

# Space-Time Tradeoffs for Conjunctive Queries with Access Patterns

Hangdong Zhao

University of Wisconsin-Madison  
Madison, WI, USA  
hangdong@cs.wisc.edu

Shaleen Deep

Microsoft Gray Systems Lab  
Madison, WI, USA  
shaleen.deep@microsoft.com

Paraschos Koutris

University of Wisconsin-Madison  
Madison, WI, USA  
paris@cs.wisc.edu

## ABSTRACT

In this paper, we investigate space-time tradeoffs for answering conjunctive queries with access patterns (CQAPs). The goal is to create a space-efficient data structure in an initial preprocessing phase and use it for answering (multiple) queries in an online phase. Previous work has developed data structures that trades off space usage for answering time for queries of practical interest, such as the path and triangle query. However, these approaches lack a comprehensive framework and are not generalizable. Our main contribution is a general algorithmic framework for obtaining space-time tradeoffs for any CQAP. Our framework builds upon the PANDA algorithm and tree decomposition techniques. We demonstrate that our framework captures all state-of-the-art tradeoffs that were independently produced for various queries. Further, we show surprising improvements over the state-of-the-art tradeoffs known in the existing literature for reachability queries.

## 1 INTRODUCTION

We study a class of problems that splits an algorithmic task into two phases: the *preprocessing phase*, which computes a space-efficient data structure from the input, and the *online phase*, which uses the data structure to answer requests of a specific form over the input as fast as possible. Many important algorithmic tasks such as set intersection problems [8, 15], reachability in directed graphs [2, 3, 9], histogram indexing [7, 25], and problems related to document retrieval [1, 26] can be expressed in this way. The fundamental algorithmic question related to these problems is to find *the tradeoff between the space  $S$  necessary for storing the data structures and the time  $T$  for answering a request*.

Let us look at one of the simplest tasks in this setup. Consider the 2-Set Disjointness problem: given a universe of elements  $U$  and a collection of  $m$  sets  $S_1, \dots, S_m \subseteq U$ , we want to create a data structure such that for any pair of integers  $1 \leq i, j \leq m$ , we can efficiently decide whether  $S_i \cap S_j$  is empty or not. It is well-known that the space-time tradeoff for 2-Set Disjointness is captured by the equation  $S \cdot T^2 = O(N^2)$ , where  $N$  is the total size of all sets [8, 15]. Similar tradeoffs have also been established for other

data structure problems. In the  $k$ -Reachability problem [8, 15] we are given as input a directed graph  $G = (V, E)$ , an arbitrary pair of vertices  $u, v$ , and the goal is to decide whether there exists a path of length  $k$  between  $u$  and  $v$ . The data structure obtained was conjectured to be optimal by [15], and the conjectured optimality was used to develop conditional lower bounds for other problems, such as approximate distance oracles [2, 3] where no progress has been made in improving the upper bounds in the last decade. In the edge triangle detection problem [15], we are given as input a graph  $G = (V, E)$ , and the goal is to develop a data structure that can answer whether a given edge  $e \in E$  participates in a triangle or not. Each of these problems has been studied in isolation and therefore, the algorithmic insights are not readily generalizable into a comprehensive framework.

In this paper, we cast many of the above problems into answering *Conjunctive Queries with Access Patterns* (CQAPs) over a relational database. For example, by using the relation  $R(x, y)$  to encode that element  $x$  belongs to set  $y$ , 2-Set Disjointness can be captured by the following CQAP:  $\varphi(|y_1, y_2| \leftarrow R(x, y_1) \wedge R(x, y_2))$ . The expression  $\varphi(|y_1, y_2|)$  can be interpreted as follows: given values for  $y_1, y_2$ , compute whether the query returns true or not. Different access patterns capture different ways of accessing the result of the CQ and result in different tradeoffs.

Tradeoffs for enumerating Conjunctive Query results under static and dynamic settings have been a subject of previous research [13, 17, 19–21, 30]. However, previous work either focuses on the tradeoff between preprocessing time and answering time [19–21], or the tradeoff between space and delay in enumeration [13, 30]. In this paper, we focus explicitly on the tradeoff between space and answering time, without optimizing for preprocessing time. Most closely related to our setting is the problem of answering Boolean CQs [12]. In that work, the authors slightly improve upon the data structure proposed in [13] and adapt it for Boolean CQ answering. Further, [12] identified that the conjectured tradeoff for the  $k$ -reachability problem is suboptimal by showing slightly improved tradeoffs for all  $k \geq 3$ . The techniques used in this paper are quite different and a vast generalization of the techniques used in [12]. The proposed improvements in [12] for  $k$ -reachability are already captured in this work and in many cases, surpass the ones from [12].

**Our Contribution.** Our key contribution is a general algorithmic framework for obtaining space-time tradeoffs for any CQAP. Our framework builds upon the PANDA algorithm [24] and tree decomposition techniques [16, 28]. Given any CQAP, it calculates a tradeoff that can find the best possible time for a given space budget. To achieve this goal, we need two key technical contributions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PODS '23, June 18–23, 2023, Seattle, WA, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0127-6/23/06...\$15.00  
https://doi.org/10.1145/3584372.3588675

First, we introduce the novel notion of *partially-materialized tree decompositions* (PMTDs) that allow us to capture different possible materialization strategies on a given tree decomposition (Section 3). At a high level, a PMTD augments a tree decomposition with information on which bags should be materialized and which should be computed online. To use a PMTD, we propose a variant of the Yannakakis algorithm (Subsection 3.1) such that during the online phase we incur only the cost of visiting the non-materialized bags.

The second key ingredient is an extension of the PANDA algorithm [24] that computes a disjunctive rule in two phases. The computation of a disjunctive rule allows placing an answer to any of the targets in the head of the rule. A key technical component in the PANDA algorithm is the notion of a *Shannon-flow inequality*. For any Shannon-flow inequality, one can construct a proof sequence that has a direct correspondence with relational operators. Consequently, a proof sequence can be transformed into a join algorithm. The disjunctive rules we consider are computed in two phases: in the first phase (preprocessing), we can place an answer only to targets that will be materialized during the preprocessing phase. In the second phase (online), we place an answer to the remaining targets. We call these rules *2-phase disjunctive rules* (Subsection 4.1). To achieve this 2-phase computation, we introduce a type of Shannon-flow inequalities, called *joint Shannon-flow inequalities* (Section 5), such that each inequality gives rise to a space-time tradeoff. The joint Shannon-flow inequality generates two parallel proof sequences, one proof sequence for the preprocessing phase and another proof sequence for the answering phase. This transformation allows us to use the PANDA algorithm as a blackbox on each of the proof sequences independently and is instrumental in achieving space-time tradeoffs.

We demonstrate the versatility of our framework by recovering state-of-the-art space-time tradeoffs for Boolean CQAPs, 2-Set Disjointness as well as its generalization  $k$ -Set Disjointness, and  $k$ -Reachability (Section 6). We also apply our framework to the previously unstudied setting of space-time tradeoffs (in the static setting) for access patterns over a subset of *hierarchical queries*, a fragment of acyclic CQs that is of great interest [5, 6, 10, 11, 18, 20]. Interestingly, we can recover strategies that are very similar to how specialized enumeration algorithms with provable guarantees work for this class of CQs [11, 20]. More importantly, we improve state-of-the-art tradeoffs. Our most interesting finding is that we can obtain complex tradeoffs for  $k$ -Reachability that exhibit different behavior for different regimes of  $S$ . For the 3-Reachability problem, we show how to improve the tradeoff for a significant part of the spectrum. For the 4-Reachability problem, we are able to show (via a rather involved analysis) that the space-time tradeoff can be improved *everywhere* when compared to the conjectured optimal! These results falsify the proposed optimal tradeoff of  $S \cdot T^{2/(k-1)} = \tilde{O}(|E|^2)$  for  $k$ -Reachability for regimes that are even larger than what was shown in [12].

**Organization.** We introduce the basic terminology and problem definition in Section 2. In Section 3, we describe the augmented tree decompositions that are necessary for our framework. Section 4 introduces the general framework while Section 5 presents the algorithms used in our framework. We present the

applications of the framework in Section 6. The related work is described in Section 7 and we conclude with a list of open problems in Section 8.

## 2 BACKGROUND

**Conjunctive Query.** We associate a Conjunctive Query (CQ)  $\varphi$  with a hypergraph  $\mathcal{H} = ([n], \mathcal{E})$ , where  $[n] = \{1, \dots, n\}$  and  $\mathcal{E} \subseteq 2^{[n]}$ . The body of the query has atoms  $R_F$ , where  $F \in \mathcal{E}$ . To each node  $i \in [n]$ , we associate a variable  $x_i$ . The CQ is then

$$\varphi(\mathbf{x}_H) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F),$$

where  $\mathbf{x}_I$  denotes the tuple  $(x_i)_{i \in I}$  for any  $I \subseteq [n]$ . The variables in  $\mathbf{x}_H$  are called the *head variables* of the CQ. The CQ is *full* if  $H = [n]$  and *Boolean* if  $H = \emptyset$ . We use  $\varphi$  to denote the output of the CQ  $\varphi$ .

**Degree Constraints.** A *degree constraint* is a triple  $(X, Y, N_{Y|X})$  where  $X \subset Y \subseteq [n]$  and  $N_{Y|X}$  is a natural number. A relation  $R_F$  is said to *guard* the degree constraint  $(X, Y, N_{Y|X})$  if  $X \subset Y \subseteq F$  and for every tuple  $\mathbf{t}_X$  (over  $X$ ),  $\max_{\mathbf{t}_X} \deg_F(Y|\mathbf{t}_X) \leq N_{Y|X}$ , where  $\deg_F(Y|\mathbf{t}_X) = |\Pi_{Y \setminus X} R_F|$ . We use DC to denote a set of degree constraints and say that DC is guarded by a database instance  $\mathcal{D}$  if every  $(X, Y, N_{Y|X}) \in \text{DC}$  is guarded by some relation in  $\mathcal{D}$ . A degree constraint  $(X, Y, N_{Y|X})$  is a *cardinality constraint* if  $X = \emptyset$ . Throughout this work, we make the following assumptions on DC guarded by a database instance  $\mathcal{D}$ :

- (*best constraints assumption*) w.l.o.g, for any  $X \subset Y \subseteq [n]$ , there is at most one  $(X, Y, N_{Y|X}) \in \text{DC}$ . This assumption can be maintained by only keeping the minimum  $N_{Y|X}$  if there is more than one.
- for every relation  $R_F \in \mathcal{D}$ , there is a *cardinality constraint*  $(\emptyset, F, |R_F| \stackrel{\text{def}}{=} N_{F|\emptyset}) \in \text{DC}$ . The *size* of the database  $\mathcal{D}$  is denoted as  $|\mathcal{D}| \stackrel{\text{def}}{=} \max_{R_F \in \mathcal{D}} |R_F|$ .

In this work, we use degree constraints to measure data complexity. All logs are in base 2, unless otherwise stated.

### 2.1 CQs with Access Patterns

We define CQs with access patterns following the definition from [21]:

**Definition 2.1 (CQ with access patterns).** A Conjunctive Query with *Access Patterns* (CQAP) is an expression of the form

$$\varphi(\mathbf{x}_H \mid \mathbf{x}_A) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F),$$

where  $A \subseteq [n]$  is called the *access pattern* of the query.

The access pattern tells us how a user accesses the result of the CQ. In particular, the user will provide an instance of a relation  $Q_A(\mathbf{x}_A)$ , which we call an *access request*. The task is then to return the result of the following CQ, denoted as  $\varphi$ , where

$$\varphi(\mathbf{x}_H) \leftarrow Q_A(\mathbf{x}_A) \wedge \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F).$$

We call  $\varphi$  the *access CQ*. The most natural access request is one where  $|Q_A| = 1$ ; in other words, the user provides only one fixed value for every variable  $x_i, i \in A$ . This can be thought of as using the CQ result as an index with search key  $\mathbf{x}_A$ . By allowing the

access request  $Q_A$  to consist of more tuples, we can capture other scenarios. For example, one can take a stream of access requests of size 1 and batch them together to obtain a (possibly faster) answer for all of them at once. Prior work [13, 21] has only considered the case where  $|Q_A| = 1$ .

## 2.2 Problem Statement

Let  $\varphi(\mathbf{x}_H \mid \mathbf{x}_A)$  be a CQAP under degree constraints DC guarded by the input relations. In addition, we denote by AC another set of degree constraints known in prior, guarded by any access request  $Q_A$ . Similar to DC, we assume there is a cardinality constraint  $(\emptyset, A, |Q_A| = N_A|_\emptyset) \in AC$  guarded by  $Q_A$ . For example, the case where  $|Q_A| = 1$  can be interpreted as a cardinality constraint  $(\emptyset, A, 1) \in AC$ . Given a database instance  $\mathcal{D}$  guarding DC, our goal is to construct a data structure, such that we can answer any access request as fast as possible. More formally, we split query processing into two phases:

**Preprocessing phase:** it constructs a data structure in space  $\tilde{O}(S)^1$ . The overall space cost takes the form  $\tilde{O}(S + |\mathcal{D}|)$ , where  $S$  is called the *intrinsic space cost* of the data structure and  $|\mathcal{D}|$  is the (unavoidable) space cost for storing the database.

**Online phase:** given an access request  $Q_A$  (guarding AC), it returns the results of the access CQ  $\varphi$  using the data structure built in the preprocessing phase. The (worst-case) answering time is then  $\tilde{O}(T + |Q_A|) + O(|\varphi|)$ , where  $T$  is called the *intrinsic time cost* and  $|Q_A| + |\varphi|$  is the (unavoidable) time cost of reading the access request  $Q_A$  and enumerating the output. For the Boolean case and when  $|Q_A| = 1$ , the answering time simply becomes  $\tilde{O}(T)$ .

In this work, we study the tradeoffs between the two intrinsic quantities,  $S$  and  $T$ , which we will call an *intrinsic tradeoff*. At one extreme, the algorithm stores nothing, thus  $S = O(1)$ , and we answer each access request from scratch. At the other extreme, the algorithm stores the results of the CQ  $\varphi_M(\mathbf{x}_{H \cup A}) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F)$  as a hash table with index key  $\mathbf{x}_A$ . For any access request  $Q_A$ , we simply evaluate the query  $\varphi(\mathbf{x}_H) \leftarrow Q_A \wedge \varphi_M$  in the online phase by probing each tuple of  $Q_A$  in the hash table. If  $H \supseteq A$ , then any access request can be answered in (instance-optimal) time  $O(|Q_A| + |\varphi|)$ , in which case  $T = O(1)$ .

**Example 2.2 (k-Set Disjointness).** In this problem, we are given sets  $S_1, \dots, S_m$  with elements drawn from the same universe  $U$ . Each access request asks whether the intersection between  $k$  sets is empty or not. By encoding the family of sets as a binary relation  $R(y, x)$  such that element  $y$  belongs to set  $x$ , we can express the problem as the following CQAP:

$$\varphi(\mid \mathbf{x}_{[k]} \mid) \leftarrow \bigwedge_{i \in [k]} R(y, x_i). \quad (1)$$

If we also want to enumerate the elements in their intersection, we would instead use the non-Boolean version:

$$\varphi(y \mid \mathbf{x}_{[k]}) \leftarrow \bigwedge_{i \in [k]} R(y, x_i). \quad (2)$$

<sup>1</sup>The notation  $\tilde{O}$  hides a polylogarithmic factor in  $|\mathcal{D}|$ .

**Example 2.3 (k-Reachability).** Given a direct graph  $G$ , the  $k$ -reachability problem asks, given a pair vertices  $(u, v)$ , to check whether they are connected by a path of length  $k$ . Representing the graph as a binary relation  $R(x, y)$ , we can model this problem through the following CQAP (the  $k$ -path query):

$$\phi_k(\mid \mathbf{x}_1, \mathbf{x}_{k+1} \mid) \leftarrow \bigwedge_{i \in [k]} R(x_i, x_{i+1}).$$

We can also check whether there is a path of length at most  $k$  by combining the results of  $k$  such queries (one for each  $1, \dots, k$ ).

In this work, we focus on the CQAP such that  $H \supseteq A$ . If we are given a CQAP where  $H \not\supseteq A$ , we replace the head of the CQAP with  $\varphi(\mathbf{x}_{H \cup A} \mid \mathbf{x}_A)$ , and simply project on the desired results in the end.

## 3 PARTIALLY MATERIALIZED TREE DECOMPOSITIONS

In this section, we introduce a type of tree decomposition that augments a decomposition with information about what bags we want to materialize.

**Definition 3.1 (Tree Decomposition).** A *tree decomposition* of a hypergraph  $\mathcal{H} = ([n], \mathcal{E})$  is a pair  $(\mathcal{T}, \chi)$  where (i)  $\mathcal{T}$  is an undirected tree, and (ii)  $\chi : V(\mathcal{T}) \rightarrow 2^{[n]}$  is a mapping that assigns to every node  $t \in V(\mathcal{T})$  a subset of  $[n]$ , called the *bag* of  $t$ , such that

- (1) For every hyperedge  $F \in \mathcal{E}$ , the set  $F$  is contained in some bag; and
- (2) For each vertex  $x \in [n]$ , the set of nodes  $\{t \mid x \in \chi(t)\}$  forms a (non-empty) connected subtree of  $\mathcal{T}$ .

Take a tree decomposition  $(\mathcal{T}, \chi)$  and a node  $r \in V(\mathcal{T})$ . We define  $\text{TOP}_r(x)$  as the highest node in  $\mathcal{T}$  containing  $x$  in its bag if we root the tree at  $r$ . We now say that  $(\mathcal{T}, \chi)$  is *free-connex w.r.t. r* if for any  $x \in H$  and  $y \in [n] \setminus H$ ,  $\text{TOP}_r(y)$  is not an ancestor of  $\text{TOP}_r(x)$  [34]. We say that  $(\mathcal{T}, \chi)$  is free-connex if it is free-connex w.r.t. some  $r \in V(\mathcal{T})$ .

We can now introduce our key concept of a partially materialized tree decomposition, tailored for CQAPs. Let  $\varphi(\mathbf{x}_H \mid \mathbf{x}_A)$  be a CQAP such that  $A \subseteq H$ . Let  $\mathcal{H}$  be the hypergraph associated with  $\varphi(\mathbf{x}_H)$ , the access CQ.

**Definition 3.2 (PMTD).** A *Partially Materialized Tree Decomposition (PMTD)* of the CQAP  $\varphi(\mathbf{x}_H \mid \mathbf{x}_A)$  with  $H \supseteq A$  is a tuple  $(\mathcal{T}, \chi, M, r)$  such that the following properties hold:

- (1)  $(\mathcal{T}, \chi)$  is a free-connex tree decomposition of  $\mathcal{H}$  w.r.t. node  $r$ , called the *root* ; and
- (2)  $A \subseteq \chi(r)$  ; and
- (3)  $M \subseteq V(\mathcal{T})$  such that whenever  $t \in M$  then all the nodes of its subtree (w.r.t. orienting the tree away from  $r$ ) are in  $M$ .

Given a PMTD  $(\mathcal{T}, \chi, M, r)$ , we call  $M$  the *materialization set*. We also associate with each node  $t \in V(\mathcal{T})$  a *view* with variables  $\mathbf{x}_{v(t)}$ , where the mapping  $v : V(\mathcal{T}) \rightarrow 2^{[n]}$  is defined as follows.

If the node  $t \notin M$ , then  $v(t) \stackrel{\text{def}}{=} \chi(t)$  and the view is of the form  $T_{v(t)}(\mathbf{x}_{v(t)})$ , called a *T-view*. Otherwise,  $t \in M$ . If  $t = r \in M$ , define  $v(r) \stackrel{\text{def}}{=} \chi(t) \cap H$ . Let  $p$  be the parent node of a non-root node  $t \in M$

and define

$$v(t) \stackrel{\text{def}}{=} \begin{cases} \chi(t) \cap (H \cup \chi(p)) & \text{if } p \notin M \\ \chi(t) \cap H & \text{if } p \in M \text{ and } \chi(t) \cap H \not\subseteq \chi(p) \cap H \\ \emptyset & \text{if } p \in M \text{ and } \chi(t) \cap H \subseteq \chi(p) \cap H. \end{cases}$$

The view (for each  $t \in M$ ) then is of the form  $S_{v(t)}(\mathbf{x}_{v(t)})$ , called the *S-view*. This definition of *S*-views corresponds to running a bottom-up semijoin-reduce pass of the Yannakakis algorithm in the materialization set  $M$  of the free-connex tree decomposition  $(\mathcal{T}, \chi)$ . Indeed, any variables in  $\chi(t) \setminus v(t)$  are safely projected out after the semijoin-reduce.

On a high level,  $M$  specifies the type of views associated with each bag (*S*-view or *T*-view), and  $v(\cdot)$  pinpoints the schema of that view (possibly empty). A PMTD appoints its *S*-views to be materialized in the preprocessing phase and its *T*-views to be computed in the online phase. In the case where  $M = \emptyset$ , every view in the decomposition is obtained in the online phase. When  $H = A$  or  $H = [n]$ , the free-connex property does not put any additional restrictions on the tree decompositions for a PMTD.

*Example 3.3.* We use the CQAP for 3-reachability as an example:

$$\phi_3(x_1, x_4 \mid x_1, x_4) \leftarrow R_1(x_1, x_2) \wedge R_2(x_2, x_3) \wedge R_3(x_3, x_4).$$

Here,  $(x_1, x_4)$  is the access pattern. Figure 1 shows three PMTDs for the above query, along with the associated views of each bag in each PMTD. The leftmost PMTD has an empty materialization set. The middle PMTD materializes the bag  $\{x_1, x_2, x_3\}$  but the associated view  $S_{13}$  projects out  $x_2$ . The rightmost PMTD materializes the only bag  $\{x_1, x_2, x_3, x_4\}$  but the view  $S_{14}$  keeps only the variables  $x_1, x_4$ .

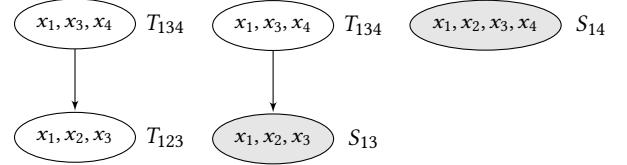
**Redundancy & Domination.** We say that a tree decomposition is *non-redundant* if no bag is a subset of another bag. We say that a tree decomposition  $(\mathcal{T}_1, \chi_1)$  is *dominated* by another tree decomposition  $(\mathcal{T}_2, \chi_2)$  if every bag of  $(\mathcal{T}_1, \chi_1)$  is a subset of some bag of  $(\mathcal{T}_2, \chi_2)$ . Here, we will generalize both notions to PMTDs.

*Definition 3.4 (PMTD Redundancy).* A PMTD  $(\mathcal{T}, \chi, M, r)$  is *non-redundant* if (1) for  $t \in M$ ,  $v(t) \neq \emptyset$  and no  $v(t)$  is a subset of another; and (2) for  $t \notin M$ , no  $v(t)$  is a subset of another.

*Definition 3.5 (PMTD Domination).* A PMTD  $(\mathcal{T}_1, \chi_1, M_1, r_1)$  is dominated by another PMTD  $(\mathcal{T}_2, \chi_2, M_2, r_2)$  if (1) for every node  $t_1 \in M_1$ , there is some node  $t_2 \in M_2$  such that  $v(t_1) \subseteq v(t_2)$ , and (2) for every node  $t_1 \in V(\mathcal{T}_1) \setminus M_1$ , there is some node  $t_2 \in V(\mathcal{T}_2) \setminus M_2$  such that  $v(t_1) \subseteq v(t_2)$ .

For PMTDs, both redundancy and domination are defined using the materialization set and views instead of the bags. For PMTDs with  $M = \emptyset$ , both PMTD redundancy and domination become equivalent to the standard definition.

*Example 3.6.* Continuing Example 3.3, suppose we consider a PMTD with that takes the same tree decomposition as the left PMTD, but with both bags in the materialization set. The *S*-view associated with the root bag is  $S_{14}$ , and  $S_\emptyset$  for the child bag; thus, this PMTD is redundant. Moreover, suppose we consider a PMTD with one bag  $\{x_1, x_2, x_3, x_4\}$  which is the root, but is not in  $M$ . The *T*-view associated with this bag is  $T_{1234}$ ; thus, this PMTD dominates the left PMTD in Figure 1. On the other hand, all PMTDs in Figure 1 are non-redundant and non-dominant to each other.



**Figure 1: Three PMTDs for the 3-reachability CQAP. The materialized nodes are shaded and labeled as *S*-views.**

As we later suggest in our general framework, we mostly focus on sets of non-redundant and non-dominant PMTDs. Note that a non-redundant PMTD  $(\mathcal{T}, \chi, M, r)$  satisfies  $v(t) \neq \emptyset$ , for any  $t \in V(\mathcal{T})$ , thus we can safely assume that all views are non-empty.

### 3.1 Online Yannakakis for PMTDs

We introduce an adaptation of the Yannakakis algorithm [37] for a non-redundant PMTD (so no empty views), called *Online Yannakakis*. Recall that for a non-redundant PMTD, the *S*-views, one for each  $t \in M$ , are stored in the preprocessing phase, while the *T*-views, one for each  $t \in V(\mathcal{T}) \setminus M$ , and the access request  $Q_A$  are accessible only in the online phase.

**THEOREM 3.7.** Consider a PMTD  $(\mathcal{T}, \chi, M, r)$  and its view  $v(\cdot)$ . Given *S*-views, we can preprocess them in space linear in their size such that we can compute the free-connex acyclic CQ

$$\psi(\mathbf{x}_H) \leftarrow Q_A \wedge \bigwedge_{t \in M} S_{v(t)} \wedge \bigwedge_{t \in V(\mathcal{T}) \setminus M} T_{v(t)} \quad (3)$$

for any *T*-view and  $Q_A$  in time  $O(\max_{t \in V(\mathcal{T}) \setminus M} |T_{v(t)}| + |Q_A| + |\psi|)$ , where  $|\psi|$  is the output size of (3).

Note that the time cost has no dependence on the size of *S*-views, because throughout Online Yannakakis, *S*-views will be only used for hash probing in semijoin operations. We defer the details of the algorithm and the proof of its correctness to Appendix A.

## 4 GENERAL FRAMEWORK

Consider a CQAP  $\phi(\mathbf{x}_H \mid \mathbf{x}_A) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F)$  with  $H \supseteq A$ . Recall that our goal is to find the best space-time tradeoffs under degree constraints DC (guarded by input relations) and AC (guarded by any access requests  $Q_A$ ), as specified in Subsection 2.2. Our main algorithm is parameterized by:

- $\mathcal{P} = \{P_i\}_{i \in I}$ , a (finite) indexed set of non-redundant and non-dominant PMTDs such that  $P_i = (\mathcal{T}_i, \chi_i, M_i, r_i)$  for every  $i \in I$ . Including all such PMTDs in  $\mathcal{P}$  (which are finite) will result in the best possible tradeoff. However, as we will see later, it is meaningful to consider smaller sets of PMTDs that result in more interpretable space-time tradeoffs.
- $S$ , the space budget.

### 4.1 2-Phase Disjunctive Rules

In this section, we define a specific type of disjunctive rule that will be necessary to acquire the *S*-views and *T*-views for PMTDs. We start by recalling the notion of a disjunctive rule. A disjunctive rule has the exact body of a CQ, while the head is a disjunction of

output relations  $T_B(\mathbf{x}_B)$ , which we call *targets*. Let  $\text{BT} \subseteq 2^{[n]}$  be a non-empty set, then a *disjunctive rule*  $\rho$  takes the form:

$$\rho : \bigvee_{B \in \text{BT}} T_B(\mathbf{x}_B) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F). \quad (4)$$

Given a database instance  $\mathcal{D}$ , a *model* of  $\rho$  is a tuple  $(T_B)_{B \in \text{BT}}$  of relations, one for each target, such that the logical implication indicated by (4) holds. More precisely, for any tuple  $\mathbf{a}$  that satisfies the body, there is a target  $T_B \in (T_B)_{B \in \text{BT}}$  such that  $\Pi_B(\mathbf{a}) \in T_B$ . The *size* of a model is defined as the maximum size of its output relations and the *output size* of a disjunctive rule  $\rho$ , denoted as  $|\rho|$ , is defined as the minimum size over all models.

For our purposes, we define a type of disjunctive rules, called *2-phase disjunctive rules*.

