections and extensions to/from such cost functions can also be distributed to the leaves in T. Thus, polynomially DAG-filterable global cost functions are tractable and also tractable projection-safe, as stated below.

Theorem 6. *A polynomially DAG-filterable global cost function $W\_GCF(S, A_1, \ldots, A_t)$ is tractable.*

Proof. Let W_S be any instance of $W\_GCF(S, A_1, \ldots, A_t)$, and $T = (V, E)$ be a safe DAG-filter for W_S. Algorithm 4 can be applied to compute $\min\{W_S\}$. The algorithm uses a bottom-up memoization approach. Algorithm 4 first sorts V topologically at line 2. After sorting, all the leaves will be grouped at the end of the sorted sequence, which is then processed in the reversed order at line 3. If the vertex is a leaf, the minimum is computed and stored in the table `Min` at line 5. Otherwise, its minimum is computed by aggregating $\{\texttt{Min}[\omega_{S_k}] \mid (\omega_{S_i}, \omega_{S_k}) \in E\}$, which have been already computed, by the function f_i at line 6. Line 7 returns the minimum of the root node.

The computation is tractable. Leaves being unary cost functions, line 5 is in $O(d)$, where d is the maximum domain size. For other vertices, line 6 calls f_i, which is assumed to be polynomial time. The result follows. □

```
   Function Minimum (W_S)
1  |  Form the corresponding filtering DAG T = (V, E);
2  |  Topologically sort V;
3  |  foreach ω_{S_i} ∈ V in reverse topological order do
4  |  |  if ω_{S_i} is a leaf of T then
5  |  |  |  Min[ω_{S_i}] := min{ω_{S_i}} ;
   |  |  else
6  |  |  |  Min[ω_{S_i}] := f_i({Min[ω_{S_k}] | (ω_{S_i}, ω_{S_k}) ∈ E}) ;
7  |  return Min[W_S];
```

Algorithm 4: Computing $\min\{W_S\}$

Note that Algorithm 4 computes the minimum from scratch each time it is called. In practice, querying the minimum of cost function W_S when x_i is assigned to v for different values v can be done more efficiently with some pre-processing. We define $Min^+[\omega_{S_j}, x_i, v]$ that stores $\min\{\omega_{S_j}(\ell) \mid x_i \in S_j \wedge \ell[x_i] = v\}$. $Min^+[\omega_{S_j}, x_i, v]$ can be computed similarly to Algorithm 4 by using the equation:

$$Min^+[\omega_{S_j}, x_i, v] = \begin{cases} \omega_{S_j}(v), & \text{if } \omega_{S_j} \text{ is a leaf of } T \text{ and } S_j = \{x_i\} \\ \min\{\omega_{S_j}\}, & \text{if } \omega_{S_j} \text{ is a leaf of } T \text{ and } S_j \neq \{x_i\} \\ f_j(\{Min^+[\omega_{S_k}, x_i, v]) \mid & (\omega_{S_i}, \omega_{S_k}) \in E\}), \text{ otherwise} \end{cases}$$

Whenever we have to compute the minimum for $x_i = v$, we simply return $Min^+[W_S, x_i, v]$. Computing $Min^+[W_S, x_i, v]$ is equivalent to running Algorithm 4 nd times, where n is the number of variables and d the maximum domain size. However, this can be reduced by incremental computations exploiting the global constraint semantics, as illustrated on the W-Grammar^{var} global cost function in Section 5.

We now show that a polynomially DAG-filterable cost function is tractable projection-safe. The following lemma will be useful. For a variable $x \in S$ and a value $v \in D(x)$, we denote as $W'\_GCF(S, A_1, \ldots, A_t)$ the cost function obtained by applying `Project`$(S, \{x\}, (v), \alpha)$ to a global cost function $W\_GCF(S, A_1, \ldots, A_t)$.

Lemma 1. *If a global cost function $W\_GCF(S, A_1, \ldots, A_t)$ is polynomially DAG-filterable, $W'\_GCF(S, A_1, \ldots, A_t)$ is polynomially DAG-filterable.*

Proof. Suppose $W\_GCF(S, A_1, \ldots, A_t)$ is polynomially DAG-filterable. Then any instance W_S of it has a safe filtering DAG $T = (V, E)$. By Theorem 5, we know that W'_S, the corresponding instance of $W'\_GCF(S, A_1, \ldots, A_t)$, has a safe DAG filter T', which is isomorphic to T, has polynomial size, and polynomial-time computable f_i associated with each internal vertex. The leaves of T' only differ from those of T by a constant. The result follows. □

Theorem 7. *A polynomially DAG-filterable global cost function $W\_GCF(S, A_1, \ldots, A_t)$ is tractable projection-safe.*

Proof. Follows directly from Theorem 6 and Lemma 1. □

As shown by Theorem 7, a polynomially DAG-filterable cost function W_S remains polynomially DAG-filterable after projection or extension. Algorithm 5 shows how the projection is performed from W_S and W_i, where $x_i \in S$. Lines 2 to 4 modify the leaves of the filtering DAG, as suggested by Theorem 5.

Lines 4 to 8 in Algorithm 5 show how incrementality can be achieved. If W_i or $D(x_i)$, $x_i \in S$, are changed we update the entry $Min[\omega_{S_i}]$ at line 4, which corresponds to the leaf ω_{S_i}, where $x_i \in S_i$. The change propagates upwards in lines 7 and 8, updating all entries related to the leaf ω_{S_i}. The table FW can be updated similarly.

```
   Precondition: W_S is polynomially DAG-filterable with the filtering DAG
   T = (V, E);
   Procedure Project (S, {x_i}, (v), α)
1  |  W_i(v) := W_i(v) ⊕ α ;
2  |  foreach ω_{S_j} ∈ V such that S_j = {x_i} and ω_{S_j} is a leaf of T do
3  |  |  ω_{S_j}(v) := ω_{S_j}(v) ⊖ α ;
4  |  |  Min[ω_{S_j}] := min{ω_{S_j}} ;
5  |  Topologically sort V;
6  |  foreach ω_{S_j} ∈ V in reverse topological order do
7  |  |  if ω_{S_j} is not a leaf and x_i ∈ S_j then
8  |  |  |  Min[ω_{S_j}] := f_i({Min[ω_{S_k}] | (ω_{S_j}, ω_{S_k}) ∈ E}) ;
```

Algorithm 5: Projection from a polynomially DAG-filterable global cost function

The time complexity of enforcing GAC* on a polynomially DAG-filterable global cost function heavily depends on preprocessing, as stated in the following corollary.

Corollary 1. *If the time complexity for pre-computing the table Min^+ for a polynomially DAG-filterable cost function W_S is $O(K(n, d))$, where K is a function of $n = |S|$ and*

maximum domain size d, then enforcing GAC on a variable $x_i \in S$ with respect to W_S requires $O(K(n,d)+d)$ time.*

Proof. Computing the minimum of W_S when $x_i = v$, where $x_i \in S$ and $v \in D(x_i)$, requires only constant time by looking up from Min^+. By Proposition 1, the time complexity is $O(K(n,d)+d)$ time. □

We have presented a new class of tractable projection-safe global cost functions. Algorithm 4 gives an efficient algorithm to compute the minimum cost. In the next Section, we give an example of such a global cost function. More examples can be found in the associated technical report [3].

5. A Polynomially DAG-filterable Global Cost Function

In the following, we show that W_GRAMMAR^{var} is polynomially DAG-filterable using the results from the previous section. Other examples of polynomially DAG-filterable global cost functions can be found in the extended version [3].

W_GRAMMAR^{var} is the cost function variant of the softened version of the hard global constraint GRAMMAR [37] defined based on a context-free language.

Definition 19. *A context-free language $L(G)$ is represented by a* context-free grammar *$G = (\Sigma, N, P, A_0)$, where:*

- *Σ is a set of terminals;*
- *N is a set of non-terminals;*
- *P is a set of production rules from N to $(\Sigma \cup N)^*$, where $*$ is the Kleene star, and;*
- *$A_0 \in N$ is a starting symbol.*

A string τ belongs to $L(G)$, written as $\tau \in L(G)$ iff τ can be derived from G.

Without loss of generality, we assume that (1) the context-free language $L(G)$ does not contain cycles, and (2) the strings are always of fixed length, representing values in tuples.

Assume $S = \{x_1, \ldots, x_n\}$. We define τ_ℓ to be a string formed by a tuple $\ell \in D^S$, where the i^{th} character of τ_ℓ is $\ell[x_i]$. The hard constraint GRAMMAR(S, G) authorizes a tuple $\ell \in D^S$ if $\tau_\ell \in L(G)$ [37]. Using the violation measure *var* by Katsirelos *et al.* [38], the W_GRAMMAR^{var} cost function is defined as follows.

Definition 20 (W_GRAMMAR^{var} [38])**.** *Given a context-free grammar $G = (\Sigma, N, P, A_0)$.* W_GRAMMAR$^{var}(S, G)$ *returns* $\min\{H(\tau_\ell, \tau_i) \mid \tau_i \in L(G)\}$ *for each tuple $\ell \in D^S$, where $H(\tau_1, \tau_2)$ returns the Hamming distance between τ_1 and τ_2.*

Example 1. *Consider $S = \{x_1, x_2, x_3, x_4\}$, where $D(x_i) = \{a, b, c\}$ for $i = 1 \ldots 4$. Given the grammar $G = (\{a, b, c\}, \{A_0, A, B, C\}, P, S)$ with the following production rules.*

$$\begin{aligned} A_0 &\rightarrow AA \\ A &\rightarrow a \mid AA \mid BC \\ B &\rightarrow b \mid BB \\ C &\rightarrow c \mid CC \end{aligned}$$

The cost returned by W_GRAMMAR$^{var}(S, G)(\ell)$ *is 1 if $\ell = (c, a, b, c)$. The assignment of x_1 needs to be changed so that $L(M)$ accepts the corresponding string aabc.*

Theorem 8. W_GRAMMAR$^{var}(S, G)$ *is a polynomially DAG-filterable and thus tractable projection-safe global cost function.*

Proof. We adopt the dynamic programming approach similar to the modified CYK parser [38]. Without loss of generality, we assume G is in Chomsky normal form, *i.e.* each production rule always has the form $A \to \alpha$ or $A \to BC$, where $A \in N$, $B, C \in N \setminus \{A_0\}$ and $\alpha \in \Sigma$.