**Definition 4.1 (2-phase Disjunctive Rules).** A *2-phase disjunctive rule*  $\rho$  defined by a CQAP  $\varphi(\mathbf{x}_H \mid \mathbf{x}_A)$  is a single disjunctive rule that takes the body of the access CQ  $\varphi$ , while the head has two sets of output relations. In other words,  $\rho$  takes the form

$$\rho : \bigvee_{B \in \text{BS}} S_B(\mathbf{x}_B) \vee \bigvee_{B \in \text{BT}} T_B(\mathbf{x}_B) \leftarrow Q_A(\mathbf{x}_A) \wedge \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F), \quad (5)$$

where  $\text{BS}, \text{BT} \subseteq 2^{[n]}$  and at most one can be empty. A *model* of  $\rho$  thus consists of two sets of output relations, i.e. the *S-targets*  $(S_B)_{B \in \text{BS}}$  and the *T-targets*  $(T_B)_{B \in \text{BT}}$ .

As the name suggests, a model of a 2-phase disjunctive rule  $\rho$  is computed in two phases, the preprocessing and online phase:

**Preprocessing phase:** we obtain the *S-targets*  $(S_B)_{B \in \text{BS}}$  using a *preprocessing disjunctive rule*

$$\rho_S : \bigvee_{B \in \text{BS}} S_B(\mathbf{x}_B) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F), \quad (6)$$

The space cost for storing the *S-targets* is  $\tilde{O}(S_\rho)$ , and the overall space cost is  $\tilde{O}(S_\rho + |\mathcal{D}|)$ . The preprocessing phase has no knowledge of  $Q_A$  except for the degree constraints AC, so as to explicitly force the *S-targets* to be universal for any instance of access request.

**Online phase:** given an access request  $Q_A$  (under AC), we obtain the *T-targets*  $(T_B)_{B \in \text{BT}}$  using an *online disjunctive rule*

$$\rho_T : \bigvee_{B \in \text{BT}} T_B(\mathbf{x}_B) \leftarrow Q_A(\mathbf{x}_A) \wedge \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F) \quad (7)$$

in time and space  $\tilde{O}(T_\rho)$ . The overall time is  $\tilde{O}(T_\rho + |Q_A|)$ .

If  $\text{BS} = \emptyset$ , then the model is computed from scratch in the online phase (and vice versa). As in Subsection 2.2, our focus is on analyzing the space-time tradeoffs between the two intrinsic quantities,  $S_\rho$  and  $T_\rho$ .

For the next part, assume that we have a 2-phase algorithm (called 2PP) that, given a space budget  $S$ , has a preprocessing procedure 2PP-Preprocess using space  $S_\rho \leq S$  and an online procedure 2PP-Online using time (and space)  $T_\rho$ . We will discuss this algorithm in the next section.

## 4.2 Preprocessing Phase

As a first step, we construct from  $\mathcal{P}$  a set of 2-phase disjunctive rules as follows. Let  $v_i$  be the mapping for associated views of  $P_i$ . Let us define the cartesian product  $\mathbf{A} = \times_{i \in I} \{V(\mathcal{T}_i)\}$  and let

$M = |\mathbf{A}|$ . Informally, every element  $\mathbf{a} \in \mathbf{A}$  picks one view from every PMTD in the indexed set. For every  $\mathbf{a} \in \mathbf{A}$ , we construct the following 2-phase disjunctive rule (recall that  $M_i$  is the materialization set of PMTD  $(\mathcal{T}_i, \chi_i, M_i, r_i)$ ):

$$\bigvee_{\mathbf{a}_i \in M_i} S_{v_i(\mathbf{a}_i)}(\mathbf{x}_{v_i(\mathbf{a}_i)}) \vee \bigvee_{\mathbf{a}_i \notin M_i} T_{v_i(\mathbf{a}_i)}(\mathbf{x}_{v_i(\mathbf{a}_i)}) \leftarrow Q_A(\mathbf{x}_A) \wedge \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F)$$

The body of the rule is the same independent of  $\mathbf{a} \in \mathbf{A}$ . The head of the rule introduces an *S-target* whenever the corresponding bag is in the materialization set of the PMTD (using the corresponding view); otherwise, it introduces a *T-target*. There are exactly  $M$  2-phase disjunctive rules constructed from the given set of PMTDs, which is a query-complexity quantity.

**Example 4.2.** Continuing our running example, consider the three PMTDs in Figure 1. These result in four 2-phase disjunctive rules (after removing redundant *T-targets* and *S-targets*):

$$\begin{aligned} & T_{134}(x_1, x_3, x_4) \vee S_{14}(x_1, x_4) \leftarrow \text{body} \\ & T_{134}(x_1, x_3, x_4) \vee S_{13}(x_1, x_3) \vee S_{14}(x_1, x_4) \leftarrow \text{body} \\ & T_{123}(x_1, x_2, x_3) \vee T_{134}(x_1, x_3, x_4) \vee S_{14}(x_1, x_4) \leftarrow \text{body} \\ & T_{123}(x_1, x_2, x_3) \vee S_{13}(x_1, x_3) \vee S_{14}(x_1, x_4) \leftarrow \text{body} \end{aligned}$$

where

$$\text{body} = Q_{14}(x_1, x_4) \wedge R_1(x_1, x_2) \wedge R_2(x_2, x_3) \wedge R_3(x_3, x_4)$$

For each 2-phase disjunctive rule  $\rho_k$ , where  $k \in [M]$ , we run 2PP-Preprocess with the space budget  $S$ . 2PP-Preprocess generates the *S-targets* for  $\rho_k$ . Next, we compute each *S-view* of a PMTD  $P_i$  by unioning all *S-targets* with the same schema as the *S-view* (possibly from outputs of different disjunctive rules). Then, we semijoin-reduce every *S-view* with the full join  $\bowtie_{F \in \mathcal{E}} R_F$ . This semijoin-reduce can be accomplished by tentatively storing  $\bowtie_{F \in \mathcal{E}} R_F$  as an intermediate truth table and remove it after the semijoin-reduce of all *S-views*. This step guarantees that any tuple in a *S-view* participates in  $\bowtie_{F \in \mathcal{E}} R_F$ . Finally, we preprocess the *S-views* as described in Theorem 3.7.

## 4.3 Online Phase

Recall that upon receiving an instance of access request  $Q_A$ , we need to return the results of the CQ,  $\varphi(\mathbf{x}_H)$ . We obtain  $\varphi(\mathbf{x}_H)$  as follows. First, we apply 2PP-Online for every  $\rho_k$  to get its *T-targets* (of size  $\tilde{O}(T_{\rho_k})$ ) in time  $\tilde{O}(T_{\rho_k} + |Q_A|)$ . Let  $T_{\max} = \max_{k \in [M]} T_{\rho_k}$ .

Next, we compute each *T-view* of a PMTD  $P_i$  by unioning all *T-targets* with the same schema as the *T-view* (possibly from outputs of different disjunctive rules). We semijoin-reduce every *T-view* (of size  $\tilde{O}(T_\rho)$ ) with every input relation and  $Q_A$ . Then, for every PMTD in  $\{P_i\}_{i \in I}$ , we compute the free-connex acyclic CQ

$$\psi_i(\mathbf{x}_H) \leftarrow Q_A \wedge \bigwedge_{t \in M_i} S_{v_i(t)} \wedge \bigwedge_{t \in V_i(\mathcal{T}_i) \setminus M_i} T_{v_i(t)} \quad (8)$$

by applying Online Yannakakis as described in Section 3.1 in time  $\tilde{O}(T_{\max}) + O(|Q_A| + |\phi_i \cap \varphi|)$ . We obtain the final result by unioning the outputs across all PMTDs in our set,  $\varphi = \bigcup_{i \in I} \psi_i$ . In total, we answer the access request  $Q_A$  in time  $\tilde{O}(T_{\max} + |Q_A|) + O(|\varphi|)$ .

## 5 CONSTRUCTING THE TRADEOFFS

Let  $\rho$  be a 2-phase disjunctive rule taking the form (5), under degree constraints DC (guarded by input relations) and degree constraints AC (guarded by  $Q_A$ ). In this section, we will discuss how we can obtain a model of  $\rho$  in two phases using PANDA, and the resulting space-time tradeoff. Due to limited space, we will keep the presentation informal and introduce the key ideas through an example. The full details and proofs are deferred to Appendix C and D. We will use the following rule as our running example, where  $|R_1| = |R_2| = |\mathcal{D}|$ :

$$T_{123} \vee S_{13} \leftarrow Q_{13}(x_1, x_3), R_1(x_1, x_2), R_2(x_2, x_3).$$

This rule is the only rule we obtain from considering two PMTDs for the 2-reachability query. To compute a disjunctive rule, PANDA starts with a *Shannon-flow inequality*, which is an inequality over set functions  $h : 2^{[n]} \rightarrow \mathbb{R}_+$  that must hold for any set function that is a polymatroid<sup>2</sup>. For our purposes, we need a *joint Shannon-flow inequality*, which holds over two set functions  $h_S, h_T$  that must be polymatroids. Intuitively,  $h_S$  governs the preprocessing phase, while  $h_T$  governs the online phase. The joint Shannon-flow inequality for our example is:

$$\underbrace{h_S(1) + h_T(2|1)}_{R_1} + \underbrace{h_S(3) + h_T(2|3)}_{R_2} + 2\underbrace{h_T(13)}_{Q_{13}} \geq \underbrace{h_S(13)}_{S_{13}} + 2\underbrace{h_T(123)}_{T_{123}}$$

where  $h(Y|X) = h(Y) - h(X)$ . The right-hand side includes terms of  $h_S$  that correspond to  $S$ -targets and terms of  $h_T$  that correspond to  $T$ -targets. The left-hand side includes a term of  $h_T$  that corresponds to the access request  $Q_A$ , and possibly terms of  $h_S$  ( $h_T$ ) that encode the degree constraints DC (DC  $\cup$  AC). More importantly, it contains terms that correlate the two polymatroids by splitting an input relation with attributes  $Y$  into two parts, either (i)  $h_S(X) + h_T(Y|X)$ , or (ii)  $h_T(X) + h_S(Y|X)$ , where  $X \subseteq Y$ . Intuitively, the first split materializes the heavy  $X$ -values and sends everything else to the online phase, while the second split preprocesses the light  $X$ -values and sends the heavy  $X$ -values to the online phase. In our example, relation  $R_1$  is split into  $h_S(1) + h_T(2|1)$ , and each part is sent to a different polymatroid. Using the coefficients of the above joint Shannon-flow inequality, we get the following intrinsic space-time tradeoff:

$$S \cdot T^2 \cong |Q_{13}|^2 \cdot |\mathcal{D}|^2$$

We will use the  $\cong$  notation to mean that  $S \cdot T^2 = \tilde{O}(|Q_{13}|^2 \cdot |\mathcal{D}|^2)$ . Generally, we show (for a formal definition, see Theorem D.6):

**THEOREM 5.1 (INFORMAL).** *Every joint Shannon-flow inequality for a 2-phase disjunctive rule implies a space-time tradeoff computed by reading the coefficients of the inequality.*

The above theorem requires that we are given a joint Shannon-flow inequality to obtain a space-time tradeoff. We additionally show that, given a space budget  $S$ , we can also compute via a linear program the optimal inequality that will result in the best possible answering time.

**The 2PP algorithm.** We now present how our main algorithm works (see Appendix D for a detailed description). For the running example, we take  $|Q_{13}| = 1$ , and  $S$  is a fixed space budget.

<sup>2</sup>A polymatroid is a set function  $h : 2^{[n]} \rightarrow \mathbb{R}_+$  that is non-negative, monotone, and submodular, with  $h(\emptyset) = 0$ .

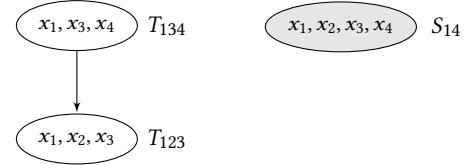


Figure 2: Two PMTDs for the square CQAP. The materialized nodes are shaded and labeled as  $S$ -views.

As a first step, 2PP scans the joint Shannon-flow inequality and partitions  $R_1(x_1, x_2)$  (on  $x_1$ ) into  $R_1^H$  and  $R_1^L$ , where  $R_1^H$  contains all  $(x_1, x_2)$  tuples where  $|\sigma_{x_1=t}(R_1)| \geq |\mathcal{D}|/\sqrt{S}$ , and  $R_1^L$  contains the tuples that satisfy  $\deg_{12}(x_2|x_1) \leq |\mathcal{D}|/\sqrt{S}$ .  $R_2$  is partitioned symmetrically (on  $x_3$ ) into  $R_2^H$  and  $R_2^L$ . This creates four subproblems,  $\{R_1^H, R_2^H\}$ ,  $\{R_1^H, R_2^L\}$ ,  $\{R_1^L, R_2^H\}$  and  $\{R_1^L, R_2^L\}$ . In general, these splits will be done according to the correlated terms in the joint flow.

The preprocessing phase (2PP-Preprocess) is governed by the Shannon-flow inequality for  $h_S$ , which is  $h_S(1) + h_S(3) \geq h_S(13)$ . We now follow PANDA and construct a *proof sequence* for this inequality. A proof sequence proves the inequality via a sequence of smaller steps, such that each step can be interpreted as a relational operator. The proof sequence for our case is:

$$\begin{aligned} h_S(1) + h_S(3) &\geq h_S(13|3) + h_S(3) && (\text{submodularity}) \\ &= h_S(13) && (\text{composition}) \end{aligned}$$

In this case, PANDA attempts to join the two relations in each subproblem. However, we allow this to happen only if the resulting space is at most  $S$ . Because  $R_1^H$  and  $R_2^H$  have size at most  $|\mathcal{D}|/(|\mathcal{D}|/\sqrt{S}) = \sqrt{S}$  values for  $x_1, x_3$  respectively, the subproblem  $\{R_1^H, R_2^H\}$  can be stored in  $S_{13}$  in space at most  $\sqrt{S} \cdot \sqrt{S} = S$ .

The online phase (2PP-Online) takes an access request  $Q_{13}(x_1, x_3)$  that contains one tuple. Now, 2PP-Online follows the second proof sequence for the polymatroid  $h_T$ :

$$\begin{aligned} h_T(2|1) + h_T(2|3) + 2h_T(13) &\geq 2h_T(2|13) + 2h_T(13) && (\text{submod.}) \\ &= 2h_T(123) && (\text{comp.}) \end{aligned}$$

For the other 3 subproblems, 2PP-Online computes the following 3 joins:  $Q_{13}(x_1, x_3) \bowtie R_2^L(x_2, x_3)$ ,  $Q_{13}(x_1, x_3) \bowtie R_1^L(x_1, x_2)$  and  $Q_{13}(x_1, x_3) \bowtie R_1^L(x_1, x_2)$ . In the submodularity step, 2PP-Online identifies that for the first join,  $\deg_{23}(x_2|x_3) \leq |\mathcal{D}|/\sqrt{S}$ , so this join takes time  $|Q_{13}| \cdot |\mathcal{D}|/\sqrt{S} \leq |\mathcal{D}|/\sqrt{S}$ ; and since  $\deg_{12}(x_2|x_1) \leq |\mathcal{D}|/\sqrt{S}$ , the last two identical joins take time  $|Q_{13}| \cdot \deg_{12}(x_2|x_1) \leq |\mathcal{D}|/\sqrt{S}$ . Therefore, the overall online computing time is  $|\mathcal{D}|/\sqrt{S}$ .

**Example 5.2 (The square query).** We now give a comprehensive example of how to construct tradeoffs for the following CQAP:

$$\varphi(x_1, x_3 \mid x_1, x_3) \leftarrow R_1(x_1, x_2) \wedge R_2(x_2, x_3) \wedge R_3(x_3, x_4) \wedge R_4(x_4, x_1).$$

This captures the following task: given two vertices of a graph, decide whether they occur in two opposite corners of a square. We consider two PMTDs. The first PMTD has a root bag  $\{1, 3, 4\}$  associated with a  $T$ -view  $T_{134}$ , and a bag  $\{1, 3, 2\}$  associated with a  $T$ -view  $T_{132}$ . The second PMTD has one bag  $\{1, 2, 3, 4\}$  associated

with an  $S$ -view  $S_{13}$ . The two PMTDs are depicted in Figure 2. This in turn generates two disjunctive rules:

$$T_{134} \vee S_{13} \leftarrow \text{body}, \quad T_{132} \vee S_{13} \leftarrow \text{body}$$

where  $\text{body} = Q_{13}(x_1, x_3) \wedge R_1(x_1, x_2) \wedge R_2(x_2, x_3) \wedge R_3(x_3, x_4) \wedge R_4(x_4, x_1)$ . We can construct the following joint Shannon-flow inequality (and its proof sequence) for the first rule:

$$\begin{aligned}
& \underbrace{\cancel{h_S(1)} + \cancel{h_T(4|1)} + \cancel{h_S(3)} + \cancel{h_T(4|3)} + 2 \cdot \cancel{h_T(13)}}_{R_4} \\
& \geq \cancel{h_S(13)} + \cancel{h_T(4|1)} + \cancel{h_T(4|3)} + 2 \cdot \cancel{h_T(13)} \\
& \geq \cancel{h_S(13)} + h_T(4|13) + h_T(13) + h_T(4|13) + h_T(13) \\
& = \underbrace{\cancel{h_S(13)} + 2 \cdot \cancel{h_T(134)}}_{S_{13}} + \underbrace{h_T(134)}_{T_{134}}.
\end{aligned}$$

For the second rule, we symmetrically construct a proof sequence for  $2 \log |\mathcal{D}| + 2 \log |Q_{13}| \geq h_S(13) + 2 \cdot h_T(132)$ . Hence, reading the coefficients of the above joint Shannon-flow inequalities, we obtain the following intrinsic space-time tradeoff  $S \cdot T^2 \cong |\mathcal{D}|^2 \cdot |Q_{13}|^2$  for the given square CQAP.

## 6 APPLICATIONS

In this section, we apply our framework to obtain state-of-the-art space-time tradeoffs for several specific problems, as well as obtain new tradeoff results. We defer the discussion for hierarchical CQAPs to the full version of the paper [38].

## 6.1 Tradeoffs for $k$ -Set Intersection

We will first study the CQAP (2) that corresponds to the non-Boolean  $k$ -Set Disjointness problem (set  $y = x_{k+1}$ )

$$\varphi(\mathbf{x}_{[k+1]} \mid \mathbf{x}_{[k]}) \leftarrow \bigwedge_{i \in [k]} R(x_{k+1}, x_i)$$

From the decomposition with a single node  $t$  with  $\chi(t) = [k+1]$ , we construct two PMTDs, one with  $M_1 = \emptyset$ , another with  $M_2 = \{t\}$ . Thus,  $v_1(t) = v_2(t) = [k+1]$ . This gives rise to the following (only) two-phase disjunctive rule:

$$T_{[k+1]} \vee S_{[k+1]} \leftarrow Q_{[k]}(\mathbf{x}_{[k]}) \wedge \bigwedge_{i \in [k]} R(x_{k+1}, x_i)$$

For this rule, we have the following joint Shannon-flow inequality:

$$\begin{aligned} h_S(k, k+1) + \sum_{i \in [k-1]} \{h_S(i|k+1) + h_T(k+1)\} + (k-1) \cdot h_T([k]) \\ \geq h_S([k+1]) + (k-1) \cdot h_T([k+1]). \end{aligned}$$

By Theorem 5.1, we get the tradeoff  $S \cdot T^{k-1} \cong |\mathcal{D}|^k \cdot |Q_A|^{k-1}$ .

## 6.2 Tradeoffs via Fractional Edge Covers

Let  $\varphi(\mathbf{x}_A \mid \mathbf{x}_B)$  be a CQAP with hypergraph  $([n], \mathcal{E})$  of  $\varphi$ . A fractional edge cover of  $S \subseteq [n]$  is an assignment  $\mathbf{u} = (u_F)_{F \in \mathcal{E}}$  such that (i)  $u_F \geq 0$ , and (ii) for every  $i \in S$ ,  $\sum_{F:i \in F} u_F \geq 1$ . For any fractional edge cover  $\mathbf{u}$  of  $[n]$ , we define the *slack* of  $\mathbf{u}$  w.r.t.  $A \subseteq [n]$ :

$$\alpha(\mathbf{u}, A) \stackrel{\text{def}}{=} \min_{i \notin A} \sum_{F \in \mathcal{E}: i \in F} u_F.$$

In other words, the slack is the maximum factor by which we can scale down the fractional cover  $\mathbf{u}$  so that it remains a valid edge

cover of the variables not in  $A$ . Hence  $(u_F/\alpha(\mathbf{u}, A))_{F \in \mathcal{E}}$  is a fractional edge cover of  $[n] \setminus A$ . We always have  $\alpha(\mathbf{u}, A) \geq 1$ .

**THEOREM 6.1.** Let  $\varphi(x_A \mid x_A)$  be a CQAP. Let  $\mathbf{u}$  be any fractional edge cover of the hypergraph of  $\varphi$ . Then, for any input database  $\mathcal{D}$ , and any access request, the following intrinsic tradeoff holds:

$$S \cdot T^{\alpha(\mathbf{u}, A)} \cong |Q_A|^{\alpha(\mathbf{u}, A)} \cdot \prod_{F \in \mathcal{E}} |R_F|^{u_F}$$

The above theorem can also be shown as a corollary of Theorem 1 in [13]. However, the data structure used in [13] is much more involved, since its goal is to also bound the delay during enumeration (while we are interested in total time instead). A simpler construction with the same tradeoff was shown in [12]. Our framework recovers the same result using a simple materialization strategy with two PMTDs.

*Example 6.2.* Consider  $\varphi(\mathbf{x}_{[k]} \mid \mathbf{x}_{[k]}) \leftarrow \bigwedge_{i \in [k]} R(y, x_i)$  (corresponds to the  $k$ -Set Disjointness problem) with the fractional edge cover  $\mathbf{u}$ , where  $u_j = 1$  for  $j \in \{1, \dots, k\}$ . The slack w.r.t.  $[k]$  is  $k$ , since the fractional edge cover  $\hat{\mathbf{u}}$ , where  $\hat{u}_i = u_i/k = 1/k$  covers  $x$ . Applying Theorem 6.1, we obtain a tradeoff of  $S \cdot T^k \cong |Q_A|^k \cdot |\mathcal{D}|^k$ . When  $|Q_A| = 1$ , this matches the best-known space-time tradeoff for the  $k$ -Set Disjointness problem.

### 6.3 Tradeoffs via Tree Decompositions

Let  $\varphi(\mathbf{x}_A \mid \mathbf{x}_B)$  be a CQAP. In the previous section, we recovered a space-time tradeoff using two trivial PMTDs. Here, we will show how our framework recovers a better space-time tradeoff by considering a larger set of PMTDs that corresponds to one decomposition.

Pick any arbitrary non-redundant free-connex decomposition  $(\mathcal{T}, \chi, r)$ . We start by taking any set of nodes that are not ancestors of each other in the decomposition as a materialization set. Then, for each node  $t$  in the materialization set, we merge all bags in the subtree of  $t$  into the bag of  $t$  (and truncate the subtree). By ranging over all such materialization sets, we construct a fixed (finite) set of PMTDs. We say that this set of PMTDs is *induced* from  $(\mathcal{T}, \chi, r)$ . We now input the induced set of PMTDs to our general framework. To discuss the obtained space-time tradeoff, take any assignment of a fractional edge cover  $\mathbf{u}_t$  to each node  $t \in V(\mathcal{T})$  and let  $u_t^*$  be its total weight. Let  $A_t$  denote the common variables between node  $t$  and its parent (for the root,  $A_r = A$ ), and define  $\alpha_t = \alpha(\mathbf{u}_t, A_t)$  to be the slack in node  $t$  w.r.t.  $A_t$ . Now, take the nodes  $P$  of any root-to-leaf path in  $\mathcal{T}$ . We can show that any such path  $P$  generates the following intrinsic tradeoff:

$$S^{\sum_{t \in P} 1/\alpha_t} \cdot T \cong |Q_A| \cdot |\mathcal{D}|^{\sum_{t \in P} u_t^*/\alpha_t}$$

To obtain the final space-time tradeoff, we take the worst such tradeoff across all root-to-leaf paths. We show in Appendix E.2 how to obtain this tradeoff, and also show why it recovers prior results [12]. Our framework guarantees that adding PMTDs to the induced set we considered here can only make this tradeoff better.

*Example 6.3.* Consider the 4-reachability CQAP. Here,  $H = A = \{x_1, x_5\}$ . We will consider the tree decomposition with bags  $t_1 = \{x_1, x_2, x_4, x_5\} \rightarrow t_2 = \{x_2, x_3, x_4\}$ .

Take the edge cover  $u_1 = 1, u_4 = 1$  for the bag  $t_1$ , and the edge cover  $u_2 = 1, u_3 = 1$  for the bag  $t_2$ . The first bag has slack  $\alpha_1 = 1$

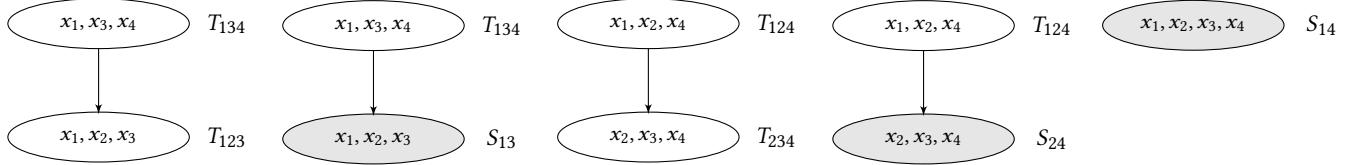


Figure 3: The PMTDs for the 3-reachability CQAP.

Table 1: 2-phase disjunctive rules for 3-reachability

rule	head	tradeoff
$\rho_1$	$T_{134} \vee T_{124} \vee S_{14}$	$S \cdot T^2 \cong  \mathcal{D} ^2 \cdot  Q_A ^2$
$\rho_2$	$T_{123} \vee S_{13} \vee T_{124} \vee S_{14}$	$\begin{cases} S^2 \cdot T^3 \cong  \mathcal{D} ^4 \cdot  Q_A ^3 \\ T \cong  \mathcal{D}  \cdot  Q_A  \end{cases}$
$\rho_3$	$T_{134} \vee T_{234} \vee S_{24} \vee S_{14}$	$\begin{cases} S^2 \cdot T^3 \cong  \mathcal{D} ^4 \cdot  Q_A ^3 \\ T \cong  \mathcal{D}  \cdot  Q_A  \end{cases}$
$\rho_4$	$T_{123} \vee S_{13} \vee T_{234} \vee S_{24} \vee S_{14}$	$\begin{cases} S \cdot T \cong  \mathcal{D} ^2 \cdot  Q_A  \\ S^4 \cdot T \cong  \mathcal{D} ^6 \cdot  Q_A  \\ T \cong  \mathcal{D}  \cdot  Q_A  \end{cases}$

(w.r.t.  $x_1, x_5$ ), while the second has slack  $\alpha_2 = 2$  (w.r.t.  $x_2, x_4$ ). Here we have one root-to-leaf path, hence we get the tradeoff  $S^{1+1/2} \cdot T \cong |Q_A| \cdot |\mathcal{D}|^{2/1+2/2}$ , or equivalently  $S^{3/2} \cdot T \cong |Q_A| \cdot |\mathcal{D}|^3$ .

#### 6.4 Tradeoffs for $k$ -Reachability

In this part, we will consider the CQAP that corresponds to the  $k$ -reachability problem described in Example 2.3:

$$\phi_k(x_1, x_{k+1} \mid x_1, x_{k+1}) \leftarrow \bigwedge_{i \in [k]} R(x_i, x_{i+1}).$$

Prior work [15] has shown the following tradeoff for a input  $\mathcal{D}$ , which was conjectured to be asymptotically optimal for  $|Q_A| = 1$ :

$$S \cdot T^{2/(k-1)} \cong |\mathcal{D}|^2 \cdot |Q_A|^{2/(k-1)}.$$

We will show that the above tradeoff can be significantly improved for  $k \geq 3$  by applying our framework.

**3-reachability.** As a first step, we consider the set of all non-redundant and non-dominant PMTDs (five in total), as seen in Figure 3. The five PMTDs will lead to  $2^4 = 16$  disjunctive rules, but we can reduce the number of rules we analyze by discarding rules with strictly more targets than other rules. For example, the disjunctive rule with head  $T_{134} \vee S_{13} \vee T_{124} \vee S_{14}$  can be ignored, since there is another disjunctive rule which has a strict subset of targets, i.e.,  $T_{134} \vee T_{124} \vee S_{14}$ . We list out the heads of the two-phase disjunctive rules we need to consider (we omit the variables for simplicity), along with the intrinsic tradeoffs for each rule in Table 1.

Note that rules can admit two (or more) tradeoffs that do not dominate each other; hence, we need to pick the best tradeoff depending on the regime we consider (see Appendix E for how we prove each tradeoff). To understand the combined tradeoff we obtain from our analysis, we plot in Figure 4a each tradeoff curve by

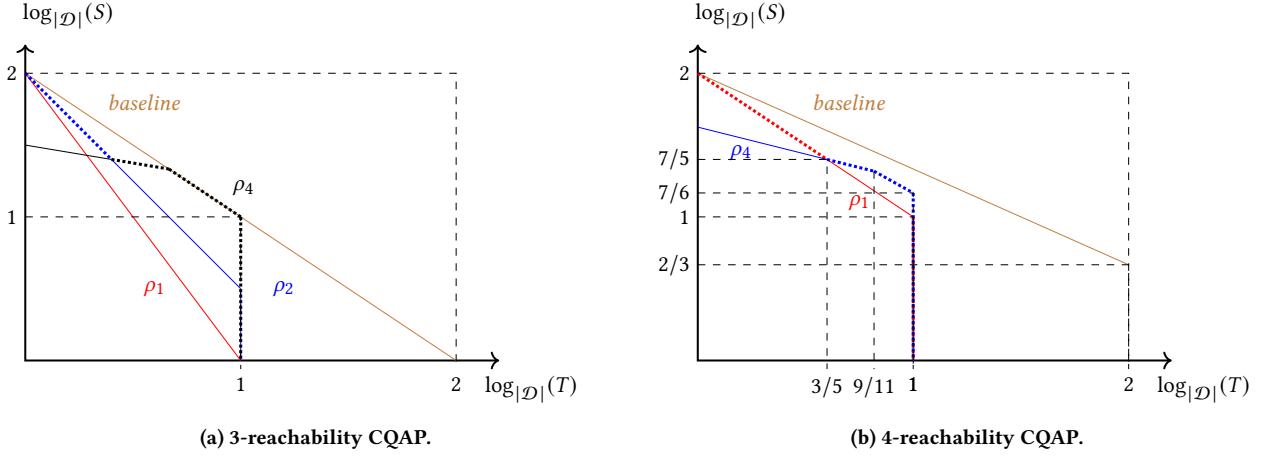
taking  $\log_{|\mathcal{D}|}$  and then taking the  $x$ -axis as  $\log_{|\mathcal{D}|}(T)$  and the  $y$ -axis as  $\log_{|\mathcal{D}|}(S)$  (fixing  $|Q_A| = 1$ ). The dotted line in the figure shows the resulting tradeoff, which is a piecewise linear function. Note that each linear segment denotes a different strategy that is optimal for that regime of space. Note that Figure 4a is not necessarily optimized for  $|Q_A| > 1$ . Suppose that  $S = |\mathcal{D}|$  and we receive  $|\mathcal{D}|$  single-tuple access requests in the online phase. Answering them one-by-one costs time  $\tilde{O}(|\mathcal{D}|^2)$  using the above tradeoffs. However, one better strategy is to batch the  $|\mathcal{D}|$  tuples (into a 4-cycle query) and use PANDA to answer it from scratch, which costs time  $\tilde{O}(|\mathcal{D}|^{3/2})$ .

**4-reachability.** We also study the CQAP for the 4-reachability problem. We leave the (quite complex) calculations to Appendix E, but we include here a plot (Figure 4b) similar to the one in Figure 4a. One surprising observation is that the new space-time tradeoff is better than the prior state-of-the-art for *every regime of space*. We should also point out that the tradeoff can possibly be further improved by including even more PMTDs (our analysis involved 11 PMTDs!), but the calculations were beyond the scope of this work.

**General reachability.** Analyzing the best possible tradeoff for  $k \geq 5$  becomes a very complex proposition. However, from Subsection 6.3 and the analysis of [12], our framework can at least obtain the  $S \cdot T^{2/(k-1)} \cong |\mathcal{D}|^2$  tradeoff, and can likely strictly improve it.

## 7 RELATED WORK

**Set Intersection and Distance Oracles.** Space-time tradeoffs for query answering (exact and approximate) has been an active area of research across multiple communities in the last decade [1, 8, 25, 26]. Cohen and Porat [8] introduced the fast intersection problem and presented a data structure to enumerate the intersection of two sets with guarantees on the total answering time. Goldstein et. al [15] formulated the  $k$ -reachability problem on graphs, and showed a simple recursive data structure which achieves the  $S \cdot T^{2/(k-1)} = O(|\mathcal{D}|^2)$  tradeoff. They also conjectured that the tradeoff is optimal and used it to justify the optimality of an approximate distance oracle proposed by [2]. The study of (approximate) distance oracles over graphs was initiated by Patrascu and Roditty [32], where lower bounds are shown on the size of a distance oracle for sparse graphs based on a conjecture about the best possible data structure for a set intersection problem. Cohen and Porat [9] also connected set intersection to distance oracles. Agarwal et al. [2, 3] introduced the notion of *stretch* of an oracle that controls the error allowed in the answer. Further, for stretch-2 and stretch-3 oracles, we can achieve tradeoffs  $S \cdot T = O(|\mathcal{D}|^2)$  and



**Figure 4: Space-time tradeoffs for the 3- and 4-reachability CQAP. The new tradeoffs obtained from our framework are depicted via the dotted segments. The brown lines (baseline) show the previous state-of-the-art tradeoffs.**

$S \cdot T^2 = O(|\mathcal{D}|^2)$  respectively, and for any integer  $k > 0$ , a stretch- $(1 + 1/k)$  oracle exhibits an  $S \cdot T^{1/k} = O(|\mathcal{D}|^2)$  tradeoff. Unfortunately, no lower bounds are known for non-constant query time.

**Space/Delay Tradeoffs.** A different line of work considers the problem of enumerating query results of a non-Boolean query, with the goal of minimizing the *delay* between consecutive tuples of the output. In constant delay enumeration [4, 33], the goal is to achieve constant delay for a CQ after a preprocessing step of linear time (and space); however, only a subset of CQs can achieve such a tradeoff. Factorized databases [31] achieve constant delay enumeration after a more expensive super-linear preprocessing step for any CQ. If we want to reduce preprocessing time further, it is necessary to increase the delay. Kara et. al [20] presented a tradeoff between preprocessing time and delay for enumerating the results of any hierarchical CQ under static (and dynamic) settings. Deng et.al [14] initiates the study of the space-query tradeoffs for range subgraph counting and range subgraph listing problems. The problem of CQs with access patterns was first introduced by Deep and Koutris [13] (under the restriction  $|Q_A| = 1$ ), but the authors only consider full CQAPs. Previous work [36] considered the problem of constructing space-efficient views of graphs to perform graph analytics, but did not offer any theoretical guarantees. More recently, Kara et. al [21] studied tradeoffs between preprocessing time, delay, and update time for CQAPs. They characterized the class of CQAPs that admit linear preprocessing time, constant delay enumeration, and constant update time. All of the above results concern the tradeoff between space (or preprocessing time) and delay, while our work focuses on the total time to answer the query. Our work is most closely related to the non-peer-reviewed work in [12]. There, the authors also study the problem of building tradeoffs for Boolean CQs. The authors propose two results that slightly improve upon [13]. They were also the first to recognize that the  $k$ -reachability tradeoff is not optimal by proposing a small improvement for  $k \geq 3$ . The results in our work are a vast generalization that is achieved using a more comprehensive

framework. For the dynamic setting, [5] initiated the study of answering CQs under updates. Recently, [19] presented an algorithm for counting the number of triangles under updates. [21] proposed dynamic algorithms for CQAPs and provided a syntactic characterization of queries that admit constant time per single-tuple update and whose output tuples can be enumerated with constant delay.

**CQ Evaluation.** Our proposed framework is based on recent advances in efficient CQ evaluation, and in particular the PANDA algorithm [24]. This powerful algorithmic result follows a long line of work on query decompositions [16, 27, 28], worst-case optimal algorithms [29], and connections between CQ evaluation and information theory [22, 23].

## 8 CONCLUSION

In this paper, we present a framework for computing general space-time tradeoffs for answering CQs with access patterns. We show the versatility of our framework by demonstrating how it can capture state-of-the-art tradeoffs for problems that have been studied separately. The application of our framework has also uncovered previously unknown tradeoffs. Many open problems remain, among which are obtaining (conditional) lower bounds that match our upper bounds, and investigating how to make our approach practical.

Many open problems remain that merit further work. In particular, there are no known lower bounds to prove the optimality of the space-time tradeoffs. The optimality of existing space-time tradeoffs for approximate distance oracles is also now an open problem again and we believe our proposed framework should be able to improve the upper bounds. It would also be very interesting to see the applicability of this framework in practice. In particular, our framework can be extended to also include views that have been precomputed, which is a common setting. In this regard, challenges remain to optimize the constants in the time complexity to ensure implementation feasibility.

## REFERENCES

- [1] Peyman Afshani and Jesper Asbjørn Sindahl Nielsen. Data structure lower bounds for document indexing problems. In *ICALP*, 2016.
- [2] Rachit Agarwal. The space-stretch-time tradeoff in distance oracles. In *ESA*, pages 49–60. Springer, 2014.
- [3] Rachit Agarwal, P Brighten Godfrey, and Sariel Har-Peled. Approximate distance queries and compact routing in sparse graphs. In *INFOCOM*, pages 1754–1762. IEEE, 2011.
- [4] Guillaume Bagan, Arnaud Durand, and Etienne Grandjean. On acyclic conjunctive queries and constant delay enumeration. In *CSL*, volume 4646 of *Lecture Notes in Computer Science*, pages 208–222. Springer, 2007.
- [5] Christoph Berkholz, Jens Keppeler, and Nicole Schweikardt. Answering conjunctive queries under updates. In *PODS*, pages 303–318. ACM, 2017.
- [6] Angela Bonifati, Wim Martens, and Thomas Timm. An analytical study of large sparql query logs. *The VLDB Journal*, 29(2):655–679, 2020.
- [7] Timothy M Chan and Moshe Lewenstein. Clustered integer 3sum via additive combinatorics. In *STOC*, pages 31–40, 2015.
- [8] Hagai Cohen and Ely Porat. Fast set intersection and two-patterns matching. *Theoretical Computer Science*, 411(40–42):3795–3800, 2010.
- [9] Hagai Cohen and Ely Porat. On the hardness of distance oracle for sparse graph. *arXiv preprint arXiv:1006.1117*, 2010.
- [10] Nilesh Dalvi, Christopher Ré, and Dan Suciu. Probabilistic databases: diamonds in the dirt. *Communications of the ACM*, 52(7):86–94, 2009.
- [11] Shaleen Deep, Xiao Hu, and Paraschos Koutris. Enumeration algorithms for conjunctive queries with projection. In *24th International Conference on Database Theory*, page 1. 2021.
- [12] Shaleen Deep, Xiao Hu, and Paraschos Koutris. Space-time tradeoffs for answering boolean conjunctive queries. *arXiv preprint arXiv:2109.10889*, 2021.
- [13] Shaleen Deep and Paraschos Koutris. Compressed representations of conjunctive query results. In *PODS*, pages 307–322. ACM, 2018.
- [14] Shiyuan Deng, Shangqi Lu, and Yufei Tao. Space-query tradeoffs in range subgraph counting and listing. In *26th International Conference on Database Theory, ICDT 2023, March 28–31, 2023, Ioannina, Greece*, pages 6:1–6:25, 2023.
- [15] Isaac Goldstein, Tsvi Kopelowitz, Moshe Lewenstein, and Ely Porat. Conditional lower bounds for space/time tradeoffs. In *WADS*, pages 421–436. Springer, 2017.
- [16] Georg Gottlob, Gianluigi Greco, and Francesco Scarsello. Treewidth and hyper-tree width. *Tractability: Practical Approaches to Hard Problems*, 1, 2014.
- [17] Gianluigi Greco and Francesco Scarsello. Structural tractability of enumerating csp solutions. *Constraints*, 18(1):38–74, 2013.
- [18] Muhammad Idris, Martin Ugarte, and Stijn Vansumeren. The dynamic yannakakis algorithm: Compact and efficient query processing under updates. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1259–1274, 2017.
- [19] Ahmet Kara, Hung Q Ngo, Milos Nikolic, Dan Olteanu, and Haozhe Zhang. Counting triangles under updates in worst-case optimal time. In *ICDT*, 2019.
- [20] Ahmet Kara, Milos Nikolic, Dan Olteanu, and Haozhe Zhang. Trade-offs in static and dynamic evaluation of hierarchical queries. In *PODS*, pages 375–392, 2020.
- [21] Ahmet Kara, Milos Nikolic, Dan Olteanu, and Haozhe Zhang. Conjunctive queries with free access patterns under updates. In *Proceedings of the 26th International Conference on Database Theory (ICDT 2023)*, 2022. The 26th International Conference on Database Theory, 2023, ICDT 2023 ; Conference date: 28-03-2023 Through 31-03-2023.
- [22] Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. Bag query containment and information theory. In *PODS*, pages 95–112. ACM, 2020.
- [23] Mahmoud Abo Khamis, Hung Q. Ngo, and Atri Rudra. FAQ: questions asked frequently. In *PODS*, pages 13–28. ACM, 2016.
- [24] Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. What do shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another? In *PODS*, pages 429–444. ACM, 2017.
- [25] Tomasz Kociumaka, Jakub Radoszewski, and Wojciech Rytter. Efficient indexes for jumbled pattern matching with constant-sized alphabet. In *ESA*, pages 625–636. Springer, 2013.
- [26] Kasper Green Larsen, J Ian Munro, Jesper Sindahl Nielsen, and Sharma V Thankachan. On hardness of several string indexing problems. *Theoretical Computer Science*, 582:74–82, 2015.
- [27] Dániel Marx. Can you beat treewidth? *Theory Comput.*, 6(1):85–112, 2010.
- [28] Dániel Marx. Tractable hypergraph properties for constraint satisfaction and conjunctive queries. *J. ACM*, 60(6):42:1–42:51, 2013.
- [29] Hung Q. Ngo, Christopher Ré, and Atri Rudra. Skew strikes back: new developments in the theory of join algorithms. *SIGMOD Rec.*, 42(4):5–16, 2013.
- [30] Dan Olteanu and Maximilian Schleich. Factorized databases. *ACM SIGMOD Record*, 45(2):5–16, 2016.
- [31] Dan Olteanu and Jakub Závodný. Size bounds for factorised representations of query results. *ACM Trans. Database Syst.*, 40(1):2:1–2:44, 2015.
- [32] Mihai Patrascu and Liam Roditty. Distance oracles beyond the thorup-zwick bound. In *FOCS*, pages 815–823. IEEE, 2010.
- [33] Luc Segoufin. Enumerating with constant delay the answers to a query. In *ICDT*, pages 10–20. ACM, 2013.
- [34] Yilei Wang and Ke Yi. Secure yannakakis: Join-aggregate queries over private data. In *SIGMOD Conference*, pages 1969–1981. ACM, 2021.
- [35] Yilei Wang and Ke Yi. Query evaluation by circuits. In *PODS '22*, pages 67–78. ACM, 2022.
- [36] Konstantinos Xirogiannopoulos and Amol Deshpande. Extracting and analyzing hidden graphs from relational databases. *CoRR*, abs/1701.07388, 2017.
- [37] Mihalis Yannakakis. Algorithms for acyclic database schemes. In *VLDB*, pages 82–94. IEEE Computer Society, 1981.
- [38] Hangdong Zhao, Shaleen Deep, and Paraschos Koutris. Space-time tradeoffs for conjunctive queries with access patterns. *arXiv preprint arXiv:2304.06221*, 2023.

## A MISSING DETAILS FROM SECTION 3

We will adapt the Yannakakis algorithm for a free-connex tree decomposition that works into two passes: the bottom-up semijoin-reduce pass and the top-down join pass. The algorithm first groups all edges by whether the edge connects two  $S$ -views, two  $T$ -views or one  $S$ -view and one  $T$ -view. We name them the  $SS$ -edges,  $TT$ -edges and  $ST$ -edges, respectively. One key observation of a PMTD is that the bottom-up order of edges on each branch is always: first some  $SS$ -edges, then at most one  $ST$ -edges followed by some  $TT$ -edges.

To preprocess the  $S$ -views, we first run a bottom-up semijoin pass on the  $SS$ -edges. Then, for each  $S$ -view, we create a hash index with search key the common variables with its (unique) parent. We now illustrate the two passes of Online Yannakakis. This allows the semijoin of a parent with a child that is an  $S$ -view to be done in time linear to the size of the parent view.

**Bottom-up Semijoin-Reduce Pass.** We first apply a bottom-up semijoin-reduce pass to remove all non-head variables in the tree by semijoins and projections. There are three scenarios as we go upwards, depending on the type of the edge we visit. Let  $(t, p) \in E(\mathcal{T})$  be the edge we are visiting, where  $t$  is the child node and  $p$  is the parent of  $t$ . We distinguish the following cases:

- (1)  $(t, p)$  is an  $SS$ -edge: we skip the edge (recall we have handled this edge during the bottom-up semijoin-reduce pass in  $M$ ).
- (2)  $(t, p)$  is an  $ST$ -edge: we update the view  $T_{v(p)} \leftarrow T_{v(p)} \ltimes S_{v(t)}$ . If every head variable in  $v(t)$  is also in  $v(p)$ , we remove  $S_{v(t)}$  from the tree.
- (3)  $(t, p)$  is a  $TT$ -edge: we update  $T_{v(p)} \leftarrow T_{v(p)} \ltimes T_{v(t)}$ . If every head variable in  $v(t)$  is also in  $v(p)$ , we remove  $T_{v(t)}$  from the tree; otherwise, we update  $v(t) \leftarrow v(t) \cap H$  and  $T_{v(t)} \leftarrow \Pi_{v(t) \cap H}(T_{v(t)})$ .

At the end of the bottom-up pass, for the root node  $r$ , if  $r \in M$ , we update  $Q_A \leftarrow Q_A \ltimes S_{v(r)}$ ; or if  $r \notin M$ , we update  $v(r) \leftarrow v(r) \cap H$ ,  $T_{v(r)} \leftarrow \Pi_{v(r) \cap H}(T_{v(r)})$  and  $Q_A \leftarrow Q_A \ltimes T_{v(r)}$ . Now, a bottom-up semi-join reducer is accomplished on the reduced tree. We prove that this reduced tree contains only head variables in Lemma A.2.

**Top-down Join Pass.** If  $r \in M$ , we compute  $Q_A \bowtie S_{v(r)}$ , or if  $r \notin M$ , we compute  $Q_A \bowtie T_{v(r)}$ . From here, we apply the exact top-down full-join pass of Yannakakis on the reduced tree (from  $r$ , use the parent view to probe the child view) to get the output  $\psi$ .

*Example A.1.* We use the non-redundant PMTD shown in Figure 5 of a CQAP  $\varphi(x_1, x_2, x_3, x_4, x_7, x_8 \mid x_1, x_2)$ , where  $(x_1, x_2)$  is the access pattern. We use the following free-connex acyclic CQ to demonstrate Online Yannakakis:

$$\psi(x_H) \leftarrow Q_{12}(x_1, x_2) \wedge T_{12}(x_1, x_2) \wedge T_{13}(x_1, x_3) \wedge T_{345}(x_3, x_4, x_5) \wedge S_{45}(x_4, x_5) \wedge S_{37}(x_3, x_7) \wedge S_{78}(x_7, x_8),$$

where  $x_H = (x_1, x_2, x_3, x_4, x_7, x_8)$ . The following is the sequence of semijoin-reduces in the bottom-up semijoin-reduce pass (the  $SS$ -edge  $(S_{37}, S_{78})$  is skipped)

$$\begin{aligned} TS\text{-edge } (T_{345}, S_{45}) : & \quad T_{345}^{(1)} \leftarrow T_{345} \ltimes S_{45}, \quad \text{remove } S_{45} \\ TS\text{-edge } (T_{13}, S_{37}) : & \quad T_{13}^{(1)} \leftarrow T_{13} \ltimes S_{37} \\ TT\text{-edge } (T_{13}, T_{345}) : & \quad T_{13}^{(2)} \leftarrow T_{13} \ltimes T_{345}^{(1)}, \quad T_{34}^{(1)} \leftarrow \Pi_{34}(T_{345}^{(1)}) \\ TT\text{-edge } (T_{12}, T_{13}) : & \quad T_{12}^{(1)} \leftarrow T_{12} \ltimes T_{13}^{(2)} \\ \text{root : } & \quad Q_A^{(1)} \leftarrow Q_A \ltimes T_{12}^{(1)} \end{aligned}$$

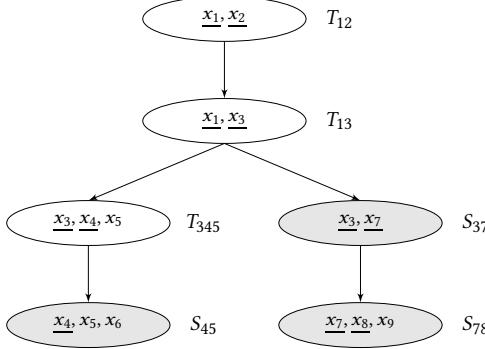
In the top-down pass, to get the result of  $\psi$ , Online Yannakakis computes the following joins from the root to the leaves, starting from  $Q_A^{(1)}$

$$Q_A^{(1)} \bowtie T_{12}^{(1)} \bowtie T_{13}^{(2)} \bowtie T_{34}^{(1)} \bowtie S_{37} \bowtie S_{78}.$$

**LEMMA A.2.** *The reduced tree after the bottom-up semijoin-reduce pass of the Online Yannakakis contains only head variables, i.e.  $x_H$ .*

**PROOF.** Obviously,  $T$ -views in the reduced tree contain only head variables. Therefore, we only need to show the property for every  $S$ -view. This is obvious for a root  $S$ -view or a non-root  $S$ -view that has a parent  $S$ -view by definition of  $v(\cdot)$ . We are left to show for a  $ST$ -edge  $(t, p)$  such that there is a head vertex  $y \in v(t) \setminus v(p)$ , which indicates that  $t$  is the top-most bag to contain  $y$ . Suppose the  $S$ -view  $S_{v(t)}$  contains some  $z \notin H$ , then  $z$  must be in  $v(p) = \chi(p)$  by definition of  $v(\cdot)$ . This contradicts the free-connex property since  $\text{TOP}_r(z)$  is an ancestor of  $\text{TOP}_r(y)$ , i.e. a non-head vertex  $z \in v(p)$  is above the head vertex  $y \in v(t)$ .  $\square$

**PROOF OF THEOREM 3.7.** Define  $T = \max_{t \in V(\mathcal{T}) \setminus M} |T_{v(t)}|$ . The bottom-up pass costs time  $O(T + |Q_A|)$  as only  $T$ -views and  $Q_A$  are semijoin-reduced ( $S$ -views, by the index construction, are also bottom-up semijoin-reduced). Moreover, by Lemma A.2, the reduced tree after the bottom-up pass contains only variables in  $H$ . So, joining top-down from the root circumvents any intermediate dangling tuples and costs time  $O(|\psi|)$ . The overall time cost is  $O(T + |Q_A| + |\psi|)$ .  $\square$



**Figure 5:** A non-redundant PMTD as an example for Online Yannakakis. The materialization set is shaded (and marked as S-views) and the head variables are underlined.

## B MISSING DETAILS FROM SECTION 4

We show that the general framework, for any access request  $Q_A$ , returns the correct results  $\varphi$ , i.e.  $\varphi = \bigcup_{i \in I} \psi_i$ . As a benefit from semijoin-reduces of every view in a PMTD, it is obvious that  $\bigcup_{i \in I} \psi_i \subseteq \varphi$ . For the opposite inclusion, we prove the following two claims, following the proof of Corollary 7.13 in [24].

**Claim 1.** Let us pick out one target (either a  $S$ -target or a  $T$ -target), denote the target as  $U_k$ , from the head of every rule  $\rho_k$ , where  $k \in [M]$ . We call the tuple  $(U_k)_{k \in [M]}$  that consists of one picked target per rule a *full-choice*. Then, for any *full choice*  $(U_k)_{k \in [M]}$ , there is a PMTD  $P \in \{P_i\}_{i \in I}$  such that for every tree node  $t \in M$ , the  $S$ -target  $S_{v(t)}$  is in the full-choice and for every tree node  $t \notin M$ , the  $T$ -target  $T_{v(t)}$  is in the full-choice. Breaking ties arbitrarily, we call this PMTD *the* PMTD associated with the full choice  $(U_k)_{k \in [M]}$ .

**PROOF OF CLAIM 1.** Fix a full-choice  $(U_k)_{k \in [M]}$ . Suppose to the contrary that for every PMTD  $P_i \in \mathcal{P}$ , there is a tree node such that its corresponding target (either a  $T$ -target or a  $S$ -target), denoted as  $U_j^*$ , is not in the full-choice. Then we examine the 2-phase disjunctive rule  $\rho^*$  that takes  $\bigvee_{i \in I} U_i^*$  as its head. By definition of a full-choice, one target must be picked from  $\rho^*$ . However, this is a contradiction because every head  $U_i^*$  in  $\rho^*$  does not show up in the fixed full-choice  $(U_k)_{k \in [M]}$ .  $\square$

**Claim 2.** For any full-choice  $U \stackrel{\text{def}}{=} (U_k)_{k \in [M]}$  with its associated PMTD  $(\mathcal{T}_i, \chi_i, M_i, r_i)$ , we define a CQ

$$\varphi_U(x_H) \leftarrow Q_A \wedge \bigwedge_{k \in [M]} U_k$$

Let  $\mathcal{U}$  denote the set of all full-choices. Then,

$$\varphi \subseteq \bigcup_{U \in \mathcal{U}} \varphi_U \subseteq \bigcup_{i \in I} \psi_i \tag{9}$$

**PROOF OF CLAIM 2.** Take any output tuple  $a_H \in \varphi$ . Then, there is some tuple  $a$  satisfying the body of  $\varphi$  such that  $a_H = \Pi_H(a)$ . For each rule  $\rho_k$ , where  $k \in [M]$ , let  $U_k^*$  be the target (either  $S$ -target or  $T$ -target) associated with the view  $v_k^*$  such that  $\Pi_{v_k^*}(a) \in U_k^*$ . Therefore,  $U^* = (U_k^*)_{k \in M}$  is a full-choice and  $a$  satisfies the body of  $\varphi_{U^*}$ . Hence,  $a_H \in \varphi_{U^*}$  and we have shown the first inclusion in (9). The second inclusion follows by dropping the atoms in the body of  $\varphi_{U^*}$  that is not any view of  $U^*$ 's associated PMTD.  $\square$

## C ALGORITHMS FOR 2-PHASE DISJUNCTIVE RULES

Let  $\rho$  be a 2-phase disjunctive rule taking the form (5), under degree constraints DC (guarded by input relations) and degree constraints AC (guarded by access request  $Q_A$ ). In this section, we introduce a naïve algorithm that uses the PANDA algorithm to obtain a model of  $\rho$  in two phases. First, we present some necessary terminologies and results.

### C.1 Background

**Entropic Functions.** Given a disjunctive rule (4), a set function  $h : 2^{[n]} \rightarrow \mathbb{R}_+$  is *entropic* if there is a joint probability distribution on  $[n]$  such that  $h(F)$  is the marginal entropy of  $F$  for any  $F \subseteq [n]$ . Let  $\Gamma_n^*$  be the set of all entropic functions and  $h(Y|X) \stackrel{\text{def}}{=} h(Y) - h(X)$ . Under a given set of degree constraints DC, any joint distribution on  $[n]$  conforms to the constraints  $h(Y|X) \leq n_{Y|X}$ , where  $n_{Y|X} \stackrel{\text{def}}{=} \log N_{Y|X}$ , for each  $(X, Y, N_{Y|X}) \in \text{DC}$ .

**Polymatroid.** A polymatroid is a set function  $h : 2^{[n]} \rightarrow \mathbb{R}_+$  that is non-negative, monotone, and submodular, with  $h(\emptyset) = 0$ . To be precise, monotonicity implies that  $h(Y) \geq h(X)$  for any  $X \subseteq Y \subseteq [n]$  and let  $h(Y|X) \stackrel{\text{def}}{=} h(Y) - h(X)$ , then submodularity implies that  $h(I|I \cap J) \geq h(I \cup J|J)$  for any  $I, J \subseteq [n]$ . Let  $\Gamma_n$  be the set of all polymatroids on  $[n]$ . As every entropic function is a polymatroid, it holds that  $\Gamma_n^* \subseteq \Gamma_n$ .