Define $\omega^A_{S_{i,j}} = \text{W\_GRAMMAR}^{var}(S_{i,j}, G_A)$, where $i \leq j$, $S_{i,j} = \{x_i \dots x_j\} \subseteq S$, and $G_A = (\Sigma, N, P, A)$ for $A \in N$. By definition,

$$\text{W\_GRAMMAR}^{var}(S, G)(\ell) = \omega^{A_0}_{S_{1,n}}(\ell)$$

The base cases $\omega^A_{S_{i,i}}$ is defined as follows. Define $\Sigma_A = \{\alpha \mid A \to \alpha\}$ to be the set of terminals that can be yielded from A.

$$\omega^A_{S_{i,i}}(\ell) = \begin{cases} \min\{U^\alpha_i(\ell[x_i]) \mid (A \to \alpha) \in P\}, & \text{if } \Sigma_A \neq \varnothing \\ \top, & \text{otherwise} \end{cases} \tag{1}$$

The unary cost function $U^\alpha_i(\ell[x_i])$ is defined as follows.

$$U^\alpha_i(v) = \begin{cases} 0, & \text{if } v = \alpha; \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

Other cost functions $\omega^A_{S_{i,j}}$, where $i < j$, are defined as follows. Let $N_A = \{(B, C) \mid A \to BC\}$ be the set of pairs of non-terminals that are yielded from A.

$$\omega^A_{S_{i,j}}(\ell) = \begin{cases} \min\limits_{k=i,\dots j-1} \{\omega^B_{S_{i,k}}(\ell[S_{i,k}]) \oplus \omega^C_{S_{k+1,j}}(\ell[S_{k+1,j}]) \mid (A \to BC) \in P\}, & \text{if } N_A \neq \varnothing \\ \top, & \text{otherwise} \end{cases} \tag{3}$$

□

The associated filtering DAG (V, E) is illustrated in Figure 1 on Example 1. In Figure 1, leaves are indicated by double circles, corresponding to the unary cost function in equation 2. Vertices with min or $\oplus$ aggregators are indicated by rectangles and circles respectively, corresponding to cost functions $\omega^A_{S_{i,j}}$ in equation 3 if $i \neq j$, or equation 1 otherwise. As shown in Figure 1, the root node W_GRAMMAR is first split by the production rule $A_0 \to AA$. One of its children $\omega^A_{S_{1,1}}$ leads to the leaf U^a_1 according to the production rule $A \to a$. The DAG uses only $\oplus$ or min as aggregations and they satisfy the preconditions that allow to apply propositions 2 and 3. The cost function is therefore safely DAG-filterable. Moreover, the corresponding DAG (V, E) has size $|V| = O(|P| \cdot |S|^3)$ polynomial in the size of the input. The leaves are unary functions $\{U^\alpha_i\}$ and by Theorem 7, the result follows.

Note that Theorem 8 also gives a proof that W_REGULAR^{var} is tractable projection-safe. Indeed, a finite state automaton, defining a regular language, can be transformed into a grammar with the number of non-terminals and production rules polynomial in the number of states in the automaton. Then, W_AMONG^{var} is also tractable projection-safe since the tuples satisfying an AMONG global constraint can be represented using a compact finite state counting automaton [8].

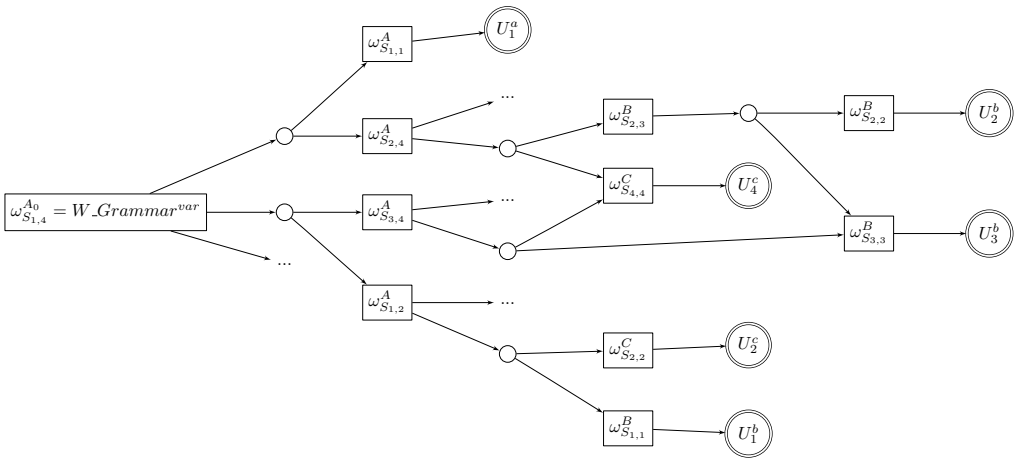

Figure 1: The DAG corresponding to W_Grammar^{var}

Function `GrammarMin` in Algorithm 6 computes the minimum of W_Grammar$^{var}(S, G)$. We first compute the minimum of the unary cost functions in the table $u[i, c]$ at lines 1 to 3. The table f of size $n \times n \times |N|$ is filled up in two separate for-loops: one at line 4 according to the equation 1, and another one at line 14 for the equation 3. The result is returned at line 8.

Theorem 9. *The function* `GrammarMin` *in Algorithm 6 computes the minimum of the global cost function* W_Grammar$^{var}(S, G = (\Sigma, N, P, A_0))$ *in time* $O(nd \cdot |\Sigma| + n^3 \cdot |P|)$, *where* $n = |S|$ *and* d *is the maximum domain size.*

Proof. Lines 1 to 3 take $O(nd \cdot |\Sigma|)$. The first for-loop at lines 4 to 7 requires $O(n \cdot |P|)$, while the second one at lines 9 to 14 requires $O(n^3 \cdot |P|)$. The overall time complexity is $O(nd \cdot |\Sigma| + n \cdot |P| + n^3 \cdot |P|) = O(nd \cdot |\Sigma| + n^3 \cdot |P|)$. □

As for incrementality, Algorithm 7 gives the pre-processing performed on top of Algorithm 6, based on the weighted CYK propagator used in Katsirelos *et al.* [38]. We compute the table f at line 1 using Algorithm 6. Then we compute the table F at lines 8 to 16 using the top-down approach. For each production $A \mapsto A_1 A_2$, lines 14 and 16 compute the maximum possible costs from their neighbors. An additional table $marked[i, j, A]$ is used to record whether the symbol A is accessible when deriving substrings at positions i to j in G. Each time we need to compute the minimum for $x_i = v$, we just return $\min\{U_i^{\alpha}(v) \ominus F[i, i, A] \oplus f[0, n-1, A_0] \mid (A \mapsto v) \in P \wedge marked[i, i, A]\}$, or $\top$ if such production does not exist.

Corollary 2. *Given* $W_S =$ W_Grammar$^{var}(S, G = (\Sigma, N, P, A_0))$. *Enforcing GAC* on a variable* $x_i \in S$ *with respect to* W_Grammar *requires* $O(nd \cdot |\Sigma| + n^3 \cdot |P|)$ *time, where* $n = |S|$ *and* d *is the maximum domain size.*

Proof. Using a similar argument to that in the proof of Theorem 9, Algorithm 7 requires $O(nd \cdot |\Sigma| + n^3 \cdot |P|)$ time. The result follows directly from Corollary 1 and Theorem 9. □

```
   Function GrammarMin(S, G)
1  |  for i := 1 to n do
2  |  |  for c ∈ Σ do
3  |  |  |  u[i, c] := min{U_i^c};
4  |  for i := 1 to n do
5  |  |  foreach A ∈ N do f[i, i, A] := ⊤;
6  |  |  foreach (A, a) such that (A ↦ a) ∈ P do
7  |  |  |  f[i, i, A] = min{f[i, i, A], u[i, a]} ;
8  |  return GrammarPartialMin(S, G, 1);

   Function GrammarPartialMin(S, G, start)
9  |  for len := 2 to n do
10 |  |  for i := start to n − len + 1 do
11 |  |  |  j := i + len − 1 ;
12 |  |  |  foreach A ∈ N do f[i, j, A] := ⊤;
13 |  |  |  foreach (A, A_1, A_2) such that (A ↦ A_1 A_2) ∈ P do
14 |  |  |  |  for k := i to j − 1 do
15 |  |  |  |  |  f[i, j, A] := min{f[i, j, A], f[i, k, A_1] ⊕ f[k + 1, j, A_2]} ;
16 |  return f[1, n, A_0];
```

Algorithm 6: Finding the minimum of W_GRAMMAR^{var}

Algorithm 8 shows how projection is performed between $W\_Grammar^{var}$ and W_p, and how incrementally can be achieved. Line 3 modifies the leaves U_p^c for each $c \in \Sigma$, while lines 4 and 5 update the corresponding entries in the tables u and f respectively. The change is propagated up in f at line 6, corresponding to derivation of sub-strings with positions from p to the end in G.

In this section, we have seen how the minimum of a polynomially DAG-filterable global cost function can be computed efficiently, leading to efficient soft local consistency enforcement. However, each newly implemented cost function requires to build a corresponding DAG structure with a dedicated dynamic programming algorithm.

In the next section, we show that, in some cases, it is also possible to avoid this by directly decomposing a global cost functions into a CFN in such a way that local consistency enforcement will emulate dynamic programming, avoiding the need for dedicated enforcement algorithms.

6. Decomposing Global Cost Functions into CFNs

In CSPs, some global constraints can be efficiently represented by a logically equivalent subnetwork of constraints of bounded arities [16, 13], and are said to be decomposable. Similarly, we will show that some global cost functions can be encoded as a sum of bounded arity cost functions. The definition below applies to any cost function, including constraints, extending the definition in [16] and [13].