**Size Bound for Disjunctive Rules.** Let  $\rho$  be a disjunctive rule of the form (4). Let  $\mathcal{D}$  be a database instance under a given set of degree constraints DC. The set

$$\text{HDC} \stackrel{\text{def}}{=} \left\{ h : 2^{[n]} \rightarrow \mathbb{R}_+ \mid \bigwedge_{(X, Y, N_{Y|X}) \in \text{DC}} h(Y|X) \leq \log N_{Y|X} \right\}$$

contains all entropic functions  $h$  on  $[n]$  satisfying the degree constraints DC. Fix a closed subset  $\mathcal{F} \subseteq \mathbb{R}_+^{2^n}$ . We define the log-size-bound with respect to  $\mathcal{F}$  of a disjunctive rule  $\rho$  to be the quantity:

$$\text{LogSizeBound}_{\mathcal{F}}(\rho) \stackrel{\text{def}}{=} \max_{h \in \mathcal{F}} \min_{B \in \text{BT}} h(B)$$

Then for the output size of  $\rho$ , we have the following theorem (see Theorem 1.5 in [24]):

**THEOREM C.1 ([24]).** *Let  $\rho$  be any disjunctive rule (4) under degree constraints DC. Then for any database instance  $\mathcal{D}$  satisfying DC, the following holds:*

$$\log |\rho| \leq \underbrace{\text{LogSizeBound}_{\Gamma_n^* \cap \text{HDC}}(\rho)}_{\text{entropic bound}} \leq \underbrace{\text{LogSizeBound}_{\Gamma_n \cap \text{HDC}}(\rho)}_{\text{polymatroid bound}},$$

The entropic bound is tight under degree constraints in the worst case. However, its computation is often hard in general. The polymatroid bound is tight if  $\rho$  is a CQ (i.e. has a single target) and there are only cardinality constraints, in which case the polymatroid bound degenerates into the AGM bound. However, it is not tight under general degree constraints.

**The PANDA Algorithm.** Given a disjunctive rule  $\rho$  of the form (4), the PANDA algorithm takes a database instance  $\mathcal{D}$ , a set of degree constraint DC (guarded by  $\mathcal{D}$ ) as inputs and computes a model in time and space predicted by its polymatroid bound:

$$\tilde{O}(2^{\text{LogSizeBound}_{\Gamma_n \cap \text{HDC}}(\rho)}).$$

The reader can refer to Theorem 1.7 in [24] for details. For now, we will treat PANDA algorithm as a blackbox; we will later present how it works.

### C.2 The 2-phase Framework

In this section, we introduce a 2-phase algorithmic framework that we will follow to design 2-phase algorithms for a 2-phase disjunctive rule of the form (5).

Let  $\mathcal{D}$  be a database instance and  $Q_A$  be an arbitrary access request. We assume that *hash tables on necessary index keys (of input relations and degrees of tuples in input relations)* can be pre-built at the start of the preprocessing phase as needed by the framework. That is, we assume constant-time accesses of tuples and degrees of tuples in the input relations during both phases. There are at most  $O(2^{2n})$  hash tables to be pre-built per input relation  $R_F$  (that is, for every  $(Y, X)$ -pair where  $X \subset Y \subseteq F \in \mathcal{E}$ ), so the space cost for storing all necessary hash tables is  $O(|\mathcal{D}|)$  in data complexity. Recall that we also assume w.l.o.g the following *best constraint assumption*: we only keep at most one  $(X, Y, N_{Y|X}) \in \text{DC}$  for each  $X \subset Y \subseteq [n]$  (keep the minimum  $N_{Y|X}$  if there are more than one).

**Split Steps.** Let  $R \in \mathcal{D}$  be the guard of a cardinality constraint  $(\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}$ . A *split step* on a  $(Y, X)$ -pair, where  $\emptyset \neq X \subset Y \subseteq Z$ , applies Lemma 6.1 of [24] and partitions  $R_Y \stackrel{\text{def}}{=} \Pi_Y(R)$  into  $k = 2 \log N_{Z|\emptyset}$  sub-tables, i.e.  $R_Y^{(1)}, \dots, R_Y^{(k)}$ , such that  $N_{X|\emptyset}^{(j)} \cdot N_{Y|X}^{(j)} \leq N_{Z|\emptyset}$ , for all  $j \in [k]$ , where

$$N_{X|\emptyset}^{(j)} \stackrel{\text{def}}{=} |\Pi_X(R_Y^{(j)})|$$

$$N_{Y|X}^{(j)} \stackrel{\text{def}}{=} \deg_{R_Y^{(j)}}(Y|X).$$

For each of these sub-tables  $R_Y^{(j)}$ , we create a subproblem with inputs  $(\mathcal{D}^{(j)}, \text{DC}^{(j)})$ , where  $\mathcal{D}^{(j)} \stackrel{\text{def}}{=} \mathcal{D} \cup \{R_Y^{(j)}\}$  denotes the input tables and

$$\text{DC}^{(j)} \stackrel{\text{def}}{=} \text{DC} \cup \{(\emptyset, X, N_{X|\emptyset}^{(j)}), (X, Y, N_{Y|X}^{(j)})\}$$

denotes the degree constraints guarded by  $\mathcal{D}^{(j)}$ . A sequence of split steps on  $(Y_1, X_1), (Y_2, X_2), \dots, (Y_\ell, X_\ell)$ , applies the first split step on the  $(Y_1, X_1)$ -pair, generating  $k_1 = O(\log |\mathcal{D}|)$  subproblems with inputs  $(\mathcal{D}^{(j)}, \text{DC}^{(j)})$ , where  $j \in [k_1]$ . Then, for each subproblem, applies the second split step on the  $(Y_2, X_2)$ -pair, generating  $O((\log |\mathcal{D}|)^2)$  subproblems. This iterative process goes on until every split step in the sequence is applied, thus it generates  $O(\text{poly}(\log |\mathcal{D}|))$  subproblems in total.

**The 2-phase Framework.** Now we formally characterize our 2-phase algorithmic framework. Let  $S$  be the given space budget. We denote the task of obtaining a model for a 2-phase disjunctive rule  $\rho$  (of the form (5)) with input relations  $\mathcal{D} \cup \{Q_A\}$  satisfying degree constraints  $\text{DC} \cup \text{AC}$  as  $\rho(\mathcal{D} \cup \{Q_A\}, \text{DC} \cup \text{AC})$ . The framework starts by using a sequence of *split steps* to partition  $\rho(\mathcal{D} \cup \{Q_A\}, \text{DC} \cup \text{AC})$  into  $O(\text{poly}(\log |\mathcal{D}|))$  subproblems. Then, the  $j$ -th subproblem with input  $\mathcal{D}^{(j)}$  and degree constraint  $\text{DC}^{(j)} \supseteq \text{DC}$ , denoted as  $\rho(\mathcal{D}^{(j)} \cup \{Q_A\}, \text{DC}^{(j)} \cup \text{AC})$ , either

- (1) generates  $S$ -targets  $(S_B^{(j)})_{B \in \text{BS}}$  as a model of the preprocessing disjunctive rule  $\rho_S$  of the form (6) using PANDA, provided that the output size of  $\rho_S$  is within  $\tilde{O}(S)$ ; or
- (2) generates  $T$ -targets  $(T_B^{(j)})_{B \in \text{BT}}$  as a model of the online disjunctive rule  $\rho_T$  of the form (7) using PANDA.

In other words, the subproblem  $\rho(\mathcal{D}^{(j)} \cup \{Q_A\}, \text{DC}^{(j)} \cup \text{AC})$  is conquered by applying PANDA to obtain either a model  $\rho_S$  with input relations  $\mathcal{D}^{(j)}$  under degree constraint  $\text{DC}^{(j)}$ , denoted as  $\rho_S(\mathcal{D}^{(j)}, \text{DC}^{(j)})$ , or a model of  $\rho_T$  with input relations  $\mathcal{D}^{(j)} \cup \{Q_A\}$  under degree constraint  $\text{DC}^{(j)} \cup \text{AC}$ , denoted as  $\rho_T(\mathcal{D}^{(j)} \cup \{Q_A\}, \text{DC} \cup \text{AC})$ . After all subproblems are computed, the model of  $\rho$  is simply the union over  $S$ -targets and  $T$ -targets generated from all subproblems.

**Analysis of the 2-phase Framework.** Next, we analyze the intrinsic space-time tradeoff (specified in Subsection 4.1, between  $S_\rho$  and  $T_\rho$ ) that can be obtained by the 2-phase framework introduced above. First, the split steps incur a poly-logarithmic factor on both  $S_\rho$  and  $T_\rho$  and spawn  $O(\text{poly}(\log |\mathcal{D}|))$  subproblems. Then, the  $j$ -th spawned subproblem  $\rho(\mathcal{D}^{(j)} \cup \{Q_A\}, \text{DC}^{(j)} \cup \text{AC})$  is conquered by PANDA in one of the two phases. Note that  $\text{DC}^{(j)} \supseteq \text{DC}$  contains extra degree constraints due to the split steps. For ease of analysis, we separate out the extra constraints by defining a set  $\text{SC}^{(j)} \stackrel{\text{def}}{=} \text{DC}^{(j)} \setminus \text{DC}$ . To apply the polymatroid bound, we use two polymatroids,  $\textcolor{blue}{h}_S \in \Gamma_n$  to represent the preprocessing phase and  $\textcolor{red}{h}_T \in \Gamma_n$  to represent the online phase. We define

$$\begin{aligned} \text{HDC} &\stackrel{\text{def}}{=} \left\{ h : 2^{[n]} \rightarrow \mathbb{R}_+ \mid \bigwedge_{(X, Y, N_{Y|X}) \in \text{DC}} h(Y|X) \leq \log N_{Y|X} \right\} \\ \text{HSC}^{(j)} &\stackrel{\text{def}}{=} \left\{ h : 2^{[n]} \rightarrow \mathbb{R}_+ \mid \bigwedge_{(X, Y, N_{Y|X}) \in \text{SC}^{(j)}} h(Y|X) \leq \log N_{Y|X} \right\} \\ \text{HAC} &\stackrel{\text{def}}{=} \left\{ h : 2^{[n]} \rightarrow \mathbb{R}_+ \mid \bigwedge_{(X, Y, N_{Y|X}) \in \text{AC}} h(Y|X) \leq \log N_{Y|X} \right\} \end{aligned}$$

where  $h(Y|X) = h(Y) - h(X)$ , to denote collections of set functions that satisfy DC,  $\text{SC}^{(j)}$  and AC, respectively. The 2-phase framework enforces that  $\textcolor{blue}{h}_S$  conforms to  $\text{HDC} \cap \text{HSC}^{(j)}$  and  $\textcolor{red}{h}_T$  conforms to  $\text{HDC} \cap \text{HSC}^{(j)} \cap \text{HAC}$ . Thus, the  $j$ -th subproblem costs space  $\tilde{O}(S_\rho^{(j)})$  in the preprocessing phase, where

$$\log S_\rho^{(j)} \stackrel{\text{def}}{=} \text{LogSizeBound}_{\textcolor{blue}{h}_S \in \Gamma_n \cap \text{HDC} \cap \text{HSC}^{(j)}}(\rho_S) \quad (10)$$

provided that  $S_\rho^{(j)} \leq S$ . Otherwise,  $j$ -th subproblem costs time (and space)  $\tilde{O}(T_\rho^{(j)})$ , where

$$\log T_\rho^{(j)} \stackrel{\text{def}}{=} \text{LogSizeBound}_{\textcolor{red}{h}_T \in \Gamma_n \cap \text{HDC} \cap \text{HSC}^{(j)} \cap \text{HAC}}(\rho_T) \quad (11)$$

By conquering all subproblems, we conclude that  $S_\rho = \max_j S_\rho^{(j)} \leq S$  and  $T_\rho = \max_j T_\rho^{(j)}$ .

### C.3 A Naïve 2-phase Algorithm

In this section, we use the 2-phase framework to design a 2-phase algorithm for a 2-phase disjunctive rule  $\rho$  that attains the smallest possible  $T_\rho$  for a fixed space budget  $S$ .

Recall that for each  $(\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}$ , there are at most  $2^{2n}$   $(Y, X)$ -pairs with  $\emptyset \neq X \subset Y \subseteq Z$ . Thus, the total number of distinct split steps is a constant in data complexity. To exploit the full potential of split steps, we design a naïve 2-phase algorithm that applies a sequence of all distinct split steps. Also, recall that a finite sequence of split steps spawns  $O(\text{poly}(\log |\mathcal{D}|))$  subproblems.

**The Naïve Algorithm.** As said, the naïve algorithm first applies a sequence of all distinct split steps. Intuitively, this partitions  $\mathcal{D}$  into its most fine-grained pieces. Let  $\rho(\mathcal{D}^{(j)} \cup \{Q_A\}, \text{DC}^{(j)} \cup \text{AC})$  be the  $j$ -th subproblem spawned after the sequence of all distinct split steps. Recall that  $\text{SC}^{(j)} = \text{DC}^{(j)} \setminus \text{DC}$ . The following *splitting property* is a direct result of a sequence of all distinct split steps on  $\text{DC}$ : for any  $(\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}$  and  $(Y, X)$ -pair with  $\emptyset \neq X \subset Y \subseteq Z$ , there are some  $(\emptyset, X, N_{X|\emptyset}^{(j)})$ ,  $(X, Y, N_{Y|X}^{(j)}) \in \text{SC}^{(j)}$  such that  $N_{X|\emptyset}^{(j)} \cdot N_{Y|X}^{(j)} \leq N_{Z|\emptyset}$ . Though each subproblem varies in its own  $\text{SC}^{(j)}$ , the splitting property holds across all subproblems. To encode the splitting property, we define split constraints.

**Definition C.2 (Split Constraints).** Let  $\text{DC}$  be a set of degree constraints. A *split constraint* is a triple  $(X, Y|X, N_{Z|\emptyset})$  where  $\emptyset \neq X \subset Y \subseteq Z$ ,  $(\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}$ . A relation  $R_F$  is said to guard a split constraint  $(X, Y|X, N_{Z|\emptyset})$  if  $R_F$  guards  $(\emptyset, Z, N_{Z|\emptyset})$ . The set of all split constraints spanned from  $\text{DC}$ , is denoted as

$$\text{SC} \stackrel{\text{def}}{=} \{(X, Y|X, N_{Z|\emptyset}) \mid \emptyset \neq X \subset Y \subseteq Z, (\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}\}.$$

Intuitively, each triple  $(X, Y|X, N_{Z|\emptyset}) \in \text{SC}$  encodes the splitting property for the  $(Y, X)$ -pair on  $(\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}$ . Since we assume that every  $(\emptyset, Z, N_{Z|\emptyset}) \in \text{DC}$  has at least one guard, every  $(X, Y|X, N_{Z|\emptyset}) \in \text{SC}$  is guarded by some  $R_F \in \mathcal{D}$ .

The naïve algorithm stores  $S$ -views for  $\rho_S(\mathcal{D}^{(j)}, \text{DC}^{(j)})$  whenever its polymatroid bound, as specified in (10), is no larger than  $\log S$ . Otherwise, it applies PANDA algorithm as a black box for  $\rho_T(\mathcal{D}^{(j)} \cup \{Q_A\}, \text{DC}^{(j)} \cup \text{AC})$  in the online phase in time as specified in (11).

**Analysis of the Naïve Algorithm.** The naïve algorithm exploits the full potential of the framework (by exhausting all possible split steps) and gets the best possible  $T_\rho$  when  $S_\rho \leq S$  (up to a poly-logarithmic factor). In particular, we state the following theorem for the best possible  $T_\rho$ :

**THEOREM C.3.** *For a 2-phase disjunctive rule (5), under space budget  $S$ , the naïve algorithm obtains  $T_\rho = 2^{\text{OBJ}(S)}$ , where*

$$\begin{aligned} \text{OBJ}(S) = & \max_{\substack{\mathbf{h}_S \in \text{HDC}, \mathbf{h}_T \in \text{HDC} \cap \text{HAC} \\ (\mathbf{h}_S, \mathbf{h}_T) \in (\Gamma_n \times \Gamma_n) \cap \text{HSC}}} \min_{B \in \text{BT}} \mathbf{h}_T(B) \\ \text{s.t.} & \quad \mathbf{h}_S(B) > \log S, \quad B \in \text{BS}, \end{aligned} \tag{12}$$

assuming that  $\text{OBJ}(S)$  is positive and bounded.

**PROOF.** Merging (10) and (11), we get that  $\log T_\rho^{(j)}$  for the  $j$ -th subproblem can be expressed as

$$\begin{aligned} \log T_\rho^{(j)} = & \max_{\substack{\mathbf{h}_S \in \text{HDC} \cap \text{HSC}^{(j)} \\ \mathbf{h}_T \in \text{HDC} \cap \text{HSC}^{(j)} \cap \text{HAC} \\ \mathbf{h}_S, \mathbf{h}_T \in \Gamma_n}} \min_{B \in \text{BT}} \mathbf{h}_T(B) \\ \text{s.t.} & \quad \mathbf{h}_S(B) > \log S, \quad B \in \text{BS}. \end{aligned} \tag{13}$$

Note that the maximin optimization (13) is subproblem-dependent, since it is constrained on  $\text{HSC}^{(j)}$ . To avoid this dependency, we recall that the naïve algorithm dictates the splitting property. Thus, we define the (subproblem-independent) set  $\text{SC}$  as follows:

$$\text{HSC} \stackrel{\text{def}}{=} \left\{ (\mathbf{h}_S, \mathbf{h}_T) : 2^{[n]} \times 2^{[n]} \rightarrow \mathbb{R}_+^2 \mid \bigwedge_{(X, Y|X, N_{Z|\emptyset}) \in \text{SC}} (\mathbf{h}_S(X) + \mathbf{h}_T(Y|X) \leq \log N_{Z|\emptyset}) \wedge (\mathbf{h}_S(Y|X) + \mathbf{h}_T(X) \leq \log N_{Z|\emptyset}) \right\},$$

where  $\mathbf{h}_S(Y|X) \stackrel{\text{def}}{=} \mathbf{h}_S(Y) - \mathbf{h}_S(X)$ ,  $\mathbf{h}_T(Y|X) \stackrel{\text{def}}{=} \mathbf{h}_T(Y) - \mathbf{h}_T(X)$ .  $\text{HSC}$  is a universal collection of set functions pairs satisfying the splitting property and thus, it correlates  $\mathbf{h}_S$  and  $\mathbf{h}_T$ . Since every subproblem satisfies the splitting property, it holds that  $\text{HSC}^{(j)} \times \text{HSC}^{(j)} \subseteq \text{HSC}$ . By relaxing  $\text{HSC}^{(j)} \times \text{HSC}^{(j)}$  to  $\text{HSC}$ , we get the desired upper bound (12) for  $T_\rho$ .  $\square$

By assigning  $\mathbf{h}_T$  to be always 0, it is easy to see that the feasible region of (12) is empty if and only if

$$\text{LogSizeBound}_{\mathbf{h}_S \in \Gamma_n \cap \text{HDC}}(\rho_S) \leq \log S,$$

in which case we can simply store the  $S$ -views within space  $\tilde{O}(S)$ . Otherwise, the feasibility of (12) is guaranteed. However, the naïve algorithm has the following drawbacks in terms of practicality: (1) the exhaustive splitting steps can incur a large poly-logarithmic factor; (2) for every subproblem, we need to run PANDA from scratch with a new instance, (3) the space-time tradeoff obtained is hard to interpret. In the following sections, we introduce the 2-phase PANDA algorithm, called 2PP, that also attains the intrinsic tradeoff as specified in (12), while efficiently addressing these two drawbacks and it obtains much practical/interpretable intrinsic tradeoff(s).

## D THE 2PP ALGORITHM

The 2-phase PANDA (2PP) algorithm, similar to PANDA, is built on a class of inequalities called the joint Shannon-flow inequalities. In this section, we first present some background on Shannon-flow inequalities, and then our extension to joint inequalities. Finally, we present the 2PP algorithm.

### D.1 Shannon-flow Inequalities

The following inequality

$$\sum_{X \subset Y \subseteq [n]} \delta_{Y|X} \cdot h(Y|X) \stackrel{\text{def}}{=} \sum_{X \subset Y \subseteq [n]} \delta_{Y|X} \cdot (h(Y) - h(X)) \geq \sum_{\emptyset \neq Z \subseteq [n]} \lambda_{Z|\emptyset} \cdot h(Z|\emptyset), \quad (14)$$

is called a *Shannon-flow inequality* if it holds for any polymatroid function  $h \in \Gamma_n$  and all  $\delta_{Y|X}$  and  $\lambda_{Z|\emptyset}$  are some non-negative rational coefficients, i.e.,  $\delta_{Y|X}, \lambda_{Z|\emptyset} \in \mathbb{Q}_+$ . To concisely represent Shannon-flow inequalities, we define the conditional polymatroid as in [24].

**Conditional Polymatroids.** Let  $C \subseteq 2^{[n]} \times 2^{[n]}$  denote the set of all pairs  $(X, Y)$  such that  $\emptyset \subseteq X \subset Y \subseteq [n]$ . A vector  $\mathbf{h} \in \mathbb{R}_+^C$  has coordinates indexed by pairs  $(X, Y) \in C$  and we denote the corresponding coordinate value of  $\mathbf{h}$  by  $h(Y|X)$ . A vector  $\mathbf{h}$  is called a *conditional polymatroid* if and only if there is a polymatroid  $h$  such that  $h(Y|X) = h(Y) - h(X)$ ; and, we say that the polymatroid  $h$  defines the conditional polymatroid  $\mathbf{h}$ . In particular,  $\mathbf{h} = (h(Y|X))_{(X,Y) \in C}$ .

If for (14), we define  $\lambda_{Z|X} = 0$  for any  $\emptyset \neq X \subset Z \subseteq [n]$ , then each of  $\{\delta_{Y|X}\}, \{\lambda_{Z|\emptyset}\}$  can be interpreted as vectors over  $(X, Y)$  pairs, where  $\emptyset \subseteq X \subset Y \subseteq [n]$ . We denote them as  $\boldsymbol{\delta}, \boldsymbol{\lambda}$ , respectively. Thus, the Shannon-flow inequality (14) can be re-written into an inequality on conditional polymatroids, i.e. the  $\mathbb{Q}_+^C$  space, as  $\langle \boldsymbol{\delta}, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$ , where  $\langle \cdot, \cdot \rangle$  denotes dot product.

**Proof Sequences.** One of the major contributions from [24] says that any Shannon-flow inequality  $\langle \boldsymbol{\delta}, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$  can be proved by just applying the following 4 rules:

(R1) submodularity rule	$h(I \cup J J) - h(I I \cap J) \leq 0, \quad I \perp J$
(R2) monotonicity rule	$-h(Y \emptyset) + h(X \emptyset) \leq 0, \quad X \subset Y$
(R3) composition rule	$h(Y \emptyset) - h(Y X) - h(X \emptyset) \leq 0, \quad X \subset Y$
(R4) decomposition rule	$-h(Y \emptyset) + h(Y X) + h(X \emptyset) \leq 0, \quad X \subset Y$

where  $I \perp J$  means  $I \not\subseteq J$  and  $J \not\subseteq I$ . (R1) and (R2) come exactly from the submodularity and monotonicity properties of polymatroids. (R3) and (R4) simply follow from the definition of  $h(Y|X) = h(Y) - h(X)$ . All the rules can also be vectorized over all  $(X, Y)$  pairs, where  $\emptyset \subseteq X \subset Y \subseteq [n]$ . For every  $I \perp J$ , we define a vector  $\mathbf{s}_{I,J}$ , and for every  $X \subset Y$ , we define three vectors  $\mathbf{m}_{X,Y}, \mathbf{c}_{X,Y}, \mathbf{d}_{Y,X}$  such that the linear rules above can be written using dot-products:

(R1) submodularity rule	$\langle \mathbf{s}_{I,J}, \mathbf{h} \rangle \leq 0, \quad I \perp J$
(R2) monotonicity rule	$\langle \mathbf{m}_{X,Y}, \mathbf{h} \rangle \leq 0, \quad X \subset Y$
(R3) composition rule	$\langle \mathbf{c}_{X,Y}, \mathbf{h} \rangle \leq 0, \quad X \subset Y$
(R4) decomposition rule	$\langle \mathbf{d}_{Y,X}, \mathbf{h} \rangle \leq 0, \quad X \subset Y$

A *proof sequence* of a Shannon-flow inequality  $\langle \boldsymbol{\delta}, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$  is a sequence  $\text{ProofSeq} \stackrel{\text{def}}{=} (w_1 \mathbf{f}_1, w_2 \mathbf{f}_2, \dots, w_\ell \mathbf{f}_\ell)$  of length  $\ell$  satisfying all of the following:

- (1)  $\mathbf{f}_i \in \{\mathbf{s}_{I,J}, \mathbf{m}_{X,Y}, \mathbf{c}_{X,Y}, \mathbf{d}_{Y,X}\}$  for every  $i \in [\ell]$ , called a *proof step*;
- (2)  $w_i \in \mathbb{Q}_+$  for every  $i \in [\ell]$ , called the *weight* of the proof step  $\mathbf{f}_i$ ;
- (3) the vectors  $\boldsymbol{\delta}_0 \stackrel{\text{def}}{=} \boldsymbol{\delta}, \boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_\ell$  defined by  $\boldsymbol{\delta}_i = \boldsymbol{\delta}_{i-1} + w_i \cdot \mathbf{f}_i$  are non-negative rational vectors;
- (4)  $\boldsymbol{\delta}_\ell \geq \boldsymbol{\lambda}$  (element-wise comparison).

A proof sequence  $(w_1 \mathbf{f}_1, w_2 \mathbf{f}_2, \dots, w_\ell \mathbf{f}_\ell)$  implies the following inequalities for all polymatroids  $h \in \Gamma_n$ ,

$$\langle \boldsymbol{\delta}, \mathbf{h} \rangle = \langle \boldsymbol{\delta}_0, \mathbf{h} \rangle \geq \dots \geq \langle \boldsymbol{\delta}_\ell, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle,$$

which provides a step-by-step proof for the Shannon-flow inequality  $\langle \boldsymbol{\delta}, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$ . The following theorem (also Theorem 2 in [35]) is a direct result from Theorem B.12 and Proposition B.13 in [24].

**THEOREM D.1** ([24]). *For any Shannon-flow inequality  $\langle \boldsymbol{\delta}, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$  such that  $\|\boldsymbol{\lambda}\|_1 = 1$ , there is a proof sequence of length  $O(\text{poly}(2^n))$ .*

Theorem D.1 implies that there is a proof sequence for a Shannon-flow inequality that is exponentially long in the number of variables  $n$ , but it is of constant length under fixed query sizes.

## D.2 Joint Shannon-flow Inequalities

To motivate our study of joint Shannon-flow inequalities, we first give a characterization of the optimal objective value of the maximin optimization problem (12),  $\text{OBJ}(S)$ , through the following class of LPs

$$L(\lambda_{BT}, \theta_{BS}, S) \stackrel{\text{def}}{=} \max_{\substack{h_S \in HDC, h_T \in HDC \cap HAC \\ (h_S, h_T) \in (\Gamma_n \times \Gamma_n) \cap HSC}} \sum_{B \in BT} \lambda_B \cdot h_T(B) + \sum_{B \in BS} \theta_B \cdot h_S(B) - (\log S) \cdot \|\theta_{BS}\|_1, \quad (15)$$

parameterized by two vectors,  $\lambda_{BT} \stackrel{\text{def}}{=} (\lambda_B)_{B \in BT} \in \mathbb{Q}_+^{BT}$  with  $\|\lambda_{BT}\|_1 = 1$  and  $\theta_{BS} \stackrel{\text{def}}{=} (\theta_B)_{B \in BS} \in \mathbb{Q}_+^{BS}$ . Equivalently, any

$$(\lambda_{BT}, \theta_{BS}) \in \left\{ (\alpha_T, \alpha_S) \mid \alpha_T \in \mathbb{Q}_+^{BT}, \alpha_S \in \mathbb{Q}_+^{BS}, \|\alpha_T\|_1 = 1 \right\} \quad (16)$$

gives rise to an LP of the form (15) with optimal objective value  $L(\lambda_{BT}, \theta_{BS}, S)$ .

LEMMA D.2. *Let  $S$  be a fixed quantity. For any  $(\lambda_{BT}, \theta_{BS})$  as specified in (16), the optimal objective value of (12),  $\text{OBJ}(S)$  satisfies,*

$$\text{OBJ}(S) \leq L(\lambda_{BT}, \theta_{BS}, S)$$

Moreover, assuming that  $\text{OBJ}(S)$  is positive and bounded, there is an optimal  $(\lambda_{BT}^*, \theta_{BS}^*)$  satisfying (16) such that

$$\text{OBJ}(S) = L(\lambda_{BT}^*, \theta_{BS}^*, S).$$

Instead of proving Lemma D.2 directly, we prove a slightly more general lemma (Lemma D.3) that may be of independent interest.

LEMMA D.3. *Let  $A \in \mathbb{Q}^{\ell \times m}$ ,  $b \in \mathbb{R}^\ell$ ,  $D \in \mathbb{Q}_+^{m \times q}$ ,  $C \in \mathbb{Q}_+^{m \times p}$  be a matrix with columns  $c_1, \dots, c_p$  and polyhedron  $P = \{x \in \mathbb{R}^m \mid Ax \leq b, x \geq 0\}$ . Let  $w^*$  be the optimal objective value of the following optimization problem (assume  $S$  as a fixed quantity)*

$$\begin{aligned} & \max_{x \in P} \min_{k \in [p]} c_k^\top x \\ & \text{s.t. } D^\top x \geq 1_q \log S. \end{aligned} \quad (17)$$

If  $w^*$  is positive and bounded, then for any vectors  $z, u \in \mathbb{Q}_+^p$  with  $\|z\|_1 = 1$ , the following linear program:

$$L(z, u) \stackrel{\text{def}}{=} \max_{x \in P} (Cz)^\top x + (D^\top x - 1_q \log S)^\top u \quad (18)$$

satisfies  $w^* \leq L(z, u)$ . In particular, there is a pair of vectors  $z^* \in \mathbb{Q}_+^p$ ,  $u^* \in \mathbb{Q}_+^q$  with  $\|z^*\|_1 = 1$  such that  $w^* = L(z^*, u^*)$ .

PROOF. First, we introduce  $u \in \mathbb{Q}_+^q$  as the Lagrange multiplier for (17) and obtain

$$\begin{aligned} w^* & \leq \max_{x \in P} \min_{u \in \mathbb{Q}_+^q} \min_{k \in [p]} (c_k^\top x + (D^\top x - 1_q \log S)^\top u) \\ & \leq \max_{x \in P} \min_{u \in \mathbb{Q}_+^q, z \in \mathbb{Q}_+^p, \|z\|_1=1} (Cz)^\top x + (D^\top x - 1_q \log S)^\top u \\ & \leq \min_{u \in \mathbb{Q}_+^q, z \in \mathbb{Q}_+^p, \|z\|_1=1} \max_{x \in P} (Cz)^\top x + (D^\top x - 1_q \log S)^\top u \\ & = \min_{u \in \mathbb{Q}_+^q, z \in \mathbb{Q}_+^p, \|z\|_1=1} L(z, u) \end{aligned}$$

where the third equality is because of the Minimax Inequality. Therefore, we have shown that for any vectors  $u \in \mathbb{Q}_+^q$ ,  $z \in \mathbb{Q}_+^p$  with  $\|z\|_1 = 1$ ,  $L(z, u) \geq w^*$ . Next, we show that there are vectors  $u^* \in \mathbb{Q}_+^q$ ,  $z^* \in \mathbb{Q}_+^p$  with  $\|z^*\|_1 = 1$  such that  $L(z^*, u^*) = w^*$ . We first re-write (12) as the following equivalent linear program:

$$\begin{aligned} & \max_{x, w} w \\ & \text{s.t. } Ax \leq b \\ & \quad D^\top x \geq 1_q \log S \\ & \quad C^\top x \geq 1_p w \\ & \quad x \geq 0, w \geq 0 \end{aligned} \quad (19)$$

Let  $(w^*, x^*)$  be an optimal solution for (19). Then, the dual of (19) can be written as the following linear program:

$$\begin{aligned} & \min_{y, u, z} b^\top y - (1_q \log S)^\top u \\ & \text{s.t. } A^\top y - Du - Cz \geq 0 \\ & \quad 1_p^\top z \geq 1 \\ & \quad y, u, z \geq 0 \end{aligned} \quad (20)$$

Let  $(\mathbf{y}^*, \mathbf{u}^*, \mathbf{z}^*)$  be an extreme point of the (rational) dual polyhedron that attains the optimal objective value for (20), so  $\mathbf{z}^* \in \mathbb{Q}_+^P, \mathbf{u}^* \in \mathbb{Q}_+^q$ . The complementary slackness conditions of the (19) and (20) primal-dual pair and the assumption  $w^* > 0$  imply that  $\mathbf{1}_P^\top \mathbf{z}^* = \|\mathbf{z}^*\|_1 = 1$ ,  $(\mathbf{1}_P w^* - C^\top \mathbf{x}^*)^\top \mathbf{z}^* = 0$  and  $(D^\top \mathbf{x}^* - \mathbf{1}_q \log S)^\top \mathbf{u}^* = 0$ . We then show that  $L(\mathbf{z}^*, \mathbf{u}^*) = w^*$ . First, we note that  $\mathbf{x}^*$  is feasible for (18) with objective value  $(C\mathbf{z}^*)^\top \mathbf{x}^* + (D^\top \mathbf{x}^* - \mathbf{1}_q \log S)^\top \mathbf{u}^* = (C\mathbf{z}^*)^\top \mathbf{x}^* = (C^\top \mathbf{x}^*)^\top \mathbf{z}^* = (\mathbf{1}_P w^*)^\top \mathbf{z}^* = w^*$ . Furthermore, for any feasible  $\mathbf{x}$  to (18), we have that

$$\begin{aligned}
 (C\mathbf{z}^*)^\top \mathbf{x} + (D^\top \mathbf{x} - \mathbf{1}_q \log S)^\top \mathbf{u}^* &= (C\mathbf{z}^*)^\top \mathbf{x} + (\mathbf{D}\mathbf{u}^*)^\top \mathbf{x} - (\mathbf{1}_q \log S)^\top \mathbf{u}^* && \text{re-arrange} \\
 &\leq (\mathbf{A}^\top \mathbf{y}^*)^\top \mathbf{x} - (\mathbf{1}_q \log S)^\top \mathbf{u}^* && \text{dual feasibility and non-negativity} \\
 &= (\mathbf{A}\mathbf{x})^\top \mathbf{y}^* - (\mathbf{1}_q \log S)^\top \mathbf{u}^* \\
 &\leq \mathbf{b}^\top \mathbf{y}^* - (\mathbf{1}_q \log S)^\top \mathbf{u}^* && \text{primal feasibility} \\
 &= w^* && \text{strong duality}
 \end{aligned}$$

This implies that  $L(\mathbf{z}^*, \mathbf{u}^*) \leq w^*$ . Together, we get  $L(\mathbf{z}^*, \mathbf{u}^*) = w^*$ .  $\square$

Now, Lemma D.2 is a direct corollary of Lemma D.3 by setting  $\lambda_{BT}$  as  $\mathbf{z}^*$ ,  $\theta_{BS}$  as  $\mathbf{u}^*$ .

In the remainder of the section, we implicitly assume that  $S$  is a fixed quantity. Moreover, we fix a pair  $(\lambda_{BT}, \theta_{BS})$  satisfying (16). We can now re-write (15), listing out all its constraints and ignore the constant factor  $(\log S) \cdot \|\theta_{BS}\|_1$ :

$$\begin{aligned}
 \ell(\lambda_{BT}, \theta_{BS}) &\stackrel{\text{def}}{=} \max_{\mathbf{h}_S, \mathbf{h}_T} \sum_{B \in BT} \lambda_B \cdot \mathbf{h}_T(B) + \sum_{B \in BS} \theta_B \cdot \mathbf{h}_S(B) \\
 \text{s.t. } &\mathbf{h}_S(Y) - \mathbf{h}_S(X) \leq n_{Y|X}, & (X, Y, N_{Y|X}) \in DC \\
 &\mathbf{h}_T(Y) - \mathbf{h}_T(X) \leq n_{Y|X}, & (X, Y, N_{Y|X}) \in DC \cup AC \\
 &\mathbf{h}_S(I \cup J|J) - \mathbf{h}_S(I|I \cap J) \leq 0, & I \perp J \\
 &\mathbf{h}_T(I \cup J|J) - \mathbf{h}_T(I|I \cap J) \leq 0, & I \perp J \\
 &\mathbf{h}_S(X) - \mathbf{h}_S(Y) \leq 0, & \emptyset \neq X \subset Y \subseteq [n] \\
 &\mathbf{h}_T(X) - \mathbf{h}_T(Y) \leq 0, & \emptyset \neq X \subset Y \subseteq [n] \\
 &\mathbf{h}_S(X) + \mathbf{h}_T(Y|X) \leq n_{Z|\emptyset}, & (X, Y|X, N_{Z|\emptyset}) \in SC \\
 &\mathbf{h}_S(Y|X) + \mathbf{h}_T(X) \leq n_{Z|\emptyset}, & (X, Y|X, N_{Z|\emptyset}) \in SC \\
 &\mathbf{h}_S(Z) \geq 0, \quad \mathbf{h}_T(Z) \geq 0, & \emptyset \neq Z \subseteq [n]
 \end{aligned} \tag{21}$$

Recall that implicitly we have  $\mathbf{h}_S(\emptyset) = \mathbf{h}_T(\emptyset) = 0$  and that  $n_{Y|X} = \log N_{Y|X}$ . Then we write down the dual LP for (21). We associate a dual variable  $(\delta_S)_{Y|X}$  to  $\mathbf{h}_S(Y) - \mathbf{h}_S(X) \leq n_{Y|X}$  for each  $(X, Y, N_{Y|X}) \in DC$  and a dual variable  $(\delta_T)_{Y|X}$  to  $\mathbf{h}_T(Y) - \mathbf{h}_T(X) \leq n_{Y|X}$  for each  $(X, Y, N_{Y|X}) \in DC \cup AC$ . For each  $I \perp J$ , where  $I, J \subseteq [n]$ , we associate a dual variable  $(\sigma_S)_{I,J}$  to the submodularity constraint of  $\mathbf{h}_S$  and  $(\sigma_T)_{I,J}$  to the submodularity constraint of  $\mathbf{h}_T$ . For each  $\emptyset \neq X \subset Y \subseteq [n]$ , we associate a dual variable  $(\mu_S)_{X,Y}$  to the monotonicity constraint of  $\mathbf{h}_S$  and a dual variable  $(\mu_T)_{X,Y}$  to the monotonicity constraint of  $\mathbf{h}_T$ . Lastly, for each  $(X, Y|X, N_{Z|\emptyset}) \in SC$ , we associate a dual variable  $\gamma_{X,Y|X}$  to  $\mathbf{h}_S(X) + \mathbf{h}_T(Y) - \mathbf{h}_T(X) \leq n_{Z|\emptyset}$  and a dual variable  $\gamma_{Y|X,X}$  to  $\mathbf{h}_S(Y) - \mathbf{h}_S(X) + \mathbf{h}_T(X) \leq n_{Z|\emptyset}$ . Moreover, we extend vectors  $(\lambda_B)_{B \in BT}$  and  $(\theta_B)_{B \in BS}$  to every  $Z \in 2^{[n]}$  in the obvious way:

$$\lambda_Z \stackrel{\text{def}}{=} \begin{cases} \lambda_B & \text{when } Z = B \in BT \\ 0 & \text{otherwise} \end{cases} \quad \theta_Z \stackrel{\text{def}}{=} \begin{cases} \theta_B & \text{when } Z = B \in BS \\ 0 & \text{otherwise} \end{cases}$$

Abusing notations, we write  $(X, Y) \in DC$  whenever  $(X, Y, N_{Y|X}) \in DC$  and  $(X, Y|X) \in SC$  whenever  $(X, Y, N_{Z|\emptyset}) \in SC$ . Note that by maintaining the best constraint assumption, we can always recover the only  $N_{Y|X}$  or  $N_{Z|\emptyset}$  from a given  $(Y, X)$ -pair. The dual of (21) can now be written as

$$\begin{aligned}
 \min & \sum_{(X, Y) \in DC} n_{Y|X} \cdot (\delta_S)_{Y|X} + \sum_{(X, Y) \in DC \cup AC} n_{Y|X} \cdot (\delta_T)_{Y|X} \\
 & + \sum_{(X, Y|X) \in SC} n_{Z|\emptyset} \cdot (\gamma_{X,Y|X} + \gamma_{Y|X,X}) \\
 \text{s.t. } & \text{inflow}_S(Z) \geq \theta_Z & \emptyset \neq Z \subseteq [n] \\
 & \text{inflow}_T(Z) \geq \lambda_Z & \emptyset \neq Z \subseteq [n] \\
 & (\delta_S)_{Y|X}, (\mu_S)_{X,Y}, (\sigma_S)_{I,J} \geq 0 \\
 & (\delta_T)_{Y|X}, (\mu_T)_{X,Y}, (\sigma_T)_{I,J} \geq 0 \\
 & \gamma_{X,Y|X}, \gamma_{Y|X,X} \geq 0
 \end{aligned} \tag{22}$$

where for each  $\emptyset \neq Z \subseteq [n]$ ,

$$\begin{aligned} \text{inflow}_S(Z) &\stackrel{\text{def}}{=} \left( \sum_{X:(X,Z) \in \text{DC}} (\delta_S)_Z|X - \sum_{Y:(Z,Y) \in \text{DC}} (\delta_S)_Y|Z \right) + \left( -\sum_{X:X \subset Z} (\mu_S)_{X,Z} + \sum_{Y:Z \subset Y} (\mu_S)_{Z,Y} \right) \\ &\quad + \left( \sum_{\substack{I \perp J \\ I \cap J = Z}} (\sigma_S)_{I,J} + \sum_{\substack{I \perp J \\ I \cup J = Z}} (\sigma_S)_{I,J} - \sum_{J:J \perp Z} (\sigma_S)_{Z,J} \right) \\ &\quad + \left( \sum_{\substack{Z:(Z,Y|Z) \in \text{SC} \\ Z \subset Y}} \gamma_{Z,Y|Z} - \sum_{\substack{Z:(Z,Y|Z) \in \text{SC} \\ Z \subset Y}} \gamma_{Y|Z,Z} + \sum_{\substack{Z:(X,Z|X) \in \text{SC} \\ X \subset Z}} \gamma_{Z|X,X} \right) \\ \text{inflow}_T(Z) &\stackrel{\text{def}}{=} \left( \sum_{X:(X,Z) \in \text{DC} \cup \text{AC}} (\delta_T)_Z|X - \sum_{Y:(Z,Y) \in \text{DC} \cup \text{AC}} (\delta_T)_Y|Z \right) + \left( -\sum_{X:X \subset Z} (\mu_T)_{X,Z} + \sum_{Y:Z \subset Y} (\mu_T)_{Z,Y} \right) \\ &\quad + \left( \sum_{\substack{I \perp J \\ I \cap J = Z}} (\sigma_T)_{I,J} + \sum_{\substack{I \perp J \\ I \cup J = Z}} (\sigma_T)_{I,J} - \sum_{J:J \perp Z} (\sigma_T)_{Z,J} \right) \\ &\quad + \left( \sum_{\substack{Z:(Z,Y|Z) \in \text{SC} \\ Z \subset Y}} \gamma_{Y|Z,Z} - \sum_{\substack{Z:(Z,Y|Z) \in \text{SC} \\ Z \subset Y}} \gamma_{Z,Y|Z} + \sum_{\substack{Z:(X,Z|X) \in \text{SC} \\ X \subset Z}} \gamma_{X,Z|X} \right) \end{aligned}$$

Next, we introduce the *joint Shannon-flow inequalities*.

*Definition D.4 (Joint Shannon-flow Inequality).* The inequality

$$\begin{aligned} \sum_{(X,Y) \in \text{DC}} \textcolor{blue}{h}_S(Y|X) \cdot (\delta_S)_{Y|X} + \sum_{(X,Y) \in \text{DC} \cup \text{AC}} \textcolor{red}{h}_T(Y|X) \cdot (\delta_T)_{Y|X} + \sum_{(X,Y|X) \in \text{SC}} (\textcolor{blue}{h}_S(X) + \textcolor{red}{h}_T(Y|X)) \cdot \gamma_{X,Y|X} \\ + \sum_{(X,Y|X) \in \text{SC}} (\textcolor{blue}{h}_S(Y|X) + \textcolor{red}{h}_T(X)) \cdot \gamma_{Y|X,X} \geq \sum_{B \in \text{BS}} \theta_B \cdot \textcolor{blue}{h}_S(B) + \sum_{B \in \text{BT}} \lambda_B \cdot \textcolor{red}{h}_T(B), \end{aligned} \tag{23}$$

is called a *joint Shannon-flow inequality* if it holds for all  $(\textcolor{blue}{h}_S, \textcolor{red}{h}_T) \in \Gamma_n \times \Gamma_n$  and all coefficients are non-negative rational numbers.

By assigning either polymatroid to be always 0, the above inequality implies the following two Shannon-flow inequalities, called the *participating Shannon-flow inequalities*.

$$\sum_{(X,Y) \in \text{DC}} \textcolor{blue}{h}_S(Y|X) \cdot (\delta_S)_{Y|X} + \sum_{(X,Y|X) \in \text{SC}} \textcolor{blue}{h}_S(X) \cdot \gamma_{X,Y|X} + \sum_{(X,Y|X) \in \text{SC}} \textcolor{blue}{h}_S(Y|X) \cdot \gamma_{Y|X,X} \geq \sum_{B \in \text{BS}} \theta_B \cdot \textcolor{blue}{h}_S(B) \tag{24}$$

$$\sum_{(X,Y) \in \text{DC} \cup \text{AC}} \textcolor{red}{h}_T(Y|X) \cdot (\delta_T)_{Y|X} + \sum_{(X,Y|X) \in \text{SC}} \textcolor{red}{h}_T(Y|X) \cdot \gamma_{X,Y|X} + \sum_{(X,Y|X) \in \text{SC}} \textcolor{red}{h}_T(X) \cdot \gamma_{Y|X,X} \geq \sum_{B \in \text{BT}} \lambda_B \cdot \textcolor{red}{h}_T(B) \tag{25}$$

It will be convenient to write a joint Shannon-flow inequality as inequalities over conditional polymatroid. The polymatroid  $\textcolor{blue}{h}_S$  defines a conditional polymatroid  $\textcolor{blue}{h}_S$  and the polymatroid  $\textcolor{red}{h}_T$  defines a conditional polymatroid  $\textcolor{red}{h}_T$ . More precisely, we define the vectors  $\lambda, \theta \in \mathbb{Q}_+^C$  (extend to  $(X, Y)$  pairs where  $\emptyset \subseteq X \subset Y \subseteq [n]$ ) with coordinate values assigned as the following:

$$\lambda(Y|X) \stackrel{\text{def}}{=} \begin{cases} \lambda_B & \text{when } Y = B, X = \emptyset \\ 0 & \text{otherwise} \end{cases} \quad \theta(Y|X) \stackrel{\text{def}}{=} \begin{cases} \theta_B & \text{when } Y = B, X = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

and similarly, we define the vectors  $\delta_S, \delta_T \in \mathbb{Q}_+^C$  with coordinate values:

$$\delta_S(Y|X) \stackrel{\text{def}}{=} \begin{cases} (\delta_S)_{Y|X} & \text{when } (X, Y) \in \text{DC} \\ 0 & \text{otherwise} \end{cases} \quad \delta_T(Y|X) \stackrel{\text{def}}{=} \begin{cases} (\delta_T)_{Y|X} & \text{when } (X, Y) \in \text{DC} \cup \text{AC} \\ 0 & \text{otherwise} \end{cases}$$

Lastly, the coefficients  $\{\gamma_{X,Y|X} \mid (X, Y|X) \in \text{SC}\} \cup \{\gamma_{Y|X,X} \mid (X, Y|X) \in \text{SC}\}$  contributes to the coefficients of both inequalities (24) and (25). We define a pair of vectors  $\gamma_S, \gamma_T \in \mathbb{Q}_+^C$  as the following:

$$\gamma_S(Y|X) \stackrel{\text{def}}{=} \begin{cases} \gamma_{U,V|U} & \text{when } Y = U, X = \emptyset, (U, V|U) \in \text{SC} \\ \gamma_{V|U,U} & \text{when } Y = V, X = U, (U, V|U) \in \text{SC} \\ 0 & \text{otherwise} \end{cases}$$

$$\gamma_T(Y|X) \stackrel{\text{def}}{=} \begin{cases} \gamma_{U,V|U} & \text{when } Y = V, X = U, (U, V|U) \in \text{SC} \\ \gamma_{V|U,U} & \text{when } Y = U, X = \emptyset, (U, V|U) \in \text{SC} \\ 0 & \text{otherwise} \end{cases}$$

for tracking the contributions to  $\mathbf{h}_S, \mathbf{h}_T$ , respectively. Let  $\mathbf{g}_S \stackrel{\text{def}}{=} \boldsymbol{\delta}_S + \gamma_S$  and  $\mathbf{g}_T \stackrel{\text{def}}{=} \boldsymbol{\delta}_T + \gamma_T$ , then the two inequalities (24) and (25) can be re-written using dot-products:

$$\langle \mathbf{g}_S, \mathbf{h}_S \rangle \stackrel{\text{def}}{=} \langle \boldsymbol{\delta}_S, \mathbf{h}_S \rangle + \langle \gamma_S, \mathbf{h}_S \rangle \geq \langle \boldsymbol{\theta}, \mathbf{h}_S \rangle \quad (26)$$

$$\langle \mathbf{g}_T, \mathbf{h}_T \rangle \stackrel{\text{def}}{=} \langle \boldsymbol{\delta}_T, \mathbf{h}_T \rangle + \langle \gamma_T, \mathbf{h}_T \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h}_T \rangle \quad (27)$$

Moreover, the joint Shannon-flow inequality (23) can be written as

$$\langle \mathbf{g}_S, \mathbf{h}_S \rangle + \langle \mathbf{g}_T, \mathbf{h}_T \rangle \geq \langle \boldsymbol{\theta}, \mathbf{h}_S \rangle + \langle \boldsymbol{\lambda}, \mathbf{h}_T \rangle \quad (28)$$

**THEOREM D.5.** *The inequality (28) is a joint Shannon-flow inequality if and only if there are non-negative vectors of rationals  $\sigma_S, \mu_S, \sigma_T, \mu_T$  such that all constraints of the dual (22) are satisfied. In particular, we call  $(\sigma_S, \mu_S, \sigma_T, \mu_T)$  a witness for the joint Shannon-flow inequality.*

**PROOF.** Proposition 5.6 in [24] states that: (26) is a Shannon-flow inequality if and only if there is a  $(\sigma_S, \mu_S) \geq \mathbf{0}$  (called a witness in [24]) such that  $(\mathbf{g}_S, \sigma_S, \mu_S)$  satisfies the set of constraints:

$$\{\text{inflow}_S(Z) \geq \theta_Z \mid \emptyset \neq Z \subseteq [n]\};$$

similarly, (27) is a Shannon-flow inequality if and only if there is a (witness)  $(\sigma_T, \mu_T) \geq \mathbf{0}$  such that  $(\mathbf{g}_T, \sigma_T, \mu_T)$  satisfies the set of constraints:

$$\{\text{inflow}_T(Z) \geq \lambda_Z \mid \emptyset \neq Z \subseteq [n]\}.$$

Recall the formulation of (22), these two sets of constraints form exactly all the constraints in (22). Thus, (28) is a joint Shannon-flow inequality if and only if there is a  $(\sigma_S, \mu_S, \sigma_T, \mu_T) \geq \mathbf{0}$  such that  $(\mathbf{g}_S, \sigma_S, \mu_S, \mathbf{g}_T, \sigma_T, \mu_T)$  satisfies all the constraints of the dual (22).  $\square$

Theorem D.5 implies that a feasible solution of the dual (22), and in particular the component

$$\{(\delta_S)_{Y|X} \mid (X, Y) \in \text{DC}\} \cup \{(\delta_T)_{Y|X} \mid (X, Y) \in \text{DC}\} \cup \{\gamma_{X,Y|X} \mid (X, Y|X) \in \text{SC}\} \cup \{\gamma_{Y|X,X} \mid (X, Y|X) \in \text{SC}\}$$

in conjunction with  $(\boldsymbol{\lambda}_{\text{BT}}, \boldsymbol{\theta}_{\text{BS}})$ , defines a joint Shannon-flow inequality (23). The component  $(\sigma_S, \mu_S, \sigma_T, \mu_T)$  of the dual, by Theorem D.5, is a witness for the joint Shannon-flow inequality. Note that the joint Shannon-flow inequality implies that

$$\langle \boldsymbol{\theta}, \mathbf{h}_S \rangle + \langle \boldsymbol{\lambda}, \mathbf{h}_T \rangle \leq \sum_{(X, Y) \in \text{DC}} n_{Y|X} \cdot (\delta_S)_{Y|X} + \sum_{(X, Y) \in \text{DC} \cup \text{AC}} n_{Y|X} \cdot (\delta_T)_{Y|X} + \sum_{(X, Y|X) \in \text{SC}} n_{Z|\emptyset} \cdot (\gamma_{X,Y|X} + \gamma_{Y|X,X})$$

where the right-hand side is exactly  $\ell(\boldsymbol{\lambda}_{\text{BT}}, \boldsymbol{\theta}_{\text{BS}})$  by taking the optimal solution of the dual (22) (by strong duality). We have established that: for an arbitrary  $(\boldsymbol{\lambda}_{\text{BT}}, \boldsymbol{\theta}_{\text{BS}})$  satisfying (16), we can construct a joint Shannon-flow inequality,

$$\langle \mathbf{g}_S, \mathbf{h}_S \rangle + \langle \mathbf{g}_T, \mathbf{h}_T \rangle \geq \langle \boldsymbol{\theta}, \mathbf{h}_S \rangle + \langle \boldsymbol{\lambda}, \mathbf{h}_T \rangle,$$

having a witness  $(\sigma_S, \mu_S, \sigma_T, \mu_T)$ , such that its implied upper bound coincides with  $\ell(\boldsymbol{\lambda}_{\text{BT}}, \boldsymbol{\theta}_{\text{BS}})$ .

### D.3 A Brief Review/Augmentation of the PANDA Algorithm

This section provides a brief review of the PANDA algorithm. Let  $\rho$  be a disjunctive rule (4) under degree constraints DC. At a high level, PANDA does the following:

- Step 1. find a vector of non-negative rationals  $\boldsymbol{\lambda}_{\text{BT}} = (\lambda_B)_{B \in \text{BT}}$  with  $\|\boldsymbol{\lambda}_{\text{BT}}\|_1 = 1$ , extend it to a vector (over conditional polymatroid)  $\boldsymbol{\lambda} \in \mathbb{Q}_+^C$  and  $\text{LogSizeBound}_{\Gamma_n \cap \text{HDC}}(\rho) = \max_{h \in \Gamma_n \cap \text{HDC}} \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$ , where the right-hand side is a linear program;
- Step 2. find an optimal dual solution  $(\boldsymbol{\delta}, \boldsymbol{\sigma}, \boldsymbol{\mu})$  to the linear program,  $\max_{h \in \Gamma_n \cap \text{HDC}} \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$ , so that

$$\text{OBJ} \stackrel{\text{def}}{=} \sum_{(X, Y) \in \text{DC}} \log N_{Y|X} \cdot \delta_{Y|X} = \text{LogSizeBound}_{\Gamma_n \cap \text{HDC}}(\rho)$$

and  $\langle \boldsymbol{\delta}, \mathbf{h} \rangle \geq \langle \boldsymbol{\lambda}, \mathbf{h} \rangle$  forms a Shannon-flow inequality;

- Step 3. construct a proof sequence ProofSeq of length  $O(\text{poly}(2^n))$  for the Shannon-flow inequality  $\langle \delta, h \rangle \geq \langle \lambda, h \rangle$  (see Theorem D.1);  
 Step 4. run a PANDA instance, denoted as  $\text{PANDA}(\mathcal{D}, \text{DC}, (\lambda, \delta), \text{ProofSeq})$ , which interprets each proof step of ProofSeq as a relational operation on the input relation, where each operation is guaranteed to take time  $\tilde{O}(2^{\text{OBJ}})$ . Overall, the PANDA instance runs in time  $O(2^{\text{OBJ}})$  and computes a model of size  $\tilde{O}(2^{\text{OBJ}})$ .

In particular, the PANDA instance maintains the following 4 invariants on its inputs, i.e.  $(\mathcal{D}, \text{DC}, (\lambda, \delta), \text{ProofSeq})$ :

**PANDA invariants.**

- (1) *Degree-support invariant:* For every  $\delta_{Y|X} > 0$ , there exist  $Z \subseteq X, W \subseteq Y$  such that  $W - Z = Y - X$  and  $(Z, W, N_{W|Z}) \in \text{DC}$ . The degree constraint is said to support the positive  $\delta_{Y|X}$ . If there are more than one  $(Z, W, N_{W|Z}) \in \text{DC}$  supporting  $\delta_{Y|X}$ , we choose the one with minimum  $N_{W|Z}$  and call it the supporting constraint of  $\delta_{Y|X}$ .
- (2)  $0 < \|\lambda\|_1 \leq 1$
- (3) The Shannon flow inequality along with the supporting degree constraints satisfy  $\sum_{(X,Y)} n(\delta_{Y|X}) \leq \|\lambda\|_1 \cdot \text{OBJ}$  where

$$n(\delta_{Y|X}) \stackrel{\text{def}}{=} \begin{cases} \delta_{Y|X} \cdot n_{W|Z} & \text{if } \delta_{Y|X} > 0 \text{ and} \\ & (Z, W, N_{W|Z}) \text{ supports it} \\ 0 & \text{if } \delta_{Y|X} = 0. \end{cases}$$

and we call the quantity  $\sum_{(X,Y)} n(\delta_{Y|X})$  the *potential*.