Procedure `GrammarPreCompute`(S, G)

1 $F[1, n, A_0] :=$ `GrammarMin`(S, G);
2 **for** $i := 1$ **to** n **do**
3 **for** $j := i$ **to** n **do**
4 **foreach** $A \in N$ **do**
5 $F[i, j, A] := -\top$;
6 $marked[i, j, A] :=$ **false**;
7 $marked[1, n, A_0] :=$ **true**;
8 **for** $len := n$ **down to** 2 **do**
9 **for** $i := 1$ **to** $n - len + 1$ **do**
10 $j := i + len - 1$;
11 **foreach** (A, A_1, A_2) such that $(A \mapsto A_1 A_2) \in P \wedge marked[i, j, A]$ **do**
12 **for** $k := i$ **to** j **do**
13 $marked[i, k, A_1] :=$ **true**;
14 $F[i, k, A_1] :=$ `max`$(F[i, k, A_1], F[i, j, A] \ominus f[k + 1, j, A_2])$;
15 $marked[k + 1, j, A_2] :=$ **true**;
16 $F[k + 1, j, A_2] :=$ `max`$(F[k + 1, j, A_2], F[i, j, A] \ominus f[i, k, A_1])$;

Algorithm 7: Pre-computation for W_Grammar^{var}

Definition 21. *For a given integer p, a* p-network-decomposition *of a global cost function $W\_GCF(S, A_1, \ldots, A_k)$ is a polynomial transformation δ_p that returns a CFN $\delta_p(S, A_1, \ldots, A_k) = (S \cup E, \mathcal{F}, \top)$, where $S \cap E = \varnothing$, such that $\forall W_T \in \mathcal{F}, |T| \leq p$ and $\forall \ell \in D^S, W\_GCF(S, A_1, \ldots, A_k)(\ell) = \min_{\ell' \in D^{S \cup E}, \ell'[S] = \ell} \bigoplus_{W_{S_i} \in \mathcal{F}} W_{S_i}(\ell'[S_i])$.*

Definition 21 above allows for the use of extra variables E, which do not appear in the original cost function scope and are eliminated by minimization. We assume, without loss of generality, that every extra variable $x \in E$ is involved in at least two cost functions in the decomposition.[1] Clearly, if $W\_GCF(S, A_1, \ldots, A_k)$ appears in a CFN $P = (\mathcal{X}, \mathcal{W}, \top)$ and decomposes into $(S \cup E, \mathcal{F}, \top)$, the optimal solutions of P can directly be obtained by projecting the optimal solutions of the CFN $P' = (\mathcal{X} \cup E, \mathcal{W} \setminus \{W\_GCF(S, A_1, \ldots, A_k)\} \cup \mathcal{F}, \top)$ on $\mathcal{X}$.

6.1. Building network-decomposable global cost functions

A global cost function can be shown to be network-decomposable by exhibiting a bounded arity network decomposition of the global cost function. There is a simple way of deriving network-decomposable cost functions from known decomposable global constraints. The process goes directly from a known decomposable global constraint to a network-decomposable global cost function and does not require to use an intermediate

[1] Otherwise, such a variable can be removed by variable elimination: remove x from E and replace the W_T involving x by the cost function $\min_x W_T$ on $T \setminus \{x\}$. This preserves the Berge-acyclicity of the network if it exists.

Procedure *GrammarProject* $(S, \{x_p\}, (v), \alpha)$

$W_p(v) := W_p(v) \oplus \alpha$;
for $c \in \Sigma$ **do**
  $U_p^c(v) := U_p^c(v) \ominus \alpha$;
  $u[p, c] := \min\{U_p^c\}$;
foreach (A, a) such that $(A \mapsto a) \in P$ **do** $f[p, p, A] = \min\{f[p, p, A], u[i, a]\}$;
GrammarPartialMin (S, G, p);
GrammarPreCompute (S, G);

Algorithm 8: Projection from W_GRAMMAR$^{var}(S, G = (\Sigma, N, P, A_0))$

soft global constraint with an associated violation measure μ. Instead, the global cost function will use any relaxation of the decomposed global constraint.

We say that the cost function W_S is a *relaxation* of W'_S if for all $\ell \in D^S, W_S(\ell) \leq W'_S(\ell)$. We then write $W_S \leq W'_S$. From a network-decomposable global constraint, it is possible to define an associated network-decomposable global cost function by relaxing every constraint in the decomposition.

Theorem 10. *Let $GC(S, A_1, \ldots, A_k)$ be a global constraint that p-network decomposes into a classical constraint network $(S \cup E, \mathcal{F}, \top)$ and f_θ be a function parameterized by θ that maps every $C_T \in \mathcal{F}$ to a cost function $f_\theta(C_T)$ such that $f_\theta(C_T) \leq C_T$. The global cost function*

$$W\_GCF(S, A_1, ..., A_k, f_\theta)(\ell) = \min_{\substack{\ell' \in D^{S \cup E} \\ \ell'[S] = \ell}} \bigoplus_{C_T \in \mathcal{F}} f_\theta(C_T)(\ell'[T])$$

is a relaxation of $GC(S, A_1, \ldots, A_n)$, and is p-network-decomposable by construction.

Proof. Since $(S \cup E, \mathcal{F})$ is a network-decomposition of $GC(S, A_1, ..., A_k)$, for any tuple $\ell \in D^S$, $GC(S, A_1, ..., A_k)(\ell) = 0$ if and only if $\min_{\ell' \in D^{S \cup E}, \ell'[S] = \ell} \bigoplus_{C_T \in \mathcal{F}} C_T(\ell'[T]) = 0$. Let $\ell' \in D^{S \cup E}$ be a tuple where this minimum is reached. This implies that $\forall C_T \in \mathcal{F}$, $C_T(\ell'[T]) = 0$. Since $f_\theta(C_T) \leq C_T$, $f_\theta(C_T)(\ell'[T]) = 0$. Therefore $\bigoplus_{C_T \in \mathcal{F}} f_\theta(C_T)(\ell'[T]) = 0$ and $W\_GCF(S, A_1, \ldots, A_k, f_\theta)(\ell) = 0$. Moreover, the global cost function is p-network-decomposable by construction. □

Theorem 10 allows to immediately derive a long list of network decomposable global cost functions from existing network decompositions of global constraints such as ALLDIFFERENT, REGULAR [55], AMONG and STRETCH [14]. The parameterization through f_θ also allows a lot of flexibility.

Example 2. *Consider the softened variant* W_ALLDIFFERENT$^{dec}(S)$ *of the global constraint* ALLDIFFERENT(S) *constraint using the* decomposition *violation measure where the cost of an assignment is the number of pairs of variables taking the same value [56]. It is well known that* ALLDIFFERENT *decomposes into a set of $\frac{n.(n-1)}{2}$ binary difference constraints. Similarly, the* W_ALLDIFFERENT$^{dec}(S)$ *cost function can be decomposed into a set of $\frac{n.(n-1)}{2}$ soft difference cost functions. A soft difference cost function takes*

cost 1 iff the two involved variables have the same value and 0 otherwise. In these cases, no extra variable is required.

AllDifferent *can be softened in a different way. Take an arbitrary graph* $G = (V, E)$ *over* V*, and consider the violation measure where the cost of an assignment is the number of pairs of variables in* E *taking the same value. This gives rise to a global cost function* W_AllDifferent$^{f_G}(V)$ *that allows a zero cost assignment if and only if* G *is colorable, which is an NP-hard problem. Enforcing any soft arc consistency on that single global cost function will be intractable as well since it requires to compute the minimum of the cost function. Instead, enforcing soft arc consistencies on the network-decomposition into binary cost functions will obviously be polynomial but will achieve a lower level of filtering.*

7. Local Consistency and Network-Decompositions

As we have seen with the W_AllDifferent(V, f_G) global cost function, the use of network-decompositions instead of a monolithic variant has both advantages and drawbacks. Thanks to local reasoning, a decomposition may be filtered more efficiently, but this may hinder the level of filtering achieved. In CSP, it was observed that the structure of the decomposition has an impact on the level of consistency achieved when filtering the decomposition.

Before going further, we give some extra definitions that are useful to characterize structure of decompositions. The hypergraph (X, E) of a CFN $(\mathcal{X}, \mathcal{W}, \top)$ has one vertex per variable $x_i \in \mathcal{X}$ and one hyperedge for every scope S such that $\exists W_S \in \mathcal{W}$. The incidence graph of a hypergraph (X, E) is a bipartite graph $G = (X \cup E, E_H)$ where $\{x_i, e_j\} \in E_H$ iff $x_i \in X, e_j \in E$ and x_i belongs to the hyperedge e_j. A hypergraph (X, E) is Berge-acyclic iff its incidence graph is acyclic.

In CSP, it is known that if the decomposition is Berge-acyclic, then enforcing GAC on the decomposition enforces GAC on the global constraint itself [6]. We now show that a similar result can be obtained for cost functions using either a variant of Directional Arc Consistency or Virtual Arc Consistency (VAC), whose definitions are given in the two subsections below.

7.1. Berge-acyclicity and directional arc consistency

In this section, we will show that enforcing directional arc consistency on a Berge-acyclic network-decomposition of a cost function or on the original global cost function yields the same cost distribution on the last variable and therefore the same lower bound (obtained by node consistency) provided a correct variable ordering is used.

Directional Arc Consistency has been originally defined on binary networks. We define Terminal DAC (or T-DAC) which generalizes Directional Arc Consistency [21] by removing the requirement of having binary scopes.

Definition 22 (T-DAC)**.** *Given a CFN* $N = (\mathcal{X}, \mathcal{W}, \top)$ *a total order* $\prec$ *over variables:*

- *For a cost function* $W_S \in \mathcal{W}^+$*, a tuple* $\ell \in D^S$ *is a full support for a value* $a \in D(x_i)$ *of* $x_i \in S$ *iff* $W_S(\ell) \bigoplus_{x_j \in S, j \neq i} W_j(\ell[x_j]) = 0$.
- *A variable* $x_i \in S$ *is* star directional arc consistent (DAC*) *for* W_S *iff*
  - x_i *is NC*;*

– *each value $v_i \in D(x_i)$ has a full support ℓ for W_S.*

- *N is* Terminal Directional Arc Consistent *(T-DAC) w.r.t. the order $\prec$ iff for all cost functions $W_S \in \mathcal{W}^+$, the minimum variable in S is DAC* for W_s.*

To enforce T-DAC on a cost function W_S, it suffices to first shift the cost of every unary cost function $W_i, i \in S$ inside W_S by applying $\texttt{Project}(S, \{x_i\}, (a), -W_i(a))$ for every value $a \in D_i$. Let x_j be the minimum variable in S according to $\prec$, one can then apply $\texttt{Project}(S, \{x_j\}, (b), \alpha)$ for every value $b \in D(x_j)$ with $\alpha = \min_{\ell \in D^S, \ell[x_j]=b} W_S(\ell)$. Let ℓ be a tuple where this minimum is reached. Then either $\alpha = \top$ and the value will be deleted, or ℓ is a full support for $b \in D(x_j)$: $W_S(\ell) \bigoplus_{x_i \in S, i \neq j} W_i(\ell[x_i]) = 0$. This support can only be broken if for some unary cost function $W_i, i \in S, i \neq j$, $W_i(a)$ increases for some value $a \in D(x_i)$. Since j is minimum, $i \succ j$.