- (4) For every  $\delta_{Y|\emptyset} > 0$ , the supporting degree constraint  $(\emptyset, Y, N_{Y|\emptyset})$  satisfies  $n_{Y|\emptyset} \leq \text{OBJ}$ .

**A slight augmentation on PANDA.**

For our purposes, we augment PANDA slightly to take non-optimal proof sequences. More precisely, suppose instead of Step 1 and 2, we are given vectors  $(\lambda_{\text{BT}}, \delta_{\text{DC}})$  and witness  $(\sigma, \mu)$  such that by extending both vectors as  $\lambda, \delta \in \mathbb{Q}_+^C$ ,  $\langle \delta, h \rangle \geq \langle \lambda, h \rangle$  forms a (not necessarily optimal) Shannon-flow inequality with witness  $(\sigma, \mu)$ . Note that the implied upper bound OBJ satisfies

$$\text{LogSizeBound}_{\Gamma_n \cap \text{HDC}}(\rho) \leq \text{OBJ} \stackrel{\text{def}}{=} \sum_{(X,Y) \in \text{DC}} \log N_{Y|X} \cdot \delta_{Y|X},$$

but not necessarily coincides with  $\text{LogSizeBound}_{\Gamma_n \cap \text{HDC}}(\rho)$ . In such cases, following Step 3 and 4, we show that the PANDA instance (taking the non-optimal proof sequence for the given Shannon-flow inequality as input), runs in time as predicated by the non-optimal Shannon-flow inequality, i.e.  $\tilde{O}(2^{\text{OBJ}})$ .

The augmentation is as follows. We observe that in the proof of correctness of PANDA [24], only invariant (4) (at the beginning of the PANDA instance) relies on the optimality of proof sequences. However, as argued in Proposition 6.2, if initially for some  $\delta_{Y|\emptyset} > 0$ ,  $n_{Y|\emptyset} > \text{OBJ}$ , then we could replace the original Shannon-flow inequality with a new Shannon-flow inequality  $\langle \delta', h \rangle \geq \langle \lambda', h \rangle$  along with witness  $(\sigma', \mu')$  such that invariants (1)-(3) hold, and the length of the proof sequence decreases by at least 1. We can repeat this replacement step iteratively until invariant (4) is satisfied. If invariant (4) is never satisfied, then we will end up with a proof sequence of length 0, in which case the Shannon-flow inequality becomes a trivial one,  $\langle \delta_0, h \rangle \geq \langle \lambda_0, h \rangle$ , for some  $\lambda_0, \delta_0$  with  $0 < \|\lambda_0\|_1 \leq 1$  (by invariant (2)) and  $\delta_0 \geq \lambda_0$  (element-wise comparison). This implies that any input relation  $R_F$  where  $(\lambda_0)_{F|\emptyset} > 0$  can be a model and there is an  $n_{F|\emptyset} \leq \text{OBJ}$  that can be appointed as *the* output model (so the model size is  $\tilde{O}(2^{\text{OBJ}})$ ), because otherwise,

$$\sum_{F: (\lambda_0)_{F|\emptyset} > 0} \frac{(\lambda_0)_{F|\emptyset}}{\|\lambda_0\|_1} \cdot n_{F|\emptyset} > \sum_{F: (\lambda_0)_{F|\emptyset} > 0} \frac{(\lambda_0)_{F|\emptyset}}{\|\lambda_0\|_1} \cdot \text{OBJ} = \text{OBJ},$$

and this contradicts invariant (3). For the 2PP algorithm, we implicitly assume that PANDA is equipped with this minor augmentation.

#### D.4 The Algorithm

While 2PP follows a similar structure as the naïve algorithm (and uses the 2-phase algorithmic framework), it is guided by a joint Shannon-flow inequality to execute only the necessary split steps and PANDA instances, which provides more practicality and interpretability. In particular, we will show the following theorem for 2PP.

**THEOREM D.6.** *Let  $\rho$  be a 2-phase disjunctive rule of the form (5) satisfying degree constraints DC (guarded by the input relations) and AC (guarded by the access request). Let  $(\lambda_{\text{BT}}, \theta_{\text{BS}}) \in \{(\mathbf{a}_T, \mathbf{a}_S) \mid \mathbf{a}_T \in \mathbb{Q}_+^{\text{BT}}, \mathbf{a}_S \in \mathbb{Q}_+^{\text{BS}}, \|\mathbf{a}_T\|_1 = 1\}$ . The 2PP algorithm obtains a model of  $\rho$  in two phases and attains the following (smooth) intrinsic trade-off:*

$$S_\rho^{\|\theta_{\text{BS}}\|_1} \cdot T_\rho \cong 2^{\ell(\lambda_{\text{BT}}, \theta_{\text{BS}})} \tag{29}$$

where  $\cong$  hides a poly-logarithmic factor at the right-hand side.

Note that the intrinsic trade-off can be equivalently written as

$$S_\rho^{\|\theta_{\text{BS}}\|_1} \cdot T_\rho \cong \prod_{(X,Y) \in \text{DC}} N_{Y|X}^{(\delta_S)_{Y|X}} \cdot \prod_{(X,Y) \in \text{DC} \cup \text{AC}} N_{Y|X}^{(\delta_S)_{Y|X}} \cdot \prod_{(X,Y|X) \in \text{SC}} N_{Z|\emptyset}^{Y_{X,Y|X} + Y_{Y|X,X}} \tag{30}$$

**Algorithm 1:** 2PP-Preprocess( $\mathcal{D}$ , DC, AC,  $(\tilde{\theta}, \tilde{g}_S)$ , ProofSeq( $S$ ))

---

**Input** : a database instance  $\mathcal{D}$  and degree constraints DC guarded by  $\mathcal{D}$   
**Input** : the degree constraints AC guarded by any possible access request  $Q_A$   
**Input** : the participating Shannon-flow inequality  $\langle \tilde{g}_S, \mathbf{h}_S \rangle \geq \langle \tilde{\theta}, \mathbf{h}_S \rangle$  and its proof sequence ProofSeq( $S$ )  
1 Let SC be the set of split constraints spanned from DC  
2  $SC^+ \leftarrow \{(Y, X) \mid (X, Y|X, N_{Z|\emptyset}) \in SC, \gamma_{X,Y|X} > 0 \vee \gamma_{Y|X,X} > 0\}$   
3 Apply a sequence of split steps, one for every  $(Y, X) \in SC^+$  and spawn  $k$  subproblems with inputs  $(\mathcal{D}^{(j)}, DC^{(j)}, j \in [k])$   
 $// k = O(\text{poly}(\log |\mathcal{D}|))$   
4  $\mathcal{J} \leftarrow \emptyset$   
5 **forall**  $j \in [k]$  **do**  
6   Create a PANDA instance PANDA( $\mathcal{D}^{(j)}, DC^{(j)}, (\tilde{\theta}, \tilde{g}_S)$ , ProofSeq( $S$ ))  
7   **if** the potential satisfies (32) **then**  
8      $(S_B^{(j)})_{B \in BS} \leftarrow \text{PANDA}(\mathcal{D}^{(j)}, DC^{(j)}, (\tilde{\theta}, \tilde{g}_S), \text{ProofSeq}(S))$   
9   **else**  
10     **abort** PANDA( $\mathcal{D}^{(j)}, DC^{(j)}, (\tilde{\theta}, \tilde{g}_S)$ , ProofSeq( $S$ ))  
11     **insert**  $(\mathcal{D}^{(j)}, DC^{(j)})$  to  $\mathcal{J}$   
12 **return**  $(\mathcal{J}, (\cup_j S_B^{(j)})_{B \in BS})$

---

In particular, given a fixed  $S$  for  $S_P$ , we can construct from Lemma D.2 (using complementary slackness)  $(\lambda_{BT}^*, \theta_{BS}^*)$  such that  $\text{OBJ}(S) = L(\lambda_{BT}^*, \theta_{BS}^*, S)$ . Since  $\ell(\lambda_{BT}^*, \theta_{BS}^*) - (\log S) \cdot \|\theta_{BS}^*\|_1 = L(\lambda_{BT}^*, \theta_{BS}^*, S)$ , Theorem D.6 recovers the best possible intrinsic trade-off as specified in (12). In the rest of this section, we present the 2PP algorithm, and defer the full proof of Theorem D.6 to the next section.

**Preparation phase.** Similar to PANDA, 2PP has a preparation phase to construct the necessary inputs for a 2PP instance. First, we construct from the given  $(\lambda_{BT}, \theta_{BS})$  a joint Shannon-flow inequality,  $\langle g_S, \mathbf{h}_S \rangle + \langle g_T, \mathbf{h}_T \rangle \geq \langle \theta, \mathbf{h}_S \rangle + \langle \lambda, \mathbf{h}_T \rangle$  and a witness for it,  $(\sigma_S, \mu_S, \sigma_T, \mu_T)$ . Recall that the joint Shannon-flow inequality implies an upper bound coincides with  $\ell(\lambda_{BT}, \theta_{BS})$ .

Second, we construct a proof sequence for the joint Shannon-flow inequality. The idea is to build two parallel proof sequences for its two participating Shannon-flow inequalities and stitch them together. Recall that the two participating Shannon-flow inequalities are

$$\begin{aligned} \langle g_S, \mathbf{h}_S \rangle &\geq \langle \theta, \mathbf{h}_S \rangle \\ \langle g_T, \mathbf{h}_T \rangle &\geq \langle \lambda, \mathbf{h}_T \rangle \end{aligned}$$

From the proof of Theorem D.5,  $(\sigma_S, \mu_S)$  is a witness for  $\langle g_S, \mathbf{h}_S \rangle \geq \langle \theta, \mathbf{h}_S \rangle$  and  $(\sigma_T, \mu_T)$  is a witness for  $\langle g_T, \mathbf{h}_T \rangle \geq \langle \lambda, \mathbf{h}_T \rangle$ . We normalize the Shannon-flow inequality  $\langle g_S, \mathbf{h}_S \rangle \geq \langle \theta, \mathbf{h}_S \rangle$  into  $\langle \tilde{g}_S, \mathbf{h}_S \rangle \geq \langle \tilde{\theta}, \mathbf{h}_S \rangle$ , where  $\tilde{g}_S \stackrel{\text{def}}{=} g_S / \|\theta\|_1$  and  $\tilde{\theta} \stackrel{\text{def}}{=} \theta / \|\theta\|_1$ , with a new witness  $(\tilde{\sigma}_S, \tilde{\mu}_S) \stackrel{\text{def}}{=} (\sigma_S / \|\theta\|_1, \mu_S / \|\theta\|_1)$ . From Theorem D.1, we can construct a proof sequence ProofSeq( $S$ ) for the Shannon-flow inequality  $\langle \tilde{g}_S, \mathbf{h}_S \rangle \geq \langle \tilde{\theta}, \mathbf{h}_S \rangle$  and a proof sequence ProofSeq( $T$ ) for  $\langle g_T, \mathbf{h}_T \rangle \geq \langle \lambda, \mathbf{h}_T \rangle$ , where both proof sequences have length  $O(\text{poly}(2^n))$ ,  $n$  being the number of variables. We say that ProofSeq( $S$ ) and ProofSeq( $T$ ) are the *participating proof sequences* for the joint Shannon-flow inequality. As a brief summary, in the preparation phase, 2PP:

- (1) (see Theorem D.5) constructs a joint Shannon-flow inequality  $\langle g_S, \mathbf{h}_S \rangle + \langle g_T, \mathbf{h}_T \rangle \geq \langle \theta, \mathbf{h}_S \rangle + \langle \lambda, \mathbf{h}_T \rangle$  with a witness  $(\sigma_S, \mu_S, \sigma_T, \mu_T)$ ;
- and
- (2) (see Theorem D.1) constructs a ProofSeq( $S$ ) for the participating Shannon-flow inequality  $\langle \tilde{g}_S, \mathbf{h}_S \rangle \geq \langle \tilde{\theta}, \mathbf{h}_S \rangle$  and a ProofSeq( $T$ ) for the participating Shannon-flow inequality  $\langle g_T, \mathbf{h}_T \rangle \geq \langle \lambda, \mathbf{h}_T \rangle$

Now, we walk through both phases of 2PP. In particular, we denote 2PP-Preprocess as the preprocessing phase of 2PP and 2PP-Online as the online phase of 2PP. The sketches of 2PP-Preprocess and 2PP-Online are in the box of Algorithm 1 and Algorithm 2.

**The preprocessing phase.** We call this phase 2PP-Preprocess and it is sketched in the box of Algorithm 1. We first scan over the joint Shannon-flow inequality,  $\langle g_S, \mathbf{h}_S \rangle + \langle g_T, \mathbf{h}_T \rangle \geq \langle \theta, \mathbf{h}_S \rangle + \langle \lambda, \mathbf{h}_T \rangle$  and  $(\sigma_S, \mu_S, \sigma_T, \mu_T)$  and apply a sequence of split steps that consists of one  $(Y, X)$ -pair for every  $(Y, X)$  satisfying  $(X, Y|X, N_{Z|\emptyset}) \in SC$  and either  $\gamma_{X,Y|X} > 0$  or  $\gamma_{Y|X,X} > 0$ . The sequence of split steps spawns  $\text{poly}(\log |\mathcal{D}|)$  subproblems. Let  $\rho(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC)$  be the  $j$ -th subproblem after the sequence of split steps, having degree constraints  $DC^{(j)}$  guarded by  $\mathcal{D}^{(j)}$ . Next, we create for it a PANDA instance PANDA( $\mathcal{D}^{(j)}, DC^{(j)}, (\tilde{\theta}, \tilde{g}_S)$ , ProofSeq( $S$ )) and look at its initial potential

$$\sum_{(X, Y)} n_{W|Z}^{(j)} \cdot (\tilde{g}_S)_{Y|X} \stackrel{\text{def}}{=} \sum_{(X, Y) \in DC} n_{W|Z}^{(j)} \cdot \frac{(\delta_S)_{Y|X}}{\|\theta\|_1} + \sum_{(X, Y|X) \in SC} n_{X|\emptyset}^{(j)} \cdot \frac{\gamma_{X,Y|X}}{\|\theta\|_1} + \sum_{(X, Y|X) \in SC} n_{W|Z}^{(j)} \cdot \frac{\gamma_{Y|X,X}}{\|\theta\|_1} \quad (31)$$

**Algorithm 2:** 2PP-Online( $\mathcal{J}, \{Q_A\}, AC, (\lambda, g_T)$ , ProofSeq( $T$ ))

---

**Input** : an index  $\mathcal{J}$  containing  $O(\text{poly}(\log |\mathcal{D}|))$  entries where each entry contains input relations  $\mathcal{D}^{(j)}$  and degree constraints  $DC^{(j)}$  guarded by  $\mathcal{D}^{(j)}$

**Input** : an access request  $Q_A$  and the degree constraints  $AC$  guarded by  $Q_A$

**Input** : the participating Shannon-flow inequality  $\langle g_T, h_T \rangle \geq \langle \lambda, h_T \rangle$  and its proof sequence  $\text{ProofSeq}(T)$

**1 forall**  $(\mathcal{D}^{(j)}, DC^{(j)}) \in \mathcal{J}$  **do**

  2   Create a PANDA instance  $\text{PANDA}(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC, (\lambda, g_T), \text{ProofSeq}(T))$

  3    $(T_B^{(j)})_{B \in BT} \leftarrow \text{PANDA}(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC, (\lambda, g_T), \text{ProofSeq}(T))$

**4 return**  $(\bigcup_j T_B^{(j)})_{B \in BT}$

---

where  $n_{W|Z}^{(j)} \stackrel{\text{def}}{=} \log N_{W|Z}^{(j)}$  and  $(Z, W, N_{W|Z}^{(j)}) \in DC^{(j)}$  is the constraint that supports a positive  $(\tilde{g}_S)_{Y|X}$ . If the potential satisfies

$$\sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (\tilde{g}_S)_{Y|X} \leq \log S, \quad (32)$$

then 2PP allows this PANDA instance to run and stores its output  $(S_B^{(j)})_{B \in BS}$ . Otherwise, 2PP aborts this PANDA instance and keeps track of the input  $(\mathcal{D}^{(j)}, DC^{(j)})$  for the  $j$ -th subproblem using an index  $\mathcal{J}$ . The data structure(s) stored in the preprocessing phase are the index  $\mathcal{J}$  that tracks inputs (and its degree constraints) from aborted instances, and a set of tables  $S_B = \bigcup_j S_B^{(j)}, B \in BS$  from all succeeded PANDA instances.

**The online phase.** We call this phase 2PP-Online and it is sketched in the box of Algorithm 2. The algorithm scans over the index  $\mathcal{J}$  built in the preprocessing phase, and for each  $(\mathcal{D}^{(j)}, DC^{(j)}) \in \mathcal{J}$ , it creates and runs a PANDA instance  $\text{PANDA}(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC, (\lambda, g_T), \text{ProofSeq}(T))$ . At the termination of each PANDA instance, 2PP collects outputs  $(T_B^{(j)})_{B \in BT}$ . The overall output in the online phase is the set of tables  $T_B = \bigcup_j T_B^{(j)}, B \in BT$  from all PANDA instances created and executed online.

## D.5 Analysis of the 2PP algorithm

In the rest of the section, we formally prove Theorem D.6 for the 2PP algorithm.

**PROOF OF THEOREM D.6.** Recall that the split steps at the initial stage of 2PP spawn  $O(\text{poly}(\log |\mathcal{D}|))$  subproblems. We prove by analyzing the space and time usage for the  $j$ -th subproblem.

First, in the preprocessing phase, if the potential of the PANDA instance,  $\text{PANDA}(\mathcal{D}^{(j)}, DC^{(j)}, (\tilde{\theta}, \tilde{g}_S), \text{ProofSeq}(S))$ , is no larger than  $\log S$ , then by invariant (4) of PANDA, the output tables  $(S_B^{(j)})_{B \in BT}$  (and every intermediate view produced by this PANDA instance) have size  $\tilde{O}(S)$ . Otherwise, the  $j$ -th subproblem consumes space  $O(|\mathcal{D}|)$  for storing its input  $(\mathcal{D}^{(j)}, DC^{(j)})$  in the index. Thus, the overall space consumption for the data structure is  $O(\text{poly}(\log |\mathcal{D}|)) \cdot \tilde{O}(|\mathcal{D}| + S) = \tilde{O}(S)$ .

Next, we are left to show that for any subproblem delegated to the online phase, the PANDA instance created by 2PP,  $\text{PANDA}(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC, (\lambda, g_T), \text{ProofSeq}(T))$ , terminates in time  $\tilde{O}(T_\rho)$ , where

$$\log T_\rho = \ell(\lambda_{BT}, \theta_{BS}) - \log S \cdot \|\theta_{BS}\|_1.$$

Recall that  $n_{Y|X} \stackrel{\text{def}}{=} \log N_{Y|X}$  and by strong duality,

$$\ell(\lambda_{BT}, \theta_{BS}) = \sum_{(X,Y) \in DC} n_{Y|X} \cdot (\delta_S)_{Y|X} + \sum_{(X,Y) \in DC \cup AC} n_{Y|X} \cdot (\delta_S)_{Y|X} + \sum_{(X,Y|X) \in SC} n_{Z|\emptyset} \cdot (\gamma_{X,Y|X} + \gamma_{Y|X,X}).$$

To show this, we look at both PANDA instances of the  $j$ -th subproblem,  $\rho(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC)$ , i.e.

$$\begin{aligned} &(\text{preprocessing instance}) && \text{PANDA}(\mathcal{D}^{(j)}, DC^{(j)}, (\tilde{\theta}, \tilde{g}_S), \text{ProofSeq}(S)) \\ &(\text{online instance}) && \text{PANDA}(\mathcal{D}^{(j)} \cup \{Q_A\}, DC^{(j)} \cup AC, (\lambda, g_T), \text{ProofSeq}(T)) \end{aligned}$$

The *preprocessing instance* has the potential specified in (31), while the *online instance* has the following *potential*:

$$\sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (g_T)_{Y|X} \stackrel{\text{def}}{=} \sum_{(X,Y) \in DC \cup AC} n_{W|Z}^{(j)} \cdot (\delta_T)_{Y|X} + \sum_{(X,Y|X) \in SC} n_{W|Z}^{(j)} \cdot \gamma_{X,Y|X} + \sum_{(X,Y|X) \in SC} n_{X|\emptyset}^{(j)} \cdot \gamma_{Y|X,X}$$

where  $n_{W|Z}^{(j)} \stackrel{\text{def}}{=} \log N_{W|Z}^{(j)}$  and  $(Z, W, N_{W|Z}^{(j)}) \in DC^{(j)}$  is the constraint that supports a positive  $(g_T)_{Y|X}$ . Now we have,

$$\sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (g_T)_{Y|X} + \|\theta\|_1 \cdot \sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (\tilde{g}_S)_{Y|X} = \sum_{(X,Y) \in DC} n_{W|Z}^{(j)} \cdot (\delta_S)_{Y|X} + \sum_{(X,Y) \in DC \cup AC} n_{W|Z}^{(j)} \cdot (\delta_T)_{Y|X}$$

$$+ \sum_{(X,Y|X) \in \text{SC}} (n_{X|\emptyset}^{(j)} + n_{W|Z}^{(j)}) \cdot \gamma_{X,Y|X} + \sum_{(X,Y|X) \in \text{SC}} (n_{W|Z}^{(j)} + n_{X|\emptyset}^{(j)}) \cdot \gamma_{Y|X,X}$$

Recall that 2PP executes a split step for every  $(Y, X)$ -pair satisfying  $(X, Y|X, N_{Z|\emptyset}) \in \text{SC}$  and  $\gamma_{X,Y|X} > 0 \vee \gamma_{Y|X,X} > 0$ . So for every such  $(X, Y)$ -pair, there are some  $(\emptyset, X, N_{X|\emptyset}^{(j)}), (X, Y, N_{Y|X}^{(j)}) \in \text{DC}^{(j)}$  such that  $N_{X|\emptyset}^{(j)} \cdot N_{Y|X}^{(j)} \leq N_{Z|\emptyset}$ . It implies that  $n_{X|\emptyset}^{(j)} + n_{W|Z}^{(j)} \leq \log N_{Z|\emptyset}$ . Moreover, for every  $(X, Y) \in \text{DC}$  where  $(\delta_S)_{Y|X} > 0$  or  $(\delta_T)_{Y|X} > 0$ , it holds that  $n_{W|Z}^{(j)} \leq \log N_{Y|X}$ . Therefore, we get

$$\sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (g_T)_{Y|X} + \|\theta\|_1 \cdot \sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (\tilde{g}_S)_{Y|X} \leq \ell(\lambda_{\text{BT}}, \theta_{\text{BS}})$$

This implies that for the  $j$ -th subproblem (thus for any subproblem), either in the preprocessing phase,

$$\sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (g_T)_{Y|X} \leq \log S,$$

or in the online phase

$$\begin{aligned} \sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (g_T)_{Y|X} &\leq \ell(\lambda_{\text{BT}}, \theta_{\text{BS}}) - \|\theta\|_1 \cdot \sum_{(X,Y)} n_{W|Z}^{(j)} \cdot (\tilde{g}_S)_{Y|X} \\ &\leq \ell(\lambda_{\text{BT}}, \theta_{\text{BS}}) - \|\theta_{\text{BS}}\|_1 \cdot \log S \\ &= \log T_P. \end{aligned}$$

So, all online PANDA instances terminate in time as predicated by (30). □

## E MISSING DETAILS FROM SECTION 6

### E.1 Tradeoffs via Fractional Edge Cover (Section 6.2)

The following lemma is a generalization of Shearer's lemma (Lemma D.1 in [29]).

LEMMA E.1. *Let  $\mathcal{H} = ([n], \mathcal{E})$  be a hypergraph and  $\hat{\mathbf{u}}$  be a fractional edge cover of  $[n] \setminus A \subseteq [n]$ . Then,*

$$\sum_{F \in \mathcal{E}} \hat{u}_F \cdot h(F \mid A \cap F) + h(A) \geq h([n])$$

PROOF. Assume w.l.o.g. that  $A = \{1, \dots, \ell - 1\}$ . Then we can write:

$$\begin{aligned} h([n]) &= \sum_{j=\ell}^n h(j \mid i : i < j) + h(A) \\ &\leq \sum_{j=\ell}^n \sum_{F \in \mathcal{E}; j \in F} \hat{u}_F \cdot h(j \mid i : i < j, i \in F) + h(A) \\ &= \sum_{F \in \mathcal{E}} \hat{u}_F \sum_{j \in F \setminus A} h(j \mid i : i < j, i \in F) + h(A) \\ &= \sum_{F \in \mathcal{E}} \hat{u}_F \left( h(F) - \sum_{j \in F, j < \ell} h(j \mid i : i < j, i \in F) \right) + h(A) \\ &= \sum_{F \in \mathcal{E}} \hat{u}_F \left( h(F) - \sum_{j \in F \cap A} h(j \mid i : i < j, i \in F \cap A) \right) + h(A) \\ &= \sum_{F \in \mathcal{E}} \hat{u}_F (h(F) - h(A \cap F)) + h(A) \\ &= \sum_{F \in \mathcal{E}} \hat{u}_F \cdot h(F \mid A \cap F) + h(A) \end{aligned}$$

This completes the proof.  $\square$

PROOF OF THEOREM 6.1. To obtain the desired tradeoff, we consider two PMTDs. The first PMTD  $P_1$  has one node  $t$  with  $\chi_1(t) = [n]$  and  $M_1 = \emptyset$ , while the second PMTD  $P_2$  has also one bag  $t$  with  $\chi_2(t) = [n]$  and  $M_2 = \{t\}$ .  $P_1$  contains the  $T$ -view  $T_{[n]}(\mathbf{x}_{[n]})$ , while  $P_2$  contains the  $S$ -view  $S_A(\mathbf{x}_A)$ . These two PMTDs correspond to the following materialization policy: either store directly the answer of an access request, or compute the access request from scratch. Hence, we only need to consider one disjunctive rule (we use  $\mathbf{x}$  to denote the tuple  $\mathbf{x}_{12\dots n}$ ):

$$T_{[n]}(\mathbf{x}) \vee S_A(\mathbf{x}_A) \leftarrow Q_A(\mathbf{x}_A) \wedge \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F).$$

Define  $\alpha = \alpha(\mathbf{u}, A)$  and  $\hat{\mathbf{u}} = \mathbf{u}/\alpha$ . We can now write the following proof:

$$\begin{aligned} \sum_{F \in \mathcal{E}} u_F \cdot \log N_{F \mid \emptyset} + \alpha \cdot \log |Q_A| &\geq \sum_{F \in \mathcal{E}} u_F \cdot \{\textcolor{red}{h}_T(F \mid A \cap F) + \textcolor{blue}{h}_S(A \cap F)\} + \alpha \cdot \textcolor{red}{h}_T(A) \\ &= \sum_{F \in \mathcal{E}} u_F \cdot \textcolor{red}{h}_T(F \mid A \cap F) + \alpha \cdot \textcolor{red}{h}_T(A) + \sum_{F \in \mathcal{E}} u_F \cdot \textcolor{blue}{h}_S(A \cap F) \\ &\geq \sum_{F \in \mathcal{E}} u_F \cdot \textcolor{red}{h}_T(F \mid A \cap F) + \alpha \cdot \textcolor{red}{h}_T(A) + \textcolor{blue}{h}_S(A) && (\textit{Shearer's Lemma}) \\ &= \alpha \sum_{F \in \mathcal{E}} \hat{u}_F \cdot \textcolor{red}{h}_T(F \mid A \cap F) + \alpha \cdot \textcolor{red}{h}_T(A) + \textcolor{blue}{h}_S(A) && (\textit{Lemma E.1}) \\ &\geq \alpha \cdot \textcolor{red}{h}_T([n]) + \textcolor{blue}{h}_S(A) \end{aligned}$$

The second inequality is a direct application of Shearer's Lemma on the sub-hypergraph  $(A, \{A \cap F \mid F \in \mathcal{E}\})$  of  $H$ , since  $\mathbf{u}$  is a fractional edge cover of  $A$ . The last inequality is a direct consequence of Lemma E.1. By Theorem D.6, we obtain the desired tradeoff.  $\square$

## E.2 Tradeoffs via Tree Decompositions (Section 6.3)

Let  $\varphi(\mathbf{x}_A \mid \mathbf{x}_A)$  be a CQAP. Following Section 6.3, let  $\mathcal{P}$  be the set of all 2-phase disjunctive rules generated by the induced set of PMTDs. We start our analysis by showing that for any disjunctive rule  $\rho_a$  in this set, there is another disjunctive rule  $\rho_b$  that is no easier than  $\rho_a$  in terms of its intrinsic tradeoff. Interestingly, despite choosing a (possibly) harder rule, we are still able to recover many state-of-the-art tradeoffs. We begin by stating two key observations.

**Observation E.1.** *For any 2-phase disjunctive rules  $\rho_a$  and  $\rho_b$ ,  $\rho_a$  is said to be no harder than  $\rho_b$  (or equivalently,  $\rho_b$  is no easier than  $\rho_a$ ) if the S-targets of  $\rho_b$  are a subset of the S-targets of  $\rho_a$  and the T-targets of  $\rho_b$  are a subset of the T-targets of  $\rho_a$ .*

In other words, Observation E.1 states that adding more targets to the head of a disjunctive rule can only make its model evaluation easier since we can always ignore the new targets. The next lemma makes use of the structure of the T-views in a PMTD.

**LEMMA E.2.** *Let  $\varphi(\mathbf{x}_A \mid \mathbf{x}_A)$  be a given CQAP. Let  $(\mathcal{T}, \chi, r)$  be a fixed free-connex tree decomposition. Let  $\mathcal{P}$  be the set of PMTDs induced from  $(\mathcal{T}, \chi, r)$ . Let  $\rho_a$  be a 2-phase disjunctive rule generated from  $\mathcal{P}$ . Then, there is a 2-phase disjunctive rule  $\rho_b$  that is no easier than  $\rho_a$  such that for any two T-targets of  $\rho_b$ , their corresponding nodes in  $\mathcal{T}$  do not lie in some root-to-leaf path.*

**PROOF.** Let  $\text{BT}(\rho)$  be a set such that for every  $B \in \text{BT}(\rho)$ , there is a T-targets  $T_B$  picked by the 2-phase disjunctive rule  $\rho$  (similarly, we define  $\text{BS}(\rho)$  for S-views). Note that by definition of views, a T-view for a node implies that its ancestors are all associated with T-views.

We construct the 2-phase disjunctive rule  $\rho_b$  in the following way: for each PMTD, if  $\rho_a$  picks a S-view,  $\rho_b$  follows  $\rho_a$ 's pick; if  $\rho_a$  picks a T-view,  $\rho_b$  looks at the path from root to this T-view (picked by  $\rho_a$ ) and pick the first node  $t$  such that  $v(t) \in \text{BT}(\rho_a)$ , i.e. the top-most T-view  $\rho_a$  picked on this branch. It is easy to see that by this construction, it holds that no two bags corresponding to two different T-targets of  $\rho_b$  lie on a root-to-leaf path in  $\mathcal{T}$ . Furthermore,  $\text{BT}(\rho_b) \subseteq \text{BT}(\rho_a)$  and  $\text{BS}(\rho_b) = \text{BS}(\rho_a)$ . Thus, rule  $\rho_b$  is no easier than rule  $\rho_a$ .  $\square$

We are now ready to show the main property for any tree decomposition.

**LEMMA E.3.** *Let  $\varphi(\mathbf{x}_A \mid \mathbf{x}_A)$  be a given CQAP. Let  $(\mathcal{T}, \chi, r)$  be a fixed free-connex tree decomposition. Let  $\mathcal{P}$  be the set of PMTDs induced from  $(\mathcal{T}, \chi, r)$ . Any 2-phase disjunctive rule  $\rho_a$  generated from  $\mathcal{P}$  is no easier than a 2-phase disjunctive rule of the form*

$$T(\mathbf{x}_{\chi(t_\ell)}) \vee \bigvee_{j \in [\ell]} S(\mathbf{x}_{A_j}) \leftarrow Q_A(\mathbf{x}_A) \wedge \bigwedge_{F \in \mathcal{E}} R_F(\mathbf{x}_F), \quad (33)$$

where  $t_1 = r, t_2, \dots, t_\ell$  are the nodes of the tree  $\mathcal{T}$  that form a path starting from the root node  $r$ ; and

$$A_j \stackrel{\text{def}}{=} \begin{cases} A & \text{if } j = 1 \\ \chi(t_j) \cap \chi(t_{j-1}) & \text{if } j = 2, \dots, \ell \end{cases}$$

**PROOF.** We start by invoking Lemma E.2 with  $\rho_a$  and  $(\mathcal{T}, \chi, r)$  as input to get the rule  $\rho_b$  that satisfies that no T-target of  $\rho_b$  is an ancestor of another. We fix the T-targets of  $\rho_b$ , i.e. fix  $\text{BT}(\rho_b) = \{B_1, \dots, B_k\}$ . Our goal is to show that  $\rho_b$  must contain a subset of S-targets whose corresponding nodes in  $\mathcal{T}$  form a path starting from the root and ending at some node  $t$  such that  $\chi(t) \in \text{BT}(\rho_b)$ . In the following, we will use the function  $\chi^{-1}(B)$  to recover the node  $t \in V(\mathcal{T})$  such that  $\chi(t) = B$  and use  $\text{parent}(t)$  to denote the parent node of a non-root node  $t$ .

We prove that the property holds by allowing an adversary to pick targets in PMTDs, while we adaptively choose the PMTDs that the adversary must pick from. We can choose the ordering of the PMTDs because to construct a 2-phase disjunctive rule, one view must be picked from every PMTD in  $\mathcal{P}$  and we are only controlling the order in which they are examined.

Consider the PMTD  $P_1$  where the set for S-targets is exactly  $M_1 = \{\chi^{-1}(B_1), \dots, \chi^{-1}(B_k)\}$ . We offer the adversary to pick a target from  $P_1$ . We claim that the adversary must pick one S-view corresponding to a node in  $M_1$ . Indeed, the adversary cannot pick a T-view from  $P_1$  since the T-targets of  $\rho_b$  have already been fixed and cannot be changed. Suppose the adversary picks S-view associated with node  $\chi^{-1}(B_i)$ . For this S-view (since its parent is associated with a T-view),

$$v(\chi^{-1}(B_i)) = \chi(\chi^{-1}(B_i)) \cap \chi(\text{parent}(\chi^{-1}(B_i))) = B_i \cap \chi(\text{parent}(\chi^{-1}(B_i))). \quad (34)$$

We will now choose the PMTD  $P_2$  where  $M_2 = M_1 \cup \text{parent}(\chi^{-1}(B_i))$  and give it to the adversary. Once again, the adversary cannot pick a T-view since that will change  $\text{BT}(\rho_b)$  and must choose an S-view associated with one of the nodes in  $M_2$ . Suppose the adversary picks S-view associated with a node  $\chi^{-1}(B_{i'}) \in M_2$ . We generate the next PMTD  $P_3$  where  $M_3 = M_2 \cup \text{parent}(\chi^{-1}(B_{i'}))$ . In general, the PMTD  $P_{q+1}$  generated after  $q$  rounds has  $M_{q+1} = M_q \cup \text{parent}(\chi^{-1}(B_q))$ , where  $\chi^{-1}(B_q) \in M_q$  is the node  $\rho_b$  picked as a S-view in the last round (for PMTD  $P_q$  with materialization set  $M_q$ ). This process terminates when the S-view corresponding to the root is picked by the adversary. It is also guaranteed that this process will terminate after a finite number of steps as we always add nodes in the materialization set by moving up the tree and the tree is of finite size. At every step, the adversary picks an S-view for a bag and can only get closer to the root across all branches. Therefore, when the adversary reaches the root, there must be a subset of picked S-views corresponding to some path starting from the root  $t_1 = r$  to some node  $t_\ell$  in  $\mathcal{T}$  such that  $\chi(t_\ell) \in \text{BT}(\rho_b)$ , the desired property. Also, note that these S-views follow (34), thus we complete the proof.  $\square$

Lemma E.3 tells us that it is sufficient to analyze only the 2-phase disjunctive rules that are of the form as shown in (33). Now we proceed to find a proof sequence for (33). Let the bags for  $t_1 = r, t_2, \dots, t_\ell$  have corresponding fractional edge covers  $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(\ell)}$ , where  $\mathbf{u}^{(j)} = (u_F^{(j)})_{F \in \mathcal{E}}$ , for  $j \in [\ell]$ . We define  $\alpha_j$  to be the slack of each bag w.r.t. the variables in  $A_j$  (i.e.  $\alpha_j = \alpha(\mathbf{u}^{(j)}, A_j)$ ), and introduce a factor  $\beta_j = \beta^*/\alpha_j$  where  $\beta^* = \sum_{j \in [\ell]} \alpha_j$ . Next, we apply Theorem 6.1 for each of the  $\ell$  bags, multiply the proof sequence obtained for the  $j$ -th bag with  $\beta_j$ , and sum up terms as follows:

$$\begin{aligned}
\sum_{j \in [\ell]} \beta_j \sum_{F \in \mathcal{E}} u_F^{(j)} \cdot \log |R_F| + \beta^* \cdot \log |Q_A| &\geq \sum_{j \in [\ell]} \beta_j \sum_{F \in \mathcal{E}} u_F^{(j)} \cdot (\textcolor{red}{h_T}(F \mid A_j \cap F) + \textcolor{blue}{h_S}(A_j \cap F)) + \beta^* \cdot \textcolor{red}{h_T}(A_1) \\
&= \sum_{j \in [\ell]} \beta_j \sum_{F \in \mathcal{E}} u_F^{(j)} \cdot \textcolor{red}{h_T}(F \mid A_j \cap F) + \beta^* \cdot \textcolor{red}{h_T}(A_1) + \sum_{j \in [\ell]} \beta_j \sum_{F \in \mathcal{E}} u_F \cdot \textcolor{blue}{h_S}(A_j \cap F) \\
&\geq \sum_{j \in [\ell]} \beta_j \cdot \alpha_j \sum_{F \in \mathcal{E}} \hat{u}_F^{(j)} \cdot \textcolor{red}{h_T}(F \mid A_j \cap F) + \beta^* \cdot \textcolor{red}{h_T}(A_1) + \sum_{j \in [\ell]} \beta_j \cdot \textcolor{blue}{h_S}(A_j) \\
&= \sum_{j \in [\ell]} \beta^* \sum_{F \in \mathcal{E}} \hat{u}_F^{(j)} \cdot \textcolor{red}{h_T}(F \mid A_j \cap F) + \beta^* \cdot \textcolor{red}{h_T}(A_1) + \sum_{j \in [\ell]} \beta_j \cdot \textcolor{blue}{h_S}(A_j) \\
&= \sum_{j \in \{2, \dots, \ell\}} \beta^* \sum_{F \in \mathcal{E}} \hat{u}_F^{(j)} \cdot \textcolor{red}{h_T}(F \mid A_j \cap F) + \beta^* \left( \sum_{F \in \mathcal{E}} \hat{u}_F^{(1)} \cdot \textcolor{red}{h_T}(F \mid A_1 \cap F) + \textcolor{red}{h_T}(A_1) \right) + \sum_{j \in [\ell]} \beta_j \cdot \textcolor{blue}{h_S}(A_j) \\
&\geq \sum_{j \in \{2, \dots, \ell\}} \beta^* \sum_{F \in \mathcal{E}} \hat{u}_F^{(j)} \cdot \textcolor{red}{h_T}(F \mid A_j \cap F) + \beta^* \textcolor{red}{h_T}(\chi(t_1)) + \sum_{j \in [\ell]} \beta_j \cdot \textcolor{blue}{h_S}(A_j) \\
&\geq \sum_{j \in \{2, \dots, \ell\}} \beta^* \sum_{F \in \mathcal{E}} \hat{u}_F^{(j)} \cdot \textcolor{red}{h_T}(F \mid A_j \cap F) + \beta^* \textcolor{red}{h_T}(A_2) + \sum_{j \in [\ell]} \beta_j \cdot \textcolor{blue}{h_S}(A_j) \\
&\dots \\
&\geq \beta^* \textcolor{red}{h_T}(\chi(t_\ell)) + \sum_{j \in [\ell]} \beta_j \cdot \textcolor{blue}{h_S}(A_j),
\end{aligned}$$

where the second inequality is a direct application of Shearer's Lemma and the third inequality is a direct consequence of Lemma E.1. From this proof sequence, we obtain the following intrinsic tradeoff,

$$|Q_A|^{\beta^*} \cdot |\mathcal{D}|^{\sum_{j \in [\ell]} \beta_j \cdot u_j^*} \cong S^{\sum_{j \in [\ell]} \beta_j} \cdot T^{\beta^*}$$

where  $u_j^* = \sum_{F \in \mathcal{E}} u_F^{(j)}$ . Equivalently, we get

$$|Q_A| \cdot |\mathcal{D}|^{\sum_{j \in [\ell]} u_j^*/\alpha_j} \cong S^{\sum_{j \in [\ell]} 1/\alpha_j} \cdot T. \quad (35)$$

In the above tradeoff, for a given  $S$ , (assume  $|Q_A| = 1$ ), we get

$$\begin{aligned}
\log T &= \sum_{j \in [\ell]} \frac{u_j^*}{\alpha_j} \cdot \log |\mathcal{D}| - \sum_{j \in [\ell]} (1/\alpha_j) \log S \\
&= \sum_{j \in [\ell]} (1/\alpha_j) \cdot (u_j^* \log |\mathcal{D}| - \log S)
\end{aligned}$$

One observation is that if some bag  $t_j$  on the path  $t_1, \dots, t_\ell$  has an AGM bound that is not greater than  $S$ , then the materialization of  $t_j$ 's corresponding  $S$ -view  $S(\mathbf{x}_{A_j})$  can be fully materialized as the model for (33). Otherwise, for every  $t_j, j \in [\ell]$ , we have that  $(u_j^* \log |\mathcal{D}| - \log S)$  is non-negative for every  $j \in [\ell]$ , thus the above expression for  $\log T$  monotonically increases as  $\ell$  increases. Therefore, the most expensive tradeoff corresponds to the disjunctive rule of the form in (33) that starts from the root and ends at a leaf. To obtain the final space-time tradeoffs, we simply take the worst tradeoffs across all the root-to-leaf paths in  $\mathcal{T}$ .

Before we conclude this section, we show that the tradeoff we obtained in (35) across all root-to-leaf paths of a fixed tree decomposition  $(\mathcal{T}, \chi)$  recovers (and possibly improves over) Theorem 13 of [13], without incurring extra hyper-parameters. Indeed, in (35), the authors set a hyper-parameter  $\delta(t)$  for every  $t \in V(\mathcal{T})$  and let the online answering time to be  $T = |\mathcal{D}|^{\sum_{j \in [\ell]} \delta(t_j)}$ , where  $t_1, \dots, t_\ell$  is a root-to-leaf path of  $\mathcal{T}$  that maximizes  $\sum_{j \in [\ell]} \delta(t_j)$ . Suppose we construct a 2-phase disjunctive rule of the form (35) for this root-to-leaf path,  $t_1, \dots, t_\ell$ , then for any fractional edge cover  $u_j^*, j \in [\ell]$ , it holds that (by rewriting (35))

$$\log_{|\mathcal{D}|} S = \frac{1}{\sum_{j \in [\ell]} (1/\alpha_j)} \cdot \left( \sum_{j \in [\ell]} \frac{u_j^*}{\alpha_j} - \log_{|\mathcal{D}|} T \right)$$

$$\begin{aligned}
&= \frac{1}{\sum_{j \in [\ell]} (1/\alpha_j)} \cdot \left( \sum_{j \in [\ell]} \frac{u_j^*}{\alpha_j} - \sum_{j \in [\ell]} \delta(t_j) \right) \\
&= \frac{1}{\sum_{j \in [\ell]} (1/\alpha_j)} \cdot \sum_{j \in [\ell]} \frac{1}{\alpha_j} (u_j^* - \delta(t_j) \cdot \alpha_j) \\
&\leq \max_{j \in [\ell]} \left( \sum_{F \in \mathcal{E}} u_F^{(j)} - \delta(t_j) \cdot \alpha_j \right).
\end{aligned}$$

The last inequality holds because for all  $w_i \geq 0$  such that  $\sum_i w_i = 1$ ,  $\sum_{i=1}^n \gamma_i w_i \leq \max_i \gamma_i$ . Setting  $w_i = \frac{1}{\alpha_i} \cdot \frac{1}{\sum_{j \in [\ell]} (1/\alpha_j)}$ , we get the desired expression. Thus, it is easy to see that the tradeoff we obtained in (35) across all root-to-leaf paths of a fixed tree decomposition indeed recovers results in [13].

### E.3 Additional Examples

In this section, we present a number of concrete examples of space-time tradeoffs obtained through our framework. Across all examples, we assume the database size to be  $|\mathcal{D}|$ . Moreover, for brevity, we use an ordered tuple of views ( $T$ -views and  $S$ -views) to represent a PMTD that takes a path-like structure (the first entry of the tuple denotes the root). For ease of interpreting tradeoffs from proof sequences, we always carry the implied upper bound of its corresponding joint Shannon-flow inequalities at the left-hand side of the proof sequences. For the implied upper bound, we denote  $n_F \stackrel{\text{def}}{=} \log |R_F|$  as the log-size of the relation  $R_F$  and  $w_A \stackrel{\text{def}}{=} \log |Q_A|$  as the log-size of the access request  $Q_A$ .

*Example E.4 (The triangle query).* We take a triangle query with an *empty* access pattern, i.e.  $A = \emptyset$ :

$$\varphi(x_1, x_3 \mid \emptyset) \leftarrow R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_1).$$

We consider two PMTDs, both with the bag  $\{x_1, x_2, x_3\}$ . In the first PMTD, the bag is not materialized and hence a  $T$ -view  $T_{123}$  associated with it. In the second PMTD, the bag is materialized and has an  $S$ -view  $S_{13}$ . Thus, the two PMTDs can be denoted as

$$(T_{123}), \quad (S_{13}).$$

Hence, we obtain the following disjunctive rule (without the atom  $Q_A$ ):

$$T_{123} \vee S_{13} \leftarrow R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_1).$$

One (empty) proof sequence for it is simply  $\log |\mathcal{D}| \geq h_S(13)$ , indicating that we can store all pairs of  $(x_1, x_3)$  that participate in at least one triangle in linear space.

*Example E.5 (The square query).* For this example, we take the following CQAP:

$$\varphi(x_1, x_3 \mid x_1, x_3) \leftarrow R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4) \wedge R(x_4, x_1).$$

This captures the following task: given two vertices of a graph, decide whether they occur in two opposite corners of a square. We consider two PMTDs. The first PMTD has a root bag  $\{1, 3, 4\}$  associated with a  $T$ -view  $T_{134}$ , and a bag  $\{1, 3, 2\}$  associated with a  $T$ -view  $T_{132}$ . The second PMTD has one bag  $\{1, 2, 3, 4\}$  associated with an  $S$ -view  $S_{13}$ . The two PMTDs can be denoted as

$$(T_{134}, T_{132}), \quad (S_{13})$$

This in turn generates two disjunctive rules:

$$\begin{aligned}
T_{134} \vee S_{13} &\leftarrow Q_{13}(x_1, x_3) \wedge R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4) \wedge R(x_4, x_1) \\
T_{132} \vee S_{13} &\leftarrow Q_{13}(x_1, x_3) \wedge R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4) \wedge R(x_4, x_1).
\end{aligned}$$

We can construct the following proof sequence for the first rule:

$$\begin{aligned}
n_{14} + n_{34} + 2 \cdot w_{13} &\geq h_S(1) + h_T(4|1) + h_S(3) + h_T(4|3) + 2 \cdot h_T(13) \\
&\geq h_S(13) + h_T(4|1) + h_T(4|3) + 2 \cdot h_T(13) \\
&\geq h_S(13) + h_T(4|13) + h_T(13) + h_T(4|13) + h_T(13) \\
&= h_S(13) + 2 \cdot h_T(134)
\end{aligned}$$

For the second rule, we symmetrically construct a proof for  $n_{12} + n_{32} + 2 \cdot w_{13} \geq h_S(13) + 2 \cdot h_T(132)$ . Hence, we obtain a tradeoff of  $S \cdot T^2 \cong |\mathcal{D}|^2 \cdot |Q_A|^2$ . This tradeoff (when  $|Q_A| = 1$ ) recovers the improved one obtained in Example 15 of [12].

#### E.4 Tradeoffs for $k$ -Reachability (Section 6.4)

*Example E.6 (2-reachability).* Consider the 2-reachability CQ with the following access pattern (optimizing for  $|Q_{13}| = 1$ ):

$$\phi_2(x_1, x_3 \mid x_1, x_3) \leftarrow R_1(x_1, x_2) \wedge R_2(x_2, x_3).$$

From the tree decomposition that has one bag  $\{1, 2, 3\}$ , we construct the set of all non-redundant and non-dominant PMTDs, i.e.

$$(T_{123}), \quad (S_{13}).$$

They generate only one 2-phase disjunctive rule,

$$T_{123} \vee S_{13} \leftarrow Q_{13}(x_1, x_3), R_1(x_1, x_2), R_2(x_2, x_3),$$

for which we can construct the following proof sequence:

$$\begin{aligned} n_{12} + n_{23} + 2 \cdot w_{13} &\geq h_S(1) + h_T(2|1) + h_S(3) + h_T(2|3) + 2 \cdot h_T(13) \\ &\geq h_S(13) + h_T(2|1) + h_T(2|3) + 2 \cdot h_T(13) \\ &\geq h_S(13) + h_T(2|13) + h_T(2|13) + 2 \cdot h_T(13) \\ &\geq h_S(13) + 2 \cdot h_T(123) \end{aligned} \quad (S \cdot T^2 \cong |\mathcal{D}|^2 \cdot |Q_{13}|^2)$$

**Discussion.** Suppose  $w_{13} = \log |\mathcal{D}|$ ,  $(1, 13, N_{13|1}) \in \text{AC}$  and  $w_{3|1} \stackrel{\text{def}}{=} \log N_{13|1}$ , then  $n_{12} + w_{3|1} \geq h_T(12) + h_T(3|12) = h_T(123)$  is a (possibly desirable) proof sequence, which implies that we can answer any access request in time  $T = |\mathcal{D}| \cdot N_{13|1}$  without any materializations. Note that if  $S = |\mathcal{D}|$ ,  $N_{13|1} = |\mathcal{D}|^{1/2-\epsilon}$  for some  $\epsilon > 0$ , it is strictly better than the above proof sequence (which implies  $T = |\mathcal{D}|^{3/2}$ ).

*Example E.7 (3-reachability).* We study the 3-reachability CQAP (optimizing for  $|Q_A| = 1$ ), i.e.

$$\phi_3(x_1, x_4 \mid x_1, x_4) \leftarrow R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4).$$

Using the set of all non-redundant and non-dominant PMTDs (shown in Figure 3), we generate all 2-phase disjunctive rules (after discarding redundant rules/targets), and their corresponding proof sequences:

$$(1) \rho_1 : T_{134} \vee T_{124} \vee S_{14} \leftarrow Q_{14}(x_1, x_4) \wedge R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4)$$

$$\begin{aligned} n_{12} + n_{34} + 2w_{14} &\geq h_S(1) + h_S(4) + h_T(2|1) + h_T(3|4) + 2h_T(14) \\ &\geq h_S(14) + h_T(2|1) + h_T(3|4) + 2h_T(14) \\ &\geq h_S(14) + h_T(2|14) + h_T(3|14) + 2h_T(14) \\ &\geq h_S(14) + h_T(124) + h_T(134) \end{aligned} \quad (S \cdot T^2 \cong |\mathcal{D}|^2 \cdot |Q_A|^2)$$

$$(2) \rho_2 : T_{123} \vee S_{13} \vee T_{124} \vee S_{14} \leftarrow Q_{14}(x_1, x_4) \wedge R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4)$$

$$\begin{aligned} 2 \cdot n_{12} + n_{23} + n_{34} + 3 \cdot w_{14} &= 2(h_S(1) + h_T(2|1)) + h_S(3) + h_T(2|3) + h_S(4) + h_T(3|4) + 3 \cdot h_T(14) \\ &\geq h_S(14) + h_S(13) + 2h_T(2|14) + h_T(3|14) + h_T(2|314) + 3 \cdot h_T(14) \\ &= h_S(14) + h_S(13) + 2h_T(124) + h_T(1234) \\ &\geq h_S(14) + h_S(13) + 3 \cdot h_T(124) \end{aligned} \quad (S^2 \cdot T^3 \cong |\mathcal{D}|^4 \cdot |Q_A|^3)$$

$$(3) \rho_3 : T_{134} \vee T_{234} \vee S_{24} \vee S_{14} \leftarrow Q_{14}(x_1, x_4) \wedge R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4)$$

The proof sequence for  $\rho_3$  is omitted here because it is symmetric to rule  $\rho_2$ .

$$(4) \rho_4 : T_{123} \vee S_{13} \vee T_{234} \vee S_{24} \vee S_{14} \leftarrow Q_{14}(x_1, x_4) \wedge R(x_1, x_2) \wedge R(x_2, x_3) \wedge R(x_3, x_4)$$

For  $\rho_4$ , we show 2 proof sequences that do not dominate each other as follows:

$$\begin{aligned} n_{12} + n_{34} + w_{14} &\geq h_S(1) + h_S(4) + h_T(2|1) + h_T(3|4) + h_T(14) \\ &\geq h_S(14) + h_T(2|14) + h_T(3|214) + h_T(14) \\ &\geq h_S(14) + h_T(1234) \\ &\geq h_S(14) + h_T(123) \end{aligned} \quad (S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|)$$

$$\begin{aligned} 2n_{23} + 2n_{12} + 2n_{34} + w_{14} &\geq 2 \cdot h_S(23) + h_S(12) + h_S(34) + h_S(1) + h_T(2|1) + h_S(4) + h_T(3|4) + h_T(14) \\ &= h_S(2) + h_S(3|2) + h_S(3) + h_S(2|3) + h_S(12) + h_S(34) + h_S(1) + h_T(2|1) + h_S(4) + h_T(3|4) + h_T(14) \\ &\geq h_S(123) + h_S(234) + h_S(2) + h_S(3) + h_S(1) + h_T(2|1) + h_S(4) + h_T(3|4) + h_T(14) \\ &\geq h_S(123) + h_S(234) + h_S(24) + h_S(13) + h_T(2|1) + h_T(3|4) + h_T(14) \\ &\geq 2 \cdot h_S(24) + 2 \cdot h_S(13) + h_T(2|14) + h_T(3|124) + h_T(14) \\ &\geq 2 \cdot h_S(24) + 2 \cdot h_S(13) + h_T(1234) \end{aligned}$$

$$\geq 2 \cdot h_S(24) + 2 \cdot h_S(13) + h_T(123) \quad (S^4 \cdot T \cong |\mathcal{D}|^6 \cdot |Q_A|)$$

Lastly, for every rule above, there is a proof sequence that corresponds to applying breath-first search (BFS) from scratch in time  $O(|\mathcal{D}|)$  at the online phase. Take  $\rho_1$  as an example, we get

$$\begin{aligned} n_{23} + w_{14} &\geq h_T(23) + h_T(14) \\ &\geq h_T(1234) \\ &\geq h_T(134) \end{aligned} \quad (T = |\mathcal{D}| \cdot |Q_A|)$$

The corresponding plot of tradeoff curve is included in Figure 4a.