To enforce T-DAC on a CFN $(\mathcal{X}, \mathcal{W}, \top)$, one can simply sort $\mathcal{W}$ in a *decreasing* order of the minimum variable in the scope of each cost function, and apply the previous process on each cost function, successively. When a cost function W_S is processed, all the cost functions whose minimum variable is larger than the minimum variable of S have already been processed, which guarantees that none of the established full supports will be broken in the future. Enforcing T-DAC is therefore in $O(ed^r)$ in time, where $e = |\mathcal{W}|$ and $r = \max_{W_S \in \mathcal{W}} |S|$.

Theorem 11. *If a global cost function* W_GCF$(S, A_1, \ldots, A_k)$ *decomposes into a Berge-acyclic CFN $N = (S \cup E, \mathcal{F})$, there exists an ordering on $S \cup E$ such that the unary cost function $W_{x_{i_n}}$ on the last variable x_{i_n} of S produced by enforcing T-DAC on the sub-network $(S, \{W\_GCF(S, A_1, \ldots, A_k)\} \cup \{W_{x_i}\}_{x_i \in S})$ is identical to the unary cost function $W'_{x_{i_n}}$ produced by enforcing T-DAC on the decomposition $N = (S \cup E, \mathcal{F} \cup \{W_{x_i}\}_{x_i \in S})$.*

Proof. Consider the decomposed network N and $I_N = (S \cup E \cup \mathcal{F}, E_I)$ its incidence graph. As N is Berge-acyclic we know that I_N is a tree whose vertices are the variables and the cost functions of N. We root I_N in a variable of S. The neighbors (parent and children, if any) of cost functions W_T are the variables in T. The neighbors of a variable x_i are the cost functions involving x_i. Consider any topological ordering of the vertices of I_N. This ordering induces a variable ordering $(x_{i_1}, \ldots, x_{i_n}), x_{i_n} \in S$ which is used to enforce T-DAC on N. Notice that for any cost function $W_T \in \mathcal{F}$, the parent variable of W_T in I_N appears after all the other variables of T.

Consider a value $a \in D(x_{i_n})$ of the root. Since NC* is enforced, $W_{x_{i_n}}(a) < \top$. Let W_T be any child of x_{i_n} and ℓ a full support of value a on W_T. We have $W_{x_{i_n}}(a) = W_T(\ell) \bigoplus_{x_i \in T} W_{x_i}(\ell[x_i])$, which proves that $W_T(\ell) = 0$ and $\forall x_i \in T, i \neq i_n, W_{x_i}(\ell[x_i]) = 0$. I_N being a tree, we can inductively apply the same argument on all the descendants of x_{i_n} until leaves are reached, proving that the assignment $(x_{i_n} = a)$ can be extended to a complete assignment with cost $W_{x_{i_n}}(a)$ in N. In both cases, $W_{x_{i_n}}(a)$ is the cost of an optimal extension of $(x_{i_n} = a)$ in N.

Suppose now that we enforce T-DAC using the previous variable ordering on the undecomposed sub-network $(S, \{W\_GCF(S, A_1, \ldots, A_k)\} \cup \{W_{x_i}\}_{x_i \in S})$. Let ℓ be a full support of value $a \in D(x_{i_n})$ on W_GCF$(S, A_1, \ldots, A_k)$. By definition, $W_{x_{i_n}}(a) = W\_GCF(S, A_1, \ldots, A_k)(\ell) \bigoplus_{x_i \in S} W_{x_i}(\ell[x_i])$ which proves that $W_{x_{i_n}}(a)$ is the cost of an optimal extension of $(x_{i_n} = a)$ on $(S, \{W\_GCF(S, A_1, \ldots, A_k)\} \cup \{W_{x_i}\}_{x_i \in S})$. By definition of decomposition, and since $x_{i_n} \notin E$, this is equal to the cost of an optimal extension of $(x_{i_n} = a)$ in N. □

T-DAC has therefore enough power to handle Berge-acyclic network-decompositions without losing any filtering strength, provided a correct order is used for applying EPTs. In this case, T-DAC emulates a simple form of dynamic programming on the network-decomposition.

Example 3. *Consider the* Regular $(\{x_1, \ldots, x_n\}, M)$ *global constraint, defined by a (not necessarily deterministic) finite automaton* $M = (Q, \Sigma, \delta, q_0, F)$*, where* Q *is a set of states,* Σ *the emission alphabet,* δ *a transition function from* $\Sigma \times Q \to 2^Q$*,* q_0 *the initial state and* F *the set of final states. As shown in [15], this constraint decomposes into a constraint network* $(\{x_1, \ldots, x_n\} \cup \{Q_0, \ldots, Q_n\}, C)$ *where the extra variables* Q_i *have* Q *as their domain. The set of constraints* C *in the network decomposition contains two unary constraints restricting* Q_0 *to* $\{q_0\}$ *and* Q_n *to* F *and a sequence of identical ternary constraints* $c_{\{Q_i, x_{i+1}, Q_{i+1}\}}$ *each of which authorizes a triple* (q, s, q') *iff* $q' \in \delta(q, s)$*, thus capturing* δ*. A relaxation of this decomposition may relax each of these constraints. The unary constraints on* Q_0 *and* Q_n *would be replaced by unary cost functions* λ_{Q_0} *and* ρ_{Q_n} *stating the cost for using every state as either an initial or final state while the ternary constraints would be relaxed to ternary cost functions* $\sigma_{\{Q_i, x_{i+1}, Q_{i+1}\}}$ *stating the cost for using any* (q, s, q') *transition.*

This relaxation precisely corresponds to the use of a weighted automaton $M_W = (Q, \Sigma, \lambda, \sigma, \rho)$ *where every transition, starting and finishing state has an associated, possibly intolerable, cost defined by the cost functions* λ, σ *and* ρ *[24]. The cost of an assignment in the decomposition is equal, by definition, to the cost of an optimal parse of the assignment by the weighted automaton. This defines a* W_Regular(S, M_W) *global cost function which is parameterized by a weighted automaton. As shown in [38], a weighted automaton can encode the Hamming and Edit distances to the language of a classical automaton. We observe that the hypergraph of the decomposition of* W_Regular *is Berge-acyclic. Thus, contrary to the* AllDifferent *example, where decomposition was hindering filtering, T-DAC on the* W_Regular *network-decomposition achieves T-DAC on the original cost function.*

It should be pointed out that T-DAC is closely related to mini-buckets [26] and Theorem 11 can easily be adapted to this scheme. Mini-buckets perform a weakened form of variable elimination: when a variable x is eliminated, the cost functions linking x to the remaining variables are partitioned into sets containing at most i variables in their scopes and at most m functions (with arity > 1). If we compute mini-buckets using the same variable ordering, with $m = 1$ and unbounded i, we will obtain the same unary costs as T-DAC on the root variable r, with the same time and space complexity. Mini-buckets can be used along two main recipes: precomputed (static) mini-buckets do not require update during search but restrict search to one static variable ordering; dynamic mini-buckets allow for dynamic variable ordering (DVO) but suffer from a lack of incrementality. Soft local consistencies, being based on EPTs, always yield equivalent problems, providing incrementality during search and are compatible with DVO.

7.2. Berge-acyclicity and virtual arc consistency

Virtual Arc Consistency offers a simple and direct link between CSPs and CFNs which allows to directly lift CSP properties to CFNs, under simple conditions.

Definition 23 (VAC [19]). *Given a CFN $N = (\mathcal{X}, \mathcal{W}, \top)$, we define the constraint network $Bool(N)$ as the CSP with the same set $\mathcal{X}$ of variables with the same domains, and which contains, for each cost function $W_S \in W, |S| > 0$, a constraint c_S with the same scope, which exactly forbids all tuples $\ell \in D^S$ such that $W_S(\ell) \neq 0$. A CFN N is said to be Virtual Arc Consistent (VAC) iff the arc consistent closure of the constraint network $Bool(N)$ is non empty.*

Theorem 12. *If a global cost function* W_GCF$(S, A_1, \ldots, A_k)$ *decomposes into a Berge-acyclic CFN $N = (S \cup E, \mathcal{F}, \top)$ then enforcing VAC on either $(S \cup E, \mathcal{F} \cup \{W_{x_i}\}_{x_i \in S}, \top)$ or on $(S, \{W\_GCF(S, A_1, \ldots, A_k)\} \cup \{W_{x_i}\}_{i \in S}, \top)$ yields the same lower bound $W_\varnothing$.*

Proof. Enforcing VAC on the CFN $N = (S \cup E, \mathcal{F} \cup \{W_{x_i}\}_{x_i \in S}, \top)$ does not modify the set of scopes as it only performs 1-EPTs (See Definition 4). Hence it yields an equivalent problem N' such that $Bool(N')$ has the same hypergraph as $Bool(N)$. Since N has a Berge acyclic structure, this is also the case for $Bool(N)$ and $Bool(N')$. Now, Berge-acyclicity is a situation where arc consistency is a decision procedure. We can directly make use of Proposition 10.5 of [19], which states that if a CFN N is VAC and $Bool(N)$ is in a class of CSPs for which arc consistency is a decision procedure, N has an optimal solution of cost $w_\varnothing$.

Similarly, the network $Q = (S, \{W\_GCF(S, A_1, \ldots, A_k)\} \cup \{W_{x_i}\}_{x_i \in T}, \top)$ contains just one cost function with arity strictly above 1 and $Bool(Q)$ will be decided by arc consistency. Enforcing VAC will therefore provide a CFN which also has an optimal solution of cost $W_\varnothing$. Finally, the networks N and Q have the same optimal cost by definition of a decomposition. □

Given that VAC is both stronger and more expensive to enforce than DAC*, the added value of this theorem, compared to theorem 11, is that it does not rely on a variable ordering. Such order always exists but it is specific to each global cost function. Theorem 12 becomes interesting when a problem contains several global cost functions with intersecting scopes, for which theorem 11 may produce inconsistent orders.

8. Relation between DAG-filterability and Network-Decompositions

In this section, we show that Berge-acyclic network-decomposable global cost functions are also polynomially DAG-filterable.

Theorem 13. *Let $W\_GCF(S, A_1, \ldots, A_k)$ be a network-decomposable global cost function that decomposes into a CFN $(S \cup E, \mathcal{F}, \top)$ with a Berge-acyclic hypergraph. Then* W_GCF$(S, A_1, \ldots, A_k)$ *is polynomially DAG-filterable.*

Proof. We consider the incidence graph of the Berge-acyclic hypergraph of the CFN $(S \cup E, \mathcal{F}, \top)$ and choose a root for it in the original variables S, defining a rooted tree denoted as I. This root orients the tree I with leaves being variables in S and E. In the rest of the proof, we denote by $I(x_i)$ the subtree of I rooted in $x_i \in S \cup E$. Abusively, when the context is clear, $I(x_i)$ will also be used to denote the set of all variables in the subtree.

The proof is constructive. We will transform I into a filtering DAG (actually a tree) of nodes that computes the correct cost $\min_{\ell' \in D^{S \cup E}, \ell'[S] = \ell} \bigoplus_{W_T \in \mathcal{F}} W_T(\ell'[T])$ and

satisfies all the required properties of polynomial DAG-filters. To achieve this, we need to guarantee that the aggregation function $f_i = \oplus$ is always used on cost functions of disjoint scopes, that $f_i = \min$ is always applied on identically scoped functions and that sizes remain polynomial.

We will be using three types of DAG nodes. A first type of node will be associated with every cost function $W_T \in \mathcal{F}$ in the network-decomposition. Each cost function appears in I with a parent variable x_i and a set of children variables among which some may be leaf variables. By the assumption that extra variables belong to at least two cost functions (see paragraph below Definition 21), leaf variables necessarily belong to S. We denote by $\mathit{leaf}(T)$ the set of leaf variables in the scope T. The first type of node aims at computing the value of the cost function W_T combined with the unary cost functions on each leaf variable. This computation will be performed by a family of nodes U_T^ℓ, where $\ell \in D^{T-\mathit{leaf}(T)}$ is an assignment of non-leaf variables. Therefore, for a given cost function W_T and a given assignment ℓ of non-leaf variables, we define a DAG node with scope $leaf(T)$:

$$U_T^\ell(\ell') = W_T(\ell \cup \ell') \bigoplus_{x_j \in \mathit{leaf}(T)} W_{x_j}(\ell'[x_j])$$

These nodes will be leaf nodes of the filtering DAG. Given that all cost functions in I have bounded arity, these nodes have an overall polynomial size and can be computed in polynomial time in the size of the input global cost function.

Nodes of the second and third types are associated to every non-leaf variable x_i in I. For every value $a \in D(x_i)$, we will have a node ω_i^a with scope $I(x_i) \cap S$. x_i may have different children cost functions in I and we denote by $\mathcal{W}_i$ the set of all the children cost functions of x_i in I. For each $W_T \in \mathcal{W}_i$, we will also have a DAG node $\omega_T^{i,a}$ with scope $S'_i = (I(W_T) \cup \{x_i\}) \cap S$. Notice that even if these scopes may be large (ultimately equal to S for ω_i^a if x_i is the root of I), these nodes are not leaf nodes of the filtering DAG and do not rely on an extensional definition, avoiding exponential space.

The aim of all these nodes is to compute the cost of an optimal extension of the assignment ℓ to the subtree $I(W_T)$ (for $\omega_T^{i,a}$) or $I(x_i)$ (for ω_i^a). We therefore define:

$$\omega_i^a(\ell) = \bigoplus_{W_T \in \mathcal{W}_i} \omega_T^{i,a}(\ell[S'_i])$$

Indeed, if $\omega_T^{i,a}$ computes the cost of an optimal extension to the subtree rooted in W_T, an optimal extension to $I(x_i)$ is just the $\oplus$ of each optimal extension on each child, since the scopes S'_i do not intersect (I is a tree). The DAG node uses the $\oplus$ aggregation operator on non-intersecting scopes.

The definition of the DAG nodes $\omega_T^{i,a}$ is more involved. It essentially requires:

1. to combine the cost of W_T with the unary cost functions on leaf variables in T (this is achieved by U_T nodes) and costs of optimal extensions subtrees rooted in other non-leaf variables (this is achieved by ω_j^b nodes).
2. to eliminate in this function all extra variables in the scope T except x_i if $x_i \in E$. In this case, x_i's value will be set in ℓ and eliminated on higher levels.

If $x_i \in E$ or else if $\ell[x_i] = a$, this leads to the following definition of $\omega_T^{i,a}(\ell)$:

$$\min_{\substack{\ell' \in D^{T \cap E} \\ (x_i \in S \vee \ell'[x_i] = a)}} \left[U_T^{(\ell \cup \ell')[T - leaf(T)]}(\ell[leaf(T)]) \bigoplus_{x_j \in (T - leaf(T) - \{x_i\})} \omega_j^{\ell[x_j]}(\ell[S_j]) \right] \quad (4)$$

Otherwise ($x_i \in S$ and $\ell[x_i] \neq a$), $\omega_T^{i,a}(\ell) = \top$. This captures the fact that there is no optimal extension of ℓ that extends (x_i, a) since ℓ is inconsistent with $x_i = a$.

If we consider the root variable $x_i \in S$ of I, the ω_i^a nodes provide the cost of a best extension of any assignment ℓ (if $\ell[x_i] = a$) or $\top$ otherwise. An ultimate root DAG node using the aggregation operator min over all these ω_i^a will therefore return the optimal extension of $\ell \in D^S$ to all variables in $I(x_i)$, including extra variables.

From equation 4, one can see that nodes $\omega_T^{i,a}$ use the aggregation operator min on intermediary nodes. These intermediary nodes combine the node U_T and ω_j with $\oplus$ which have non-intersecting scopes.

Overall all those nodes form a DAG (actually a tree). In this tree, every node with the aggregation operation $\oplus$ is applied to operands with non-intersecting scopes, as required in Property 2. Similarly, every node with the min aggregation operation is applied to functions whose scope is always identical, as required by Property 3. Note that the definitions of the ω_i^a and $\omega_T^{i,a}$ are linear respectively in the number of children of W_T or x_i respectively. So, we have a filtering DAG satisfying Definition 18. □

For a global cost function which is Berge-acyclic network-decomposable, and therefore also polynomially DAG-filterable (as Theorem 13 shows), a natural question is which approach should be preferred. The main desired effect of enforcing local consistencies is that it may increase the lower bound $W_\varnothing$. From this point of view, Theorems 11 and 12 give a clear answer for a single global cost function.

- Since OSAC [19] is the strongest form of arc consistency (implying also VAC), the strongest possible lower bound will be obtained by enforcing OSAC on the network-decomposed global cost function. The size of the OSAC linear program being exponential in the arity of the cost functions, the bounded arities of the network decomposed version will define a polynomial-size linear program. This however requires an LP solver.
- If a network containing network-decomposed global cost functions is VAC, the underlying global cost functions are also VAC. As a result, good quality lower bounds can be obtained by enforcing VAC. These lower bounds are not as good as those obtained by OSAC, but VAC is usually much faster than OSAC.
- T-DAC is otherwise extremely efficient, easy to implement, offering good lower bounds and incrementality for little effort. However, when several global cost functions co-exist in a problem, a variable order that is a topological sort of all these global cost functions may not exist. In this case, using a topological order for each scope independently would lead to the creation of cycles leading to possibly infinite propagation. It may then be more attractive to use filtering DAGs to process these cost functions.

Finally, it should be noted that Theorem 11 only guarantees that T-DAC on a global cost function or its topologically sorted Berge-acyclic network-decomposition provide

the same bound contribution. If a consistency stronger than DAC* is enforced (such as FDAC* or EDAC*), it may be more powerful when enforcedt on the global cost function itself than on its network-decomposition, thus giving an advantage to filtering DAGs.

In the end, the only truly informative answer will be provided by experimental results, as proposed in Section 9.

9. Experiments

In this section, we put theory into practice and demonstrate the practicality of the transformations described in the previous sections in solving over-constrained and optimization problems. We implemented cost functions with our transformations in `toulbar2` v0.9.8[2]. For each cost function used in our benchmark problems, we implemented weak Existential Directional Generalized Arc Consistency (EDGAC*) [32, 47, 48], a local consistency combining AC, DAC and EAC, using DAG-filtering (called *DAG-based* approach in the sequel) with pre-computed tables (as described in Section 4). When possible, we also implemented a Berge-acyclic network-decomposition to be propagated using EDGAC* (called *network-based* approach). We ignore weaker forms of local consistency such as Arc Consistency or 0-inverse consistency [65] as previous experiments with global cost functions have shown that these weak local consistencies lead to much less efficient solving [48].

In the experiments, we used default options for `toulbar2`, including a new hybrid best-first search strategy introduced in [5], which finds good solutions more rapidly compared to classical depth-first search. The default variable ordering strategy is dom/wdeg [17] with Last Conflict [45], while the default value ordering consists, for each variable, in choosing first its *fully supported value* as defined by EDGAC*. At each node during search, including the root node, we eliminate dominated values using Dead End Elimination pruning [27, 44, 33] and we eliminate all variables having degree less than two using variable elimination [12, 40]. At the root node only, this is improved by pairwise decomposition [30] and we also eliminate all variables having a functional or bijective binary relation (*e.g.,* an equality constraint) with another variable. The tests are conducted on a single core of an Intel Xeon E5-2680 (2.9GHz) machine with 256GB RAM.

We performed our experiments on four different benchmark problems. For the two first benchmarks (car sequencing and nonogram), we have a model with Berge-acyclic network-decompositions, whereas for the two others (well-formed parentheses and market split), we do not. Each benchmark has a 5-minute timeout. We randomly generate 30 instances for each parameter setting of each benchmark. We first compare the number of solved instances, *i.e.* finding the optimum and proving its optimality (no initial upper bound). We report the average run-time in seconds and consider that an unsolved problem requires the maximum available time (timeout). When all instances are solved, we also report the average number of backtracks (or '–' otherwise). The best results are marked in bold (taking first into account the number of solved instances in less than 5 minutes and secondly CPU time).

[2] `http://www.inra.fr/mia/T/toulbar2/`

9.1. The Car Sequencing Problem

The car sequencing problem (prob001 in CSPLib, [54]) requires sequencing n cars of different types specified by a set of options. For any subsequence of c_i consecutive cars on the assembly line, the option o_i can be installed on at most m_i of them. This is called the *capacity* constraint. The problem is to find a production sequence on the assembly line such that each car can be installed with all the required options without violating the capacity constraint. We use n variables with domain 1 to n to model this problem. The variable x_i denotes the type of the i^{th} car in the sequence. One GCC (global cardinality [52]) constraint ensures all cars are scheduled on the assembly line. We post $n - c_i + 1$ AMONG constraints [10] for each option o_i to ensure the capacity constraint is not violated. We randomly generate 30 over-constrained instances, each of which has 5 possible options, and for each option o_i, m_i and c_i are randomly generated in such a way that $1 \leq m_i < c_i \leq 7$. Each car in each instance is randomly assigned to one type, and each type is randomly assigned to a set of options in such a way that each option has $1/2$ chance to be included in each type. To introduce costs, we randomly assign unary costs (between 0 to 9) to each variable.