*Example E.8 (4-reachability).* We study the following CQAP for 4-reachability (optimizing for  $|Q_A| = 1$ ):

$$\phi_4(x_1, x_5 \mid x_1, x_5) \leftarrow R_{12}(x_1, x_2) \wedge R_{23}(x_2, x_3) \wedge R_{34}(x_3, x_5) \wedge R_{45}(x_4, x_5)$$

We fix the following set of non-redundant and non-dominant PMTDs (11 in total), where we use an ordered tuple of views to represents a PMTD that has a path-like structure (the first entry of the tuple denotes the root). Note that including more PMTDs could potentially obtain even improved tradeoffs.

$$\begin{aligned} (T_{1235}, T_{345}), \quad (T_{1235}, S_{35}), \quad (T_{1345}, T_{123}), \quad (T_{1345}, S_{13}), \quad (T_{1245}, T_{234}), \quad (T_{1245}, S_{24}) \\ (T_{125}, T_{2345}), \quad (T_{125}, S_{25}), \quad (T_{145}, T_{1234}), \quad (T_{145}, S_{14}), \quad (S_{15}) \end{aligned}$$

Though there are  $2^{10}$  2-phase disjunctive rules generated from the above set of PMTDs, similar to 3-reachability, we can discard rules with strictly more targets than other rules. To start off, note that any rule has to have  $S_{15}$  as a target. For any disjunctive rules picking  $T_{1245}$  as a target (or  $T_{125}, T_{145}$ ), i.e.  $\rho_1 : T_{2345} \vee S_{15}$ , we always have the following proof sequence:

$$\begin{aligned} n_{12} + n_{45} + w_{15} &\geq h_S(1) + h_T(2|1) + h_S(5) + h_T(4|5) \\ &\geq h_S(15) + h_T(2|15) + h_T(4|125) + h_T(15) \\ &= h_S(15) + h_T(1245) \end{aligned} \quad (S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|)$$

Otherwise, for the last 7 PMTDs, the disjunctive rule must pick  $\{T_{234}, S_{24}, T_{2345}, S_{25}, T_{1234}, S_{14}, S_{15}\}$ , or just  $\{T_{234}, S_{24}, S_{25}, S_{14}, S_{15}\}$ , by removing the redundant term  $T_{2345}$  due to the presence of  $T_{234}$ . Now the disjunctive rules picks targets out of the first 4 PMTDs, which we can break into the following cases, discarding redundant targets:

$$(1) \rho_2 : T_{1235} \vee T_{1345} \vee (T_{234} \vee S_{24} \vee S_{25} \vee S_{14} \vee S_{15})$$

$$\begin{aligned} n_{12} + n_{23} + n_{34} + n_{45} + 2w_{15} &\geq h_S(1) + h_T(2|1) + h_S(2) + h_T(3|2) + h_S(4) + h_T(3|4) + h_S(5) + h_T(4|5) + 2h_T(15) \\ &\geq h_S(15) + h_S(24) + h_T(2|15) + h_T(3|125) + h_T(3|145) + h_T(4|15) + 2h_T(15) \\ &= h_S(15) + h_S(24) + h_T(1235) + h_T(1345) \end{aligned} \quad (S^2 \cdot T^2 \cong |\mathcal{D}|^4 \cdot |Q_A|^2)$$

$$(2) \rho_3 : T_{345} \vee S_{35} \vee T_{123} \vee S_{13} \vee (T_{234} \vee S_{24} \vee S_{25} \vee S_{14} \vee S_{15})$$

The proof sequence for  $\rho_3$  is omitted here because  $\rho_3$  is no harder than  $\rho_2$  since  $\{3, 4, 5\} \subseteq \{1, 3, 4, 5\}$  and  $\{1, 2, 3\} \subseteq \{1, 2, 3, 5\}$

$$(3) \rho_4 : T_{345} \vee S_{35} \vee T_{1345} \vee (T_{234} \vee S_{24} \vee S_{25} \vee S_{14} \vee S_{15}) = T_{345} \vee S_{35} \vee (T_{234} \vee S_{24} \vee S_{25} \vee S_{14} \vee S_{15})$$

For  $\rho_4$ , we show 2 proof sequences that do not dominate each other as follows:

$$\begin{aligned} 2n_{23} + 2n_{12} + 5n_{34} + 3n_{45} + 5w_{15} \\ \geq 2(h_S(2) + h_T(3|2)) + 2(h_S(1) + h_T(2|1)) + 2(h_S(3) + h_T(4|3)) + \\ 3(h_S(4) + h_T(3|4)) + 3(h_S(5) + h_T(4|5)) + 5h_T(15) \\ = 2(h_T(2|1) + h_T(3|2) + h_T(4|3)) + 3(h_T(4|5) + h_T(3|4)) + 5h_T(15) + \\ 2(h_S(3) + h_S(5)) + (h_S(2) + h_S(5)) + (h_S(2) + h_S(4)) + 2(h_S(1) + h_S(4)) \\ \geq 2(h_T(2|15) + h_T(3|125) + h_T(4|1235)) + 3(h_T(4|5) + h_T(3|45)) + 2h_T(15) + 3h_T(5) + \\ 2(h_S(3) + h_S(5|3)) + (h_S(2) + h_S(5|2)) + (h_S(2) + h_S(4|2)) + 2(h_S(1) + h_S(4|1)) \\ = 2h_T(12345) + 3h_T(345) + 2h_S(35) + h_S(25) + h_S(24) + 2h_S(14) \\ \geq 5h_T(345) + 2h_S(35) + h_S(25) + h_S(24) + 2h_S(14) \end{aligned} \quad (S^6 \cdot T^5 \cong |\mathcal{D}|^{12} \cdot |Q_A|^5)$$

$$\begin{aligned} 3n_{23} + 3n_{34} + 3n_{45} + n_{12} + 2n_{34} + n_{23} + 3w_{15} \\ \geq 3(h_S(3) + h_S(2|3)) + 3h_S(34) + 3(h_S(5) + h_T(4|5)) + (h_S(1) + h_T(2|1)) + 2(h_S(4) + h_T(3|4)) + (h_S(2) + h_T(3|2)) + 3h_T(15) \\ \geq 3h_S(234) + 3h_S(35) + 3h_T(4|5) + (h_S(1) + h_T(2|1)) + 2(h_S(4) + h_T(3|4)) + (h_S(2) + h_T(3|2)) + 3h_T(15) \\ \geq 3h_S(24) + 3h_S(35) + h_S(14) + h_S(24) + 3h_T(4|5) + h_T(2|1) + 2h_T(3|4) + h_T(3|2) + 3h_T(15) \\ \geq 3h_S(24) + 3h_S(35) + h_S(14) + h_S(24) + 3h_T(4|5) + h_T(2|1) + 2h_T(3|45) + h_T(3|2) + 2h_T(5) + h_T(15) \end{aligned}$$

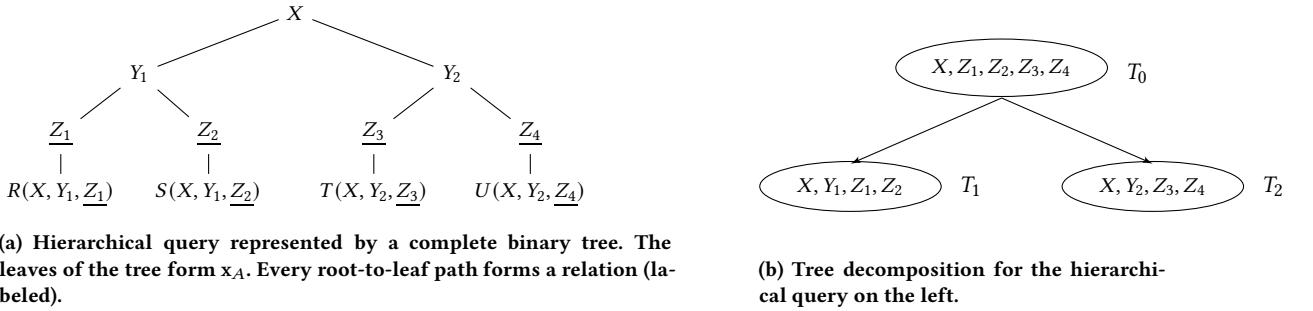


Figure 6: Example hierarchical CQAP and its tree decomposition.

$$\begin{aligned}
 &\geq 3hs(24) + 3hs(35) + hs(14) + hs(24) + 2ht(345) + ht(4|15) + ht(2|145) + ht(3|1245) + ht(15) \\
 &\geq 3hs(24) + 3hs(35) + hs(14) + hs(24) + 2ht(345) + ht(12345) \\
 &\geq 3hs(24) + 3hs(35) + hs(14) + hs(24) + 3ht(345) \\
 (4) \quad &\rho_5 : T_{1235} \vee T_{123} \vee S_{13} \vee (T_{234} \vee S_{24} \vee S_{25} \vee S_{14} \vee S_{15})
 \end{aligned}$$

The proof sequence for  $\rho_5$  is omitted here because it is symmetric to rule  $\rho_4$ .

Similar to 3-reachability, there is a proof sequence for every rule above that corresponds to breath-first search in the online phase. The corresponding plot of tradeoff curve is included in Figure 4b.

## F SPACE-TIME TRADEOFFS FOR BOOLEAN HIERARCHICAL QUERIES

In this section, we show some applications of our framework to CQ Boolean hierarchical queries, which are CQAPs defined over queries whose body is hierarchical. A query said to be *hierarchical* if for any two of its variables, either their sets of atoms are disjoint or one is contained in the other. An alternative interpretation is that each hierarchical query admits a *canonical ordering*<sup>3</sup>, which is a rooted tree where the variables of each atom in the query lie along the same root-to-leaf path in the tree and each atom is a child of its lowest variable. For example, the query shown in Figure 6a is hierarchical. Variable  $Y_1$  is present in atoms  $\{R, S\}$ , which is a subset of the atoms in which variable  $X$  is present, i.e.  $\{R, S, T, U\}$ . Note that every root-to-leaf path forms a relation.

Hierarchical queries are an interesting class of queries that captures tractability for a variety of problems such as query evaluation in probabilistic databases [10], dynamic query evaluation [5, 18], etc. Seminal work by [20] provided tradeoffs between preprocessing time and delay guarantees for enumerating the result of any (not necessarily full) hierarchical query (i.e.  $A = \emptyset$  for their setting). However, [20] does not deal with the setting where access patterns can be specified on certain variables of the hierarchical queries, which is the main focus of our work. Despite the difference in the settings, an adaptation of the main algorithm from [20] is able to provide a non-trivial baseline for tradeoffs between space usage and answering time that we highlight in this section. The adapted algorithm may be of independent interest as well. We begin by describing the main algorithm for the static enumeration of hierarchical queries.

Given a hierarchical query  $\varphi(y)$ , the preprocessing phase in [20] takes the free variables<sup>4</sup>  $y$  and constructs a list of skew-aware view trees and heavy/light indicator views. Starting from the root, for any bound variable that violates the free-connex property, two evaluation strategies are used. The first strategy materializes a subset of the query result obtained for the light values over the set of variables  $\text{anc}(W) \cup \{W\}$ <sup>5</sup> in the variable order. It also aggregates away the bound variables in the subtree rooted at  $W$ . Since the light values have a bounded degree, this materialization is inexpensive. The second strategy computes a compact representation of the rest of the query result obtained for those values over  $\text{anc}(W) \cup \{W\}$  that are heavy (i.e., have high degree) in at least one relation. This second strategy treats  $W$  as a free variable and proceeds recursively to resolve further bound variables located below  $W$ .

*Example F.1.* For the query in Figure 6a, the following views are constructed:  $V_X(Z_1, Z_2, Z_3, Z_4)$ ,  $V_{Y_1}(X, Z_1, Z_2)$ ,  $V_{Y_2}(X, Z_3, Z_4)$  for light degree threshold  $N^\epsilon$ ,  $0 \leq \epsilon \leq 1$  of  $\{X\}$ ,  $\{X, Y_1\}$ ,  $\{X, Y_2\}$  respectively (assume relation size  $N$ ). The indicator views are  $H_X(X)$  and  $L_X(X)$  that store which  $X$  values are heavy and light. Proceeding recursively, we treat  $A$  as a free variable and construct the indicator views  $H_{Y_1}(X, Y_1)$  and  $L_{Y_1}(X, Y_1)$  (and similarly for the right subtree). The heavy indicator views have size at most  $O(N^{1-\epsilon})$ .

The enumeration phase then uses the views as an input to the so called UNION and PRODUCT algorithm that combines the output of  $O(1)$  number of view trees and uses the observation that the heavy bound variables themselves form a hierarchical structure.

*Example F.2.* Continuing the example, in the enumeration phase, the output from  $V_X(Z_1, Z_2, Z_3, Z_4)$  is available with constant delay. For each heavy  $X$  in  $H_X(X)$ , we proceed recursively and either use the view  $V_{Y_1}(X, Z_1, Z_2)$ , whose output can be accessed with constant

<sup>3</sup>The original definition in [20] also has a dependency function but we omit that since we do not use it for the decomposition construction.

<sup>4</sup>The terminology in [20] calls variables in the head as free and all other variables of the query as bound (not to be confused with  $x_A$ , which we refer to as bound in this paper).

<sup>5</sup> $\text{anc}(W)$  is defined as the variables on the path from  $W$  to the root excluding  $W$ .

delay for all light  $\{X, Y_1\}$  or use the observation that the only remaining case for the left subtree is when  $\{X, Y_1\}$  is heavy, allowing us to enumerate the  $(Z_1, Z_2)$  answers from the subqueries  $Q_{Z_1}(X, Y_1, Z_1) = R(X, Y_1, Z_1)$  and  $Q_{Z_2}(X, Y_1, Z_2) = S(X, Y_1, Z_2)$ , both of which allow constant delay enumeration for a given fixing of  $\{X, Y_1\}$ . The overall delay is  $O(N^{1-\epsilon})$ .

**THEOREM F.3** ([20]). *Given a hierarchical query with static width  $w$ , a database of size  $N$ , and  $\epsilon \in [0, 1]$ , the query result can be enumerated with  $O(N^{1-\epsilon})$  delay after  $O(N^{1+(w-1)\epsilon})$  preprocessing time and space.*

The static width  $w$  is a generalization of fractional hypertree width that takes the structure of the free variables in the query into account. For the query in Figure 6a,  $w = 4$ .

**Adapted Algorithm.** In order to adapt the algorithm for enumerating hierarchical query results to apply to our setting, we make two key observations. First, we apply the preprocessing phase of [20] with the free variables as  $\mathbf{x}_A$  and construct the data structures. Then, we can construct an indexed representation of the base relations, and the view trees where the indexing variables are the set  $\mathbf{x}_A$ . In other words, for any view  $V$  and each  $v \in \Pi_{\text{vars}(V) \cap A}(V)$ , we store  $V \times v$  in a hashtable  $H_V$  with key as  $v$ . Second, given a fixing (say  $z$ ) for  $\mathbf{x}_A$ , we provide the indexed base relations, view indicators, and the views (i.e.  $H_V[\Pi_{\text{vars}(V) \cap A}(z)]$ ) to the enumeration algorithm that enumerates the query result with  $O(N^{1-\epsilon})$  delay. Since each base relation  $R' = R \times z$  provided as input has only one possible value for  $\text{vars}(R) \cap A$  (which is  $\Pi_{\text{vars}(R) \cap A}(z)$ ), the only answer the enumeration algorithm can output is either  $z$  or declare that there is no answer. This allows us to answer the Boolean hierarchical CQAP successfully. Let  $w$  denote the static width of the hierarchical query obtained from [20] with free variables as  $\mathbf{x}_A$ . We can obtain the following tradeoff.

**THEOREM F.4.** *Given a Boolean hierarchical CQAP (i.e.  $H = A$ ), a database of size  $N$ ,  $w$  as static width with free variables as  $\mathbf{x}_A$ , and  $\epsilon \in [0, 1]$ , the query can be answered in time  $O(N^{1-\epsilon})$  using a preprocessed data structure that takes space  $S = O(N^{1+(w-1)\epsilon})$ .*

*Example F.5.* Applying Theorem F.4 to the CQAP (we use  $Z$  as a shorthand for the set  $\{Z_1, Z_2, Z_3, Z_4\}$ )

$$\varphi(Z | Z) \leftarrow Q_A(Z) \wedge R(X, Y_1, Z_1) \wedge S(X, Y_1, Z_2) \wedge T(X, Y_2, Z_3) \wedge U(X, Y_2, Z_4)$$

that corresponds to Figure 6a gives us the tradeoff  $S \cdot T^3 \cong |\mathcal{D}|^4$  for any instantiation of the query. Here,  $Z$  is the access pattern.

Next, we show how the tradeoff in Example F.5 can be recovered by our framework and improved to also takes the access request  $Q_A$  into account. Figure 6b shows the (free-connex) tree decomposition we consider for the query. The set of PMTDs (5 of them) induced from this (free-connex) tree decomposition contains 5 PMTDs, by choosing the materialization set  $M$  to be one of the following:  $\{T_0\}$ ,  $\{T_1\}$ ,  $\{T_2\}$ ,  $\{T_1, T_2\}$  or  $\emptyset$ . Let

$$\text{body} = Q_A(Z) \wedge R(X, Y_1, Z_1) \wedge S(X, Y_1, Z_2) \wedge T(X, Y_2, Z_3) \wedge U(X, Y_2, Z_4),$$

we generate from the 5 PMTDs the following 2-phase disjunctive rules:

$$\begin{aligned} T_0(Z, X) \vee S_{1234}(Z) &\leftarrow \text{body} \\ T_1(Z_1, Z_2, Y_1, X) \vee S_{12}(Z_1, Z_2, X) \vee S_{1234}(Z) &\leftarrow \text{body} \\ T_1(Z_1, Z_2, Y_1, X) \vee T_2(Z_3, Z_4, Y_2, X) \vee S_{34}(Z_3, Z_4, X) \vee S_{1234}(Z) &\leftarrow \text{body} \\ T_1(Z_1, Z_2, Y_1, X) \vee T_2(Z_3, Z_4, Y_2, X) \vee S_{12}(Z_1, Z_2, X) \vee S_{1234}(Z) &\leftarrow \text{body} \end{aligned}$$

We now construct the proof sequence for each of the four rules. For the first rule, we get the tradeoff  $S \cdot T^3 \cong |\mathcal{D}|^4 \cdot |Q_A|^3$  as shown below

$$\begin{aligned} 4 \log |\mathcal{D}| + 3 \log |Q_A(Z)| &\geq 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_1Y_1X|X) + \mathbf{h}_S(Z_2Y_1X|X) + \mathbf{h}_S(Z_3Y_2X|X) + \mathbf{h}_S(Z_4Y_2X|X) + 3\mathbf{h}_T(Z) \\ &\geq 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_1Y_1X|X) + \mathbf{h}_S(Z_2Y_1X|X) + \mathbf{h}_S(Z_4Z_3Y_2X|Z_4X) + \mathbf{h}_S(Z_4X|X) + 3\mathbf{h}_T(Z) \\ &= 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_1Y_1X|X) + \mathbf{h}_S(Z_2Y_1X|X) + \mathbf{h}_S(Z_4Z_3Y_2X|X) + 3\mathbf{h}_T(Z) \\ &\geq 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_1Y_1X|X) + \mathbf{h}_S(Z_4Z_3Z_2Y_1X|Z_4Z_3X) + \mathbf{h}_S(Z_4Z_3X|X) + 3\mathbf{h}_T(Z) \\ &= 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_1Y_1X|X) + \mathbf{h}_S(Z_4Z_3Z_2Y_1X|X) + 3\mathbf{h}_T(Z) \\ &\geq 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_4Z_3Z_2Z_1Y_1X|Z_4Z_3Z_2X) + \mathbf{h}_S(Z_4Z_3Z_2X|X) + 3\mathbf{h}_T(Z) \\ &= 3\mathbf{h}_T(X) + \mathbf{h}_S(Z_4Z_3Z_2Z_1Y_1X|X) + 3\mathbf{h}_T(Z) \\ &\geq 3\mathbf{h}_T(XZ) + \mathbf{h}_S(Z) \end{aligned} \quad (S \cdot T^3 \cong |\mathcal{D}|^4 \cdot |Q_A|^3)$$

For the rest of the rules, we get the tradeoff  $S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|$ . We show the proof sequence for the second rule below. The proof sequences for the third and fourth rules are very similar and thus omitted.

$$\begin{aligned} 2 \log |\mathcal{D}| + \log |Q_A(Z)| &\geq \mathbf{h}_T(Y_1X) + \mathbf{h}_S(Z_1Y_1X|Y_1X) + \mathbf{h}_S(Z_2Y_1X|X) + \mathbf{h}_T(Z_2Z_1) \\ &\geq \mathbf{h}_T(Y_1X) + \mathbf{h}_S(Z_2Z_1Y_1X|Z_2Y_1X) + \mathbf{h}_S(Z_2Y_1X|X) + \mathbf{h}_T(Z_2Z_1) \\ &= \mathbf{h}_T(Y_1X) + \mathbf{h}_S(Z_2Z_1Y_1X|X) + \mathbf{h}_T(Z_2Z_1) \\ &\geq \mathbf{h}_T(Z_2Z_1Y_1X|X) + \mathbf{h}_S(Z_2Z_1X|X) \end{aligned} \quad (S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|)$$

Overall, we obtain the tradeoff  $S \cdot T^3 \cong |\mathcal{D}|^4 \cdot |Q_A|^3$  since  $S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|$  is dominated by it.

**Improved Tradeoffs.** We now show an alternative proof sequence that can improve upon the tradeoff  $S \cdot T^3 \cong |\mathcal{D}|^4 \cdot |Q_A|^3$  for our running example query. The key insight is to bucketize on the bound variables rather than the free variables, an idea also used by [11]. We fix the same disjunctive rules from before. For the first rule, we can obtain an improved tradeoff as follow:

$$\begin{aligned} 4 \log |\mathcal{D}| + 4 \log |Q_A(\mathbf{Z})| &= \textcolor{red}{h_T}(Y_1 X Z_1 | Z_1) + \textcolor{blue}{h_S}(Z_1) + \textcolor{red}{h_T}(Y_1 X Z_2 | Z_2) + \textcolor{blue}{h_S}(Z_2) + \textcolor{red}{h_T}(Y_2 X Z_3 | Z_3) \\ &\quad + \textcolor{blue}{h_S}(Z_3) + \textcolor{red}{h_T}(Y_2 X Z_4 | Z_4) + \textcolor{blue}{h_S}(Z_4) + 4 \textcolor{red}{h_T}(\mathbf{Z}) \\ &\geq \textcolor{blue}{h_S}(\mathbf{Z}) + \textcolor{red}{h_T}(Y_1 X Z_1 | Z_1) + \textcolor{red}{h_T}(\mathbf{Z}) + \textcolor{red}{h_T}(Y_1 X Z_2 | Z_2) + \textcolor{red}{h_T}(\mathbf{Z}) \\ &\quad + \textcolor{red}{h_T}(Y_2 X Z_3 | Z_3) + \textcolor{red}{h_T}(\mathbf{Z}) + \textcolor{red}{h_T}(Y_2 X Z_4 | Z_4) + \textcolor{red}{h_T}(\mathbf{Z}) \\ &\geq \textcolor{blue}{h_S}(\mathbf{Z}) + 4 \textcolor{red}{h_T}(X \mathbf{Z}) \end{aligned} \tag{36}$$

The above proof sequence generates the tradeoff  $S \cdot T^4 \cong |\mathcal{D}|^4 \cdot |Q_A|^4$ , a clear improvement for  $|Q_A| = 1$ . For the rest of the rules, we keep the tradeoff derived above, i.e.  $S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|^2$ .

Note that the tradeoff  $S \cdot T^3 \cong |\mathcal{D}|^4 \cdot |Q_A|^3$  dominates both  $S \cdot T^4 \cong |\mathcal{D}|^4 \cdot |Q_A|^4$  for the first rule and  $S \cdot T \cong |\mathcal{D}|^2 \cdot |Q_A|^2$  for the rest of the rules. So we get a strictly improved tradeoff across all regimes.

**Capturing Theorem F.4 in Our Framework.** Before we conclude this section, we present a general strategy to capture the tradeoff from Theorem F.4 for a subset of hierarchical queries. In particular, we show that for any Boolean hierarchical CQAP that contains  $\mathbf{x}_A$  only in the leaf variables, there is a proof sequence that recovers the tradeoff obtained from Theorem F.4. We will use  $\mathbf{Z}$  to denote the leaf variables. Recall that each hierarchical query admits a *canonical ordering*<sup>6</sup>, which is a rooted tree where the variables of each atom in the query lie along the same root-to-leaf path in the tree and each atom is a child of its lowest variable.

We begin by describing the query decomposition that we will use. Consider the canonical variable ordering of the hierarchical query. The root bag of the decomposition  $T_0$  consists of  $\mathbf{x}_A$  and the variable at the root of the variable ordering (say  $X$ ). For each child  $Y_i$  of  $X$ , we add a child bag of  $T_i$  containing the subset of  $\mathbf{x}_A$  in the subtree rooted at  $Y_i$  and  $\text{anc}(Y_i) \cup Y_i$ . We continue this procedure by traversing the variable ordering in a top-down fashion and processing all non-bound variables. It is easy to see that the tree obtained is indeed a valid decomposition.

*Example F.6.* Figure 6b shows the query decomposition generated from the canonical ordering in Figure 6a. The root bag contains all bound variables and  $X$ .  $X$  contains two children  $Y_1, Y_2$  so the decomposition contains two children of the root node. The left child contains  $X, Y_1$  and the subtree rooted at  $Y_1$  contains  $Z_1, Z_2$  as the bound variables, which are added to the left node in the decomposition. Similarly, the right node contains  $X, Y_2, Z_3, Z_4$ .

Similar to the running example, we now construct the set of PMTDs induced from this decomposition and generate the corresponding 2-phase disjunctive rules. It is easy to see that every disjunctive rule contains  $S(\mathbf{Z})$ .

Let  $U_1 \subseteq \mathbf{Z}$  and  $w_1 = |U_1|$ . For any disjunctive rule, it must contain a term of the form  $T(U_1 p)$  and  $S(\mathbf{Z})$ . Here,  $p$  denotes the set of variables other than the bound variables in the bag corresponding to the  $T$ -view that is picked. Note that the induced PMTD where  $M = \emptyset$  forces any disjunctive rule to have at least one  $T$ -view in the head of the rule. Consider some  $v^\star \in U_1$ .

$$\begin{aligned} w_1 \log |\mathcal{D}| + (w_1 - 1) \log |Q_A(\mathbf{Z})| &= \textcolor{blue}{h_S}(\text{anc}(v^\star) \cup v^\star) + \sum_{v \in \mathbf{Z} \setminus v^\star} (\textcolor{red}{h_T}(p) + \textcolor{blue}{h_S}(\text{anc}(v) \cup v \mid p)) + (w_1 - 1) \textcolor{red}{h_T}(\mathbf{Z}) \\ &\geq (w_1 - 1) \textcolor{red}{h_T}(U_1 p) + \textcolor{blue}{h_S}(\text{anc}(v^\star) \cup v^\star) + \sum_{v \in \mathbf{Z} \setminus v^\star} \textcolor{blue}{h_S}(\text{anc}(v) \cup v \mid p) \\ &= (w_1 - 1) \textcolor{red}{h_T}(U_1 p) + \textcolor{blue}{h_S}(\text{anc}(v_1) \cup v_1 \mid p) + \dots + \textcolor{blue}{h_S}(\text{anc}(v_k) \cup v_k \cup v^\star \mid v^\star \cup p) + \textcolor{blue}{h_S}(\text{anc}(v^\star) \cup v^\star) \\ &\geq (w_1 - 1) \textcolor{red}{h_T}(U_1 p) + \textcolor{blue}{h_S}(\text{anc}(v_1) \cup v_1 \mid p) + \dots + \textcolor{blue}{h_S}(\text{anc}(v_k) \cup v_k \cup v^\star \mid v^\star \cup p) + \textcolor{blue}{h_S}(p \cup v^\star) \\ &\geq (w_1 - 1) \textcolor{red}{h_T}(U_1 p) + \textcolor{blue}{h_S}(\text{anc}(v_1) \cup v_1 \mid p) + \dots + \textcolor{blue}{h_S}(\text{anc}(v_k) \cup v_k \cup v^\star) \\ &\geq (w_1 - 1) \textcolor{red}{h_T}(U_1 p) + \textcolor{blue}{h_S}(\text{anc}(v_1) \cup v_1 \dots \cup v_k \cup v^\star) \\ &\geq (w_1 - 1) \textcolor{red}{h_T}(U_1 p) + \textcolor{blue}{h_S}(\mathbf{Z}) \end{aligned} \tag{S \cdot T^{w_1-1} \cong |\mathcal{D}|^{w_1} \cdot |Q_A|^{w_1-1}}$$

The tradeoff is the most expensive when  $w_1$  is as large as possible. Thus, for  $w_1 = w$ , which corresponds to  $U_1 = \mathbf{Z}$ , we achieve the tradeoff  $S \cdot T^{w-1} \cong |\mathcal{D}|^w \cdot |Q_A(\mathbf{Z})|^{w-1}$ . However, for the same disjunctive rule  $T_0(\mathbf{Z}, A) \vee S_{\mathbf{x}_A}(\mathbf{Z})$  that gives the dominating tradeoff, we can also obtain a different proof sequence that provides an improvement, similar to (36)

$$w \log |\mathcal{D}| + w \log |Q_A(\mathbf{Z})| = \textcolor{blue}{h_S}(\mathbf{Z}) + \sum_{i=1}^w \textcolor{red}{h_T}(\text{anc}(Z_i) \cup Z_i \mid Z_i) + w \textcolor{red}{h_T}(\mathbf{Z})$$

<sup>6</sup>The original definition in [20] also has a dependency function but we omit that since we do not use it for the decomposition construction.

$$\begin{aligned}
 &= \textcolor{blue}{h}_{\mathcal{S}}(\mathbf{Z}) + \sum_{i=1}^w (\textcolor{red}{h}_{\mathcal{T}}(X \cup Z_i \mid Z_i) + \textcolor{red}{h}_{\mathcal{T}}(\mathbf{Z})) \\
 &\geq \textcolor{blue}{h}_{\mathcal{S}}(\mathbf{Z}) + w \cdot \textcolor{red}{h}_{\mathcal{T}}(X\mathbf{Z})
 \end{aligned}
 \quad (S \cdot T^w \cong |\mathcal{D}|^w \cdot |Q_A|^w)$$