The problem is then modeled in three different ways. The first model is obtained by replacing each AMONG constraint by the W_AMONG^{var} cost function and the GCC constraint by the W_GCC^{var} cost function. W_AMONG^{var} returns a cost equal to the number of variables that need to be re-assigned to satisfy the AMONG constraint. W_GCC^{var} is used as a global constraint and returns $\top$ on violation [48]. This model is called "flow&DAG-based" approach in Table 1.

The second model, identified as "DAG-based" in Table 1, uses a set of W_AMONG^{var} cost functions to encode GCC, *i.e.* replacing the single global cost function exploiting a flow network by a set of DAG-based global cost functions [3].

In the third model, identified as "network-based" in Table 1, each of the W_AMONG^{var} in the previous DAG-based model is decomposed into a set of ternary cost functions with extra variables as described in Section 6.

Table 1 gives the experimental results. Column n' indicates the sum of the number of original variables (n) and the number of extra variables added in the network-based approach. Column n'' gives the total number of unassigned variables after pre-processing. We observe that the network-based approach performed the worst among the three approaches. The DAG-based approach is up to six times faster than the flow&DAG-based approach on completely solved instances ($n \leq 13$) and solves more instances within the 5-minute time limit. Surprisingly, it also develops the least number of backtracks on completely solved instances. We found that the initial lower bound produced by weak EDGAC on the flow&DAG-based approach can be lower than the one produced by the DAG-based approach. This is due to different orders of EPTs done by the two approaches resulting in different lower bounds. Finding an optimal order of integer arc-EPTs is NP-hard [20]. Recall that EDGAC has a chaotic behavior compared to OSAC or VAC and encoding GCC into a set of W_AMONG^{var} will produce more EPTs (each W_AMONG^{var} moving unary costs differently) creating new opportunities for the overlapping W_AMONGs^{var} to deduce a better lower bound.

9.2. The Nonogram Problem

The nonogram problem (prob012 in CSPLIB [36]) is a typical board puzzle on a board of size $p \times p$. Each row and column has a specified sequence of shaded cells. For

Table 1: Car sequencing problem (timeout=5min). For each approach, we give the number of instances solved (#), the mean number of backtracks only if all the instances have been completely solved (bt.), and the mean CPU time over all the instances (in seconds).

n	flow&DAG-based			DAG-based			network-based				
	#	bt.	time	#	bt.	time	n'	n''	#	bt.	time
8	30	19.7	0.10	**30**	13.6	0.03	154	102	30	210.9	0.11
9	30	58.1	0.31	**30**	36.4	0.09	198	135	30	798.5	0.41
10	30	109.9	0.88	**30**	82.1	0.21	245	170	30	3,372	2.0
11	30	193.2	2.1	**30**	156.7	0.50	293	206	30	17,286	12.2
12	30	522.0	8.0	**30**	306.1	1.4	344	245	29	–	90.5
13	30	1,251	22.6	**30**	963.1	4.9	396	285	10	–	233.5
14	26	–	86.4	**30**	3,227	20.4	451	328	2	–	280.3
15	17	–	160.4	**29**	–	72.1	507	372	2	–	283.8
16	12	–	204.9	**23**	–	111.8	566	419	1	–	297.3

example, a row specified $(2, 3)$ contains two segments of shaded cells, one with length 2 and another with length 3. The problem is to find out which cells need to be shaded such that every row and every column contain the specific sequence of shaded cells. We model the problem by $n = p^2$ variables, in which x_{ij} denotes whether the cell at the i^{th} row and j^{th} column needs to be shaded. In the experiments, we generate random instances from perturbed white noise images. A random solution grid, with each cell colored with probability 0.5, is generated. A feasible nonogram problem instance is created from the lengths of the segments observed in this random grid. To make it infeasible, for each row and each column, the list of segment lengths is randomly permuted, *i.e.*, its elements are shuffled randomly. If a list is empty, then a segment of random length l is added ($0 < l < p$). We model and soften the restrictions on each row and column by W_REGULAR^{var}, resulting in three models: flow-based, DAG-based, and network-based. The flow-based model uses the W_REGULAR^{var} implementation based on minimum cost flows described in [48], the DAG-based version uses the filtering DAG (see [3] for implementation details), and the network-based version uses the decomposition presented in Example 3.

Table 2 shows the results of the experiments. For medium-size problems ($p \leq 9$, $n \leq 81$), the network-based approach develops the least number of backtracks on average compared to the two other approaches. Value and variable elimination at pre-processing reduces the number of variables by a factor greater than two. The flow-based and DAG-based approaches develop the same number of backtracks, producing the same EPTs, but the dynamic programming algorithm implemented in the DAG-based approach is about one order-of-magnitude faster than the minimum cost flow algorithm used in the flow-based approach. Moreover, the network-based approach is at least one order-of-magnitude faster than the DAG-based approach. On the largest instances, because of an exponential increase of the number of backtracks, the network-based approach becomes unable to solve all the instances in less than five minutes, but still outperforms the other two approaches.

9.3. The Well-formed Parentheses problem

In this experiment, we use a network-decomposition of the W_GRAMMAR constraint whose structure is depicted in Figure 2. It is obviously not Berge-acyclic. This experiment

Table 2: Nonogram (timeout=5min). For each approach, we give the number of instances solved (#), the mean number of backtracks only if all the instances have been completely solved (bt.), and the mean CPU time over all the instances (in seconds).

n	flow-based			DAG-based			network-based				
	#	bt.	time	#	bt.	time	n'	n''	#	bt.	time
36	30	11.4	0.09	30	11.4	0.01	96	18	**30**	4.4	0.00
49	30	41.8	0.29	30	41.8	0.05	133	42	**30**	22.5	0.01
64	30	186.4	2.3	30	186.4	0.26	176	64	**30**	90.3	0.01
81	30	254.4	4.5	30	254.4	0.50	225	97	**30**	248.9	0.04
100	25	–	86.0	30	3,581	10.8	280	131	**30**	3,861	0.47
121	19	–	166.8	26	–	72.8	341	171	**30**	12,919	1.6
144	3	–	279.6	9	–	233.2	408	224	**28**	–	44.1
169	0	–	300.0	5	–	266.5	481	267	**23**	–	116.4
196	0	–	300.0	1	–	297.1	560	330	**7**	–	257.1

will allow us to see the behavior of network-decompositions when they are not Berge-acyclic.

Given a set of $2p$ even length intervals within $[1, \ldots, 2p]$, the well-formed parentheses problem is to find a string of parentheses with length $2p$ such that substrings in each of the intervals are well-formed parentheses. We model this problem by a set of $n = 2p$ variables. Domains of size 6 are composed of three different parenthesis types: $()[]\{\}$. We post a W_GRAMMAR^{var} cost function on each interval to represent the requirement of well-formed parentheses. We generate $2p - 1$ even length intervals by randomly picking their end points in $[1, \ldots, 2p]$, and add an interval covering the whole range to ensure that all variables are constrained. We also randomly assign unary costs (between 0 and 10) to each variable.

We compare two models. The first model, the DAG-based approach, is obtained by modeling each W_GRAMMAR^{var} cost function using a filtering DAG approach.

In the second network-based model, we decompose each W_GRAMMAR^{var} cost function involving m variables using $m(m+1)/2$ extra variables $P_{i,j}$ ($1 \leq j \leq m, 1 \leq i \leq m - j + 1$) whose value corresponds to either a symbol value (for $j = 1$) or a pair of a symbol value S and a string length k ($1 \leq k < j$, for $j \geq 2$) associated to the substring $(i, i + j - 1)$, starting from i of length j. Ternary cost functions link every triplet $P_{i,j}$, $P_{i,k}$, $P_{i+k,j-k}$ so that there exists a compatible rule `S->AB` in order to get the substring $(i, i + j - 1)$ from the two substrings $(i, i + k - 1)$ and $(i + k, i + j - 1)$ when $P_{i,j} = (S, k)$, $P_{i,k} = (A, u)$, $P_{i+k,j-k} = (B, v)$ with $u < k$, $v < j - k$. Binary cost functions are used to encode the terminal rules between $P_{i,1}$ ($i \in [1, m]$) and the original variables.

Results are shown in Table 3. The network-based approach is clearly inefficient. It has $n' = 1,146$ variables on average for $p = 9$ ($n = 18$). The number of backtracks increases very rapidly due to the poor propagation on a non Berge-acyclic network. The DAG-based approach clearly dominates here. Notice that the DAG-based propagation of W_GRAMMAR^{var} can be very slow with around 1 backtrack per second for $p = 9$.

As a second experiment on well-formed parentheses, we generate new instances using only one hard global grammar constraint and a set of $p(2p - 1)$ binary cost functions corresponding to a complete graph. For each possible pair of positions, if a parentheses pair ((), [], or {}) is placed at these specific positions, then it incurs a randomly-generated

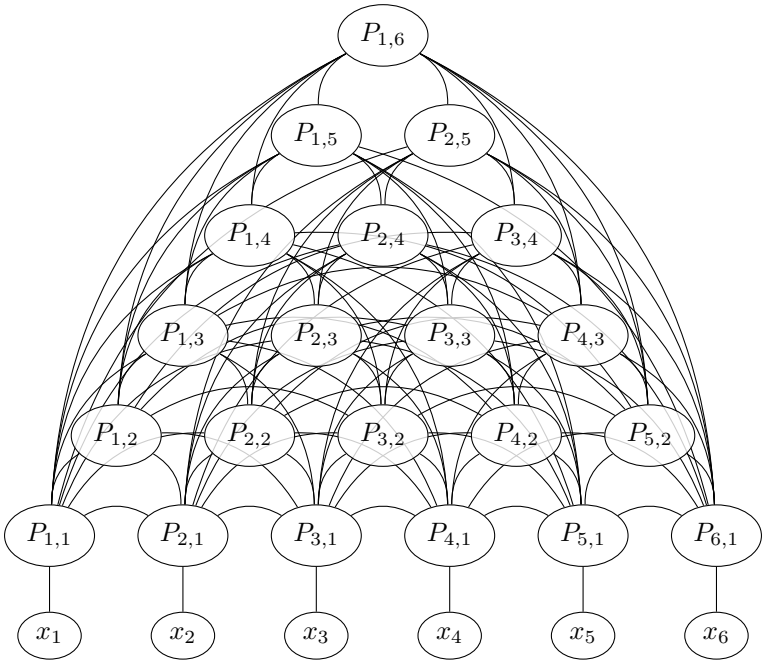

Figure 2: Network associated to the decomposition of W_Grammar$^{var}(x_1, x_2, x_3, x_4, x_5, x_6)$.

cost (between 0 to 10). A single W_Grammar^{var} cost function is placed on all the $n = 2p$ variables, which returns $\top$ on violation (a Grammar constraint), ensuring that the whole string has well-formed parentheses. As in the experiments of Table 3, the two models are characterized by how the consistency is enforced on the W_Grammar^{var} cost function: a filtering DAG for the DAG-based approach, a network-decomposition for the network-based approach.

Results are shown in Table 4. The network-based approach still develops more backtracks on average for $p \geq 6$ ($n \geq 12$) than the DAG-based approach but the difference is less important than in the previous experiment because there is a single grammar constraint. Surprisingly, for $p \leq 5$, the network-based approach develops less backtracks than the DAG-based approach. The network-based approach benefits from variable elimination that exploits bijective binary relations occurring in the decomposed hard grammar cost function. Moreover, having only one global constraint implies less extra variables for the network-based approach than in the previous experiment ($n' = 189$ for $p = 9$ instead of $n' = 1,146$). The propagation speed of the network-based approach is much better than the DAG-based approach, with $\sim$4,100 *bt./sec* instead of $\sim$23 *bt./sec* for $p = 9$, resulting in better overall time efficiency compared to the DAG-based approach, being up to 8 times faster for $p = 7$ to solve all the thirty instances.

9.4. The Market split problem

In some cases, problems may contain global cost functions which are not network-decomposable because the bounded arity cost function decomposition is not polynomial

Table 3: Soft well-formed parentheses (timeout=5min). For each approach, we give the number of instances solved (#), the mean number of backtracks only if all the instances have been completely solved (bt.), and the mean CPU time over all the instances (in seconds).

n	DAG-based			network-based				
	#	bt.	time	n'	n''	#	bt.	time
8	**30**	3.5	0.06	145	131	30	676.8	0.21
10	**30**	6.6	0.60	250	228	30	63,084	12.5
12	**30**	9.1	3.8	392	361	7	–	260.3
14	**30**	21.1	8.3	580	538	0	–	300
16	**29**	–	48.9	841	785	0	–	300
18	**23**	–	115.6	1,146	1,075	0	–	300

Table 4: Well-formed parentheses (single hard global constraint) with additional binary cost functions (timeout=5min). For each approach, we give the number of instances solved (#), the mean number of backtracks if available (bt.), and the mean CPU time (in seconds).

n	DAG-based			network-based				
	#	bt.	time	n'	n''	#	bt.	time
8	30	37.4	0.11	44	33	**30**	19.5	0.04
10	30	105.5	0.51	65	51	**30**	100.5	0.12
12	30	265.4	2.4	90	73	**30**	916.5	0.38
14	30	887.7	14.0	119	99	**30**	6,623	1.7
16	30	3,037	80.6	152	129	**30**	54,544	12.0
18	13	–	257.9	189	163	**30**	394,391	95.7

in size. However, if the network is Berge-acyclic, Theorem 11 still applies. With exponential size networks, filtering will take exponential time, but may yield strong lower bounds. The global constraint $\sum_{i=1}^{n} a_i x_i = b$ (a and b being integer coefficients) can be easily decomposed by introducing $n-3$ intermediate sum variables q_i and ternary sum constraints of the form $q_{i-1} + a_i x_i = q_i$ with $i \in [3, n-2]$ and $a_1 x_1 + a_2 x_2 = q_2$, $q_{n-2} + a_{n-1} x_{n-1} + a_n x_n = b$. More generally, ternary decompositions can be built for the more general case where the right hand side of the constraint uses any relational operator, including any Knapsack constraint. In this representation, the extra variables q_i have b values in their domain, which is exponential in the size of the representation of b (in $\log(b)$). As for the pseudo-polynomial Knapsack problem, if b is polynomially bounded by the size of the global constraint, propagation will be efficient. It may otherwise be exponential in it.

As an example, we consider a generalized version of the Knapsack problem, the Market Split problem defined in [23, 64]. The goal is to minimize $\sum_{i=1}^{n} o_i x_i$ such that $\sum_{i=1}^{n} a_{i,j} x_i = b_j$ for each $j \in [1, m]$ and x_i are Boolean variables in $\{0, 1\}$ (o, a and b being positive integer coefficients). We compared the Berge-acyclic decomposition in `toulbar2` (version 0.9.8) with a direct application of the Integer Linear Programming solver `cplex` (version 12.6.3.0). We used a depth-first search with a static variable ordering (in decreasing $\frac{o_i}{\sum_{j=1}^{m} a_{i,j}}$ order) and no pre-processing (options *-hbfs: -svo -o -nopre*) for `toulbar2`. We generated random instances with random integer coefficients in $[0, 99]$ for o and a, and $b_j = \lfloor \frac{1}{2} \sum_{i=1}^{n} a_{i,j} \rfloor$. We used a sample of 30 problems with $m = 4, n = 30$

leading to $\max b_j = 918$. The mean number of nodes developed in `toulbar2` was 29% higher than in `cplex`, which was on average 4.5 times faster than `toulbar2` on these problems. The 0/1 knapsack problem probably represents a worst case situation for `toulbar2`, given that `cplex` embeds much of what is known about 0/1 knapsacks (and only part of these extend to more complicated domains). Possible avenues to improve `toulbar2` results in this unfavorable situation would be to use a combination of the m knapsack constraints into one as suggested in [64].

10. Conclusion

Existing tools for solving optimization on graphical models are usually restricted to cost functions involving a reasonably small set of variables, often using an associated cost table. But problem modeling may require to express complex conditions on a non-bounded set of variables. This has been solved in Constraint Programming by using Global Constraints. Our results contribute to lift this approach to the more general framework of cost function networks, allowing to express and efficiently process both global constraints and global cost functions, using dedicated soft arc consistency filtering.

Our contributions are four-fold. First, we define the *tractability* of a global cost function, and study its behavior with respect to projections/extensions with different arities of cost functions. We show that tractable r-projection-safety is always possible for projections/extension to/from the nullary cost function, while it is always impossible for projections/extensions to/from r-ary cost functions for $r \geq 2$. When $r = 1$, we show that a tractable cost function may or may not be tractable 1-projection-safe. Second, we define *polynomially DAG-filterable cost functions* and show them to be tractable 1-projection-safe. We give also a polytime dynamic programming based algorithm to compute the minimum of this class of global cost functions. We also show that the cost function W_Grammar^{var}is polynomially DAG-filterable and tractable 1-projection-safe. The same results applies to W_Among^{var}, W_Regular^{var}, and W_Max/W_Min as shown in the associated technical report [3]. Third, we show that dynamic programming can be emulated by soft consistencies such as DAC and VAC if a suitable network decomposition of the global cost function into a Berge-acyclic network of bounded arity cost functions exists. In this case, local consistency on the decomposed network is essentially as strong as on the global cost function. This approach is shown to be a specific case of the previous approach in the sense that any Berge-acyclic network-decomposable cost function is also polynomially DAG-filterable. Finally, we perform experiments and compare the DAG-based and network-based approaches, in terms of run-time and search space. The DAG-based approach dominates when there are several overlapping global cost functions. On the contrary, the network-based approach performs better if there are few global cost functions resulting in a reasonable number of extra variables. This is complexified by additional techniques such as boosting search by variable elimination [40], Weighted Degree heuristics [17], and Dead-End Elimination [33] which work better with the low-arity cost functions of the network-based approach. We also compare against the flow-based approach [48] and show that our approaches are usually more competitive. On Berge acyclic network-decomposable cost function just as W_Regular^{var}, this is not unexpected as the dynamic programming based propagation or its emulation by T-DAC essentially solves a shortest path problem, which can easily be reduced to the more

general min-cost flow problem used in [48] which can itself be reduced to LP [1]. As problems become more specific, algorithmic efficiency can increase.

An immediate possible future work is to investigate other sufficient conditions for polynomially DAG-filterable and also tractable 1-projection-safety. Our results only provide a partial answer. Whether there exists necessary conditions for polynomially DAG-filterable is unknown. Besides polynomially DAG-filterable, we would like to investigate other form of tractable 1-projection-safety and techniques for enforcing typical consistency notions efficiently.

Acknowledgements

This work has been partly funded by the "Agence Nationale de la Recherche" (ANR-10-BLA-0214) and a PHC PROCORE project number 28680VH.

References

[1] Ahuja, R.K., Magnanti, T.L., Orlin, J.B., 2011. Network flows: Theory, Algorithms and Applications. Prentice Hall.

[2] Allouche, D., André, I., Barbe, S., Davies, J., de Givry, S., Katsirelos, G., O'Sullivan, B., Prestwich, S., Schiex, T., Traoré, S., 2014. Computational protein design as an optimization problem. Artificial Intelligence 212, 59–79.

[3] Allouche, D., Bessière, C., Boizumault, P., de Givry, S., Gutierrez, P., Lee, J.H.M., Leung, K., Loudni, S., Metivier, J., Schiex, T., Yi, W., 2015a. Tractability and Decompositions of Global Cost Functions. Technical Report Arxiv 1502.02414. The Chinese University of Hong Kong.

[4] Allouche, D., Bessière, C., Boizumault, P., de Givry, S., Gutierrez, P., Loudni, S., Metivier, J., Schiex, T., 2012. Decomposing Global Cost Functions, in: Proceedings of AAAI'12, pp. 407–413.

[5] Allouche, D., de Givry, S., Katsirelos, G., Schiex, T., Zytnicki, M., 2015b. Anytime Hybrid Best-First Search with Tree Decomposition for Weighted CSP, in: Proc. of CP-15, Cork, Ireland. pp. 12–28.

[6] Beeri, C., Fagin, R., Maier, D., M.Yannakakis, 1983. On the Desirability of Acyclic Database Schemes. Journal of the ACM 30, 479–513.

[7] Beldiceanu, N., 2001. Pruning for the Minimum Constraint Family and for the Number of Distinct Values Constraint Family, in: Proceedings of CP'01, pp. 211–224.

[8] Beldiceanu, N., Carlsson, M., Debruyne, R., Petit, T., 2005a. Reformulation of global constraints based on constraints checkers. Constraints 10, 339–362.

[9] Beldiceanu, N., Carlsson, M., Rampon, J., 2005b. Global Constraint Catalog. Technical Report T2005-08. Swedish Institute of Computer Science. Available at http://www.emn.fr/x-info/sdemasse/gccat/.

[10] Beldiceanu, N., Contejean, E., 1994. Introducing Global Constraints in CHIP. Mathematical and Computer Modelling 20, 97–123.

[11] Beldiceanu, N., Katriel, I., Thiel, S., 2004. Filtering Algorithms for the Same Constraints, in: Proceedings of CPAIOR'04, pp. 65–79.

[12] Bertelé, U., Brioshi, F., 1972. Nonserial Dynamic Programming. Academic Press.

[13] Bessière, C., 2006. Constraint propagation, in: Rossi, F., van Beek, P., Walsh, T. (Eds.), Handbook of Constraint Programming. Elsevier. chapter 3, pp. 29–84.

[14] Bessière, C., Hebrard, E., Hnich, B., Kiziltan, Z., Quimper, C., Walsh, T., 2007. Reformulating global constraints: the SLIDE and REGULAR constraints. Abstraction, Reformulation, and Approximation , 80–92.

[15] Bessière, C., Hebrard, E., Hnich, B., Kiziltan, Z., Walsh, T., 2008. SLIDE: A Useful Special Case of the CARDPATH Constraint, in: Proceedings of ECAI'08, pp. 475–479.

[16] Bessière, C., Van Hentenryck, P., 2003. To be or not to be ... a global constraint, in: Proceedings of CP'03, pp. 789–794.

[17] Boussemart, F., Hemery, F., Lecoutre, C., Sais, L., 2004. Boosting Systematic Search by Weighting Constraints, in: ECAI, pp. 146–150.

[18] Cabon, B., de Givry, S., Lobjois, L., Schiex, T., Warners, J., 1999. Radio link frequency assignment. Constraints Journal 4, 79–89.

[19] Cooper, M., de Givry, S., Sánchez, M., Schiex, T., Zytnicki, M., Werner, T., 2010. Soft Arc Consistency Revisited. Artificial Intelligence 174, 449–478.
[20] Cooper, M., Schiex, T., 2004. Arc Consistency for Soft Constraints. Artificial Intelligence 154, 199–227.
[21] Cooper, M.C., 2003. Reduction operations in fuzzy or valued constraint satisfaction. Fuzzy Sets and Systems 134, 311–342.
[22] Cooper, M.C., 2005. High-Order Consistency in Valued Constraint Satisfaction. Constraints 10, 283–305.
[23] Cornuéjols, G., Dawande, M., 1998. A Class of Hard Small 0-1 Programs, in: Proceedings of Integer Programming and Combinatorial Optimization 1998, pp. 284–293.
[24] Culik II, K., Kari, J., 1993. Image Compression Using Weighted Finite Automata, in: Borzyszkowski, A.M., Sokolowski, S. (Eds.), MFCS, Springer. pp. 392–402.
[25] Dasgupta, S., Papadimitriou, C.H., Vazirani, U.V., 2007. Algorithms. McGraw-Hill.
[26] Dechter, R., Rish, I., 2003. Mini-buckets: A general scheme for bounded inference. J. ACM 50, 107–153.
[27] Desmet, J., De Maeyer, M., Hazes, B., Lasters, I., 1992. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356, 539–42.
[28] Dibangoye, J.S., Amato, C., Buffet, O., Charpillet, F., 2013. Optimally solving dec-pomdps as continuous-state mdps, in: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, AAAI Press. pp. 90–96.
[29] Ermon, S., Gomes, C.P., Sabharwal, A., Selman, B., 2013. Embed and project: Discrete sampling with universal hashing, in: Advances in Neural Information Processing Systems, pp. 2085–2093.
[30] Favier, A., de Givry, S., Legarra, A., Schiex, T., 2011. Pairwise decomposition for combinatorial optimization in graphical models, in: Proc. of IJCAI-11, pp. 2126–2132.
[31] Flum, J., Grohe, M., 2006. Parameterized Complexity Theory. Springer-Verlag New York Inc.
[32] de Givry, S., Heras, F., Zytnicki, M., Larrosa, J., 2005. Existential Arc Consistency: Getting Closer to Full Arc Consistency in Weighted CSPs, in: Proceedings of IJCAI'05, pp. 84–89.
[33] de Givry, S., Prestwich, S., O'Sullivan, B., 2013. Dead-End Elimination for Weighted CSP, in: Proceedings of CP'13, pp. 263–272.
[34] van Hoeve, W.J., Pesant, G., Rousseau, L.M., 2006. On Global Warming: Flow-based Soft Global Constraints. J. Heuristics 12, 347–373.
[35] Hurley, B., O'Sullivan, B., Allouche, D., Katsirelos, G., Schiex, T., Zytnicki, M., de Givry, S., 2016. Multi-language evaluation in graphical model optimization. Constraints Initially submitted to CP-AI-OR'2016.
[36] Ishida, N., 1994. Game "NONOGRAM"(in Japanese). Mathematical Seminar 10, 21–22.
[37] Kadioglu, S., Sellmann, M., 2010. Grammar Constraints. Constraints 15, 117–144.
[38] Katsirelos, G., Narodytska, N., Walsh, T., 2011. The Weighted GRAMMAR Constraints. Annals of Operations Research 184, 179–207.
[39] Krom, M., 1967. The Decision Problem for a Class of First-Order Formulas in Which all Disjunctions are Binary. Mathematical Logic Quarterly 13, 15–20.
[40] Larrosa, J., 2000. Boosting Search with Variable Elimination, in: Proceedings of CP'00, pp. 291–305.
[41] Larrosa, J., Schiex, T., 2003. In the Quest of the Best Form of Local Consistency for Weighted CSP, in: Proceedings of IJCAI'03, pp. 239–244.
[42] Larrosa, J., Schiex, T., 2004. Solving Weighted CSP by Maintaining Arc Consistency. Artificial Intelligence 159, 1–26.
[43] Lauriere, J.L., 1978. A Language and a Program for Stating and Solving Combinatorial Problems. Artificial Intelligence 10, 29–127.
[44] Lecoutre, C., Roussel, O., Dehani, D.E., 2012. Wcsp integration of soft neighborhood substitutability, in: Proceedings of CP'12, pp. 406–421.
[45] Lecoutre, C., Saïs, L., Tabary, S., Vidal, V., 2009. Reasoning from Last Conflict(s) in Constraint Programming. Artificial Intelligence 173, 1592,1614.
[46] Lee, J.H.M., Leung, K.L., 2009. Towards Efficient Consistency Enforcement for Global Constraints in Weighted Constraint Satisfaction, in: Proceedings of IJCAI'09, pp. 559–565.
[47] Lee, J.H.M., Leung, K.L., 2010. A Stronger Consistency for Soft Global Constraints in Weighted Constraint Satisfaction, in: Proceedings of AAAI'10, pp. 121–127.
[48] Lee, J.H.M., Leung, K.L., 2012. Consistency Techniques for Global Cost Functions in Weighted Constraint Satisfaction. Journal of Artificial Intelligence Research 43, 257–292.
[49] Lee, J.H.M., Leung, K.L., Shum, Y.W., 2014. Consistency Techniques for Polytime Linear Global

Cost Functions in Weighted Constraint Satisfaction. Constraints 19, 270,308.
[50] Lee, J.H.M., Leung, K.L., Wu, Y., 2012. Polynomially Decomposable Global Cost Functions in Weighted Constraint Satisfaction, in: Proceedings of AAAI'12, pp. 507–513.
[51] Lee, J.H.M., Shum, Y.W., 2011. Modeling Soft Global Constraints as Linear Programs in Weighted Constraint Satisfaction, in: Proceedings of ICTAI'11, pp. 305–312.
[52] Oplobedu, A., Marcovitch, J., Tourbier, Y., 1989. CHARME: Un langage industriel de programmation par contraintes, illustré par une application chez Renault, in: Proceedings of the Ninth International Workshop on Expert Systems and their Applications: General Conference, p. 55–70.
[53] Papadimitriou, C., Yannakakis, M., 1991. Optimization, approximation, and complexity classes. Journal of Computer and System Sciences 43, 425–440.
[54] Parrello, B., Kabat, W., Wos, L., 1986. Job-shop Scheduling Using Automated Reasoning: A Case Study of the Car-Sequence Problem. Journal of Automated Reasoning 2, 1–42.
[55] Pesant, G., 2004. A Regular Language Membership Constraint for Finite Sequences of Variables, in: Proceedings of CP'04, pp. 482–495.
[56] Petit, T., Régin, J.C., Bessière, C., 2001a. Specific Filtering Algorithm for Over-Constrained Problems, in: Proceedings of CP'01, pp. 451–463.
[57] Petit, T., Régin, J.C., Bessiere, C., 2001b. Specific filtering algorithms for over-constrained problems, in: CP, pp. 451–463.
[58] Régin, J.C., 1996. Generalized Arc Consistency for Global Cardinality Constraints, in: Proceedings of AAAI'96, pp. 209–215.
[59] Rossi, F., van Beek, P., Walsh, T., 2006. Handbook of Constraint Programming. Elsevier.
[60] Sánchez, M., de Givry, S., Schiex, T., 2008. Mendelian Error Detection in Complex Pedigrees using Weighted Constraint Satisfaction Techniques. Constraints 13, 130–154.
[61] Schiex, T., 2000. Arc Consistency for Soft Constraints, in: Principles and Practice of Constraint Programming - CP 2000, pp. 411–424.
[62] Schiex, T., Fargier, H., Verfaillie, G., 1995. Valued Constraint Satisfaction Problems: Hard and Easy Problems, in: Proceedings of IJCAI'95, pp. 631–637.
[63] Simoncini, D., Allouche, D., de Givry, S., Delmas, C., Barbe, S., Schiex, T., 2015. Guaranteed discrete energy optimization on large protein design problems. Journal of Chemical Theory and Computation .
[64] Trick, M.A., 2003. A Dynamic Programming Approach for Consistency and Propagation for Knapsack Constraints. Annals of Operations Research 118, 73–84.
[65] Zytnicki, M., Gaspin, C., Schiex, T., 2009. Bounds Arc Consistency for Weighted CSPs. Journal of Artificial Intelligence Research 35, 593–621.