Beyond Hypergraph Dualization
Lhouari Nourine, Jean-Marc Petit

To cite this version:

HAL Id: hal-01229015
https://hal.archives-ouvertes.fr/hal-01229015
Submitted on 16 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Beyond Hypergraph Dualization

Lhouari Nourine\textsuperscript{1}, Jean Marc Petit\textsuperscript{2}

\textsuperscript{1} Clermont-Université, CNRS, Université Blaise Pascal, LIMOS, France.
\textsuperscript{2} Université de Lyon, CNRS, INSA Lyon, LIRIS, France

Years and Authors of Summarized Original Work

2012; Nourine, Petit

Keywords

Dualization; Hypergraph Transversal; Partially ordered set; Enumeration

Problem Definition

This problem concerns hypergraph dualization and generalization to poset dualization.

A hypergraph $\mathcal{H} = (V, \mathcal{E})$ consists of a finite collection $\mathcal{E}$ of sets over a finite set $V$, i.e. $\mathcal{E} \subseteq \mathcal{P}(V)$ (the powerset of $V$). The elements of $\mathcal{E}$ are called hyperedges, or simply edges. A hypergraph is said simple if none of its edges is contained within another. A transversal (or hitting set) of $\mathcal{H}$ is a set $T \subseteq V$ that intersects every edge of $\mathcal{E}$. A transversal is minimal if it does not contain any other transversal as a subset. The set of all minimal transversals of $\mathcal{H}$ is denoted by $\text{Tr}(\mathcal{H})$. The hypergraph $(V, \text{Tr}(\mathcal{H}))$ is called the transversal hypergraph of $\mathcal{H}$. Given a simple hypergraph $\mathcal{H}$, the hypergraph dualization problem ($\text{TRANS-ENUM}$ for short) concerns the enumeration without repetitions of $\text{Tr}(\mathcal{H})$.

The $\text{TRANS-ENUM}$ problem can also be formulated as a dualization problem in posets. Let $(P, \leq)$ be a poset (i.e. $\leq$ is a reflexive, antisymmetric, and transitive relation on the set $P$). For $A \subseteq P$, $\downarrow A$ (resp. $\uparrow A$) is the downward (resp. upward) closure of $A$ under the relation $\leq$ (i.e. $\downarrow A$ is an ideal and $\uparrow A$ a filter of $(P, \leq)$). Two antichains $(B^+, B^-)$ of $P$ are said to be dual if $\downarrow B^+ \cup \uparrow B^- = P$ and $\downarrow B^+ \cap \uparrow B^- = \emptyset$. Given an implicit description of a poset $P$ and an antichain $B^+$ (resp. $B^-$) of $P$, the poset dualization problem ($\text{DUAL-ENUM}$ for short) enumerates the set $B^-$ (resp. $B^+$), denoted by $\text{Dual}(B^+) = B^-$ (resp. $\text{Dual}(B^-) = B^+$). Notice that the function dual is self-dual or idempotent, i.e. $\text{Dual}(\text{Dual}(B)) = B$. 
**Key Results**

**Trans-Enum** has been intensively studied in the last two decades, and several results show that it is equivalent to many problems in computer science area (see the paper by Eiter and Gottlob [3]). The question whether **Trans-Enum** admits an output-polynomial time algorithm is still open. In fact, despite the number of papers on **Trans-Enum**, the best known algorithm is the one by Fredman and Khachiyan [8] which runs in time \( O(n \log^c(n)) \) where \( n \) is the size of the hypergraph plus the number of minimal transversals. Others results on complexity can be found in [5, 14, 6, 12, 11]. For general posets, it is shown in [7] that the dualization over the products of some posets can be done with the same complexity as **Trans-Enum**. Recently, Nourine and Petit [16] have investigated dualization problems in general posets for which the duality gap is bounded by a polynomial.

**Strong Duality**

The following characterization theorem of the zero gap, is a reformulation of a known result in [15, 10], where the poset \( Q \) is the powerset for some set.

**Theorem 1.** Let \((P, \leq_P)\) and \((Q, \leq_Q)\) be two posets. Then the duality gap is zero iff there exists a map \( f : P \to Q \) such that \( f \) is a bijective embedding, i.e. for all \( x, y \in P \) \( f(x) \leq_Q f(y) \) iff \( x \leq_P y \).

Many instances of problems have such a property, for example frequent itemsets, monotone boolean functions, minimal keys, inclusion dependencies or minimal dominating sets [15, 10, 13]. Nevertheless, the bijective embedding between two posets does not always exist. In the following we give a relaxation of the bijection embedding in order to capture some polynomial reductions between enumeration problems.

**Weak Duality**

Let \((P, \leq_P)\) and \((Q, \leq_Q)\) be posets. A function \( f : P \to Q \) is a convex embedding if for all \( x, y \in P \) and \( z \in Q \) \( x \leq_P y \) iff \( f(x) \leq_Q f(y) \) and \( f(x) \leq_Q z \leq_Q f(y) \) implies there exists \( t \in P \) such that \( f(t) = z \).
The following result can be seen as a relaxation of the bijective embedding given in Theorem 1

**Proposition 1.** Let \((P, \leq_P)\) and \((Q, \leq_Q)\) be two posets and \(f : P \rightarrow Q\) a convex embedding. Then there exist two antichains \(B^+_0, B^-_0\) of \(Q\) such that \(P \setminus \{\bot_P\}\) is isomorphic to \(Q \setminus (\downarrow B^+_0 \cup \uparrow B^-_0)\), where \(\bot_P\) is the bottom of \(P\) if it exists. Furthermore, the duality gap is bounded by \(|B^+_0| + |B^-_0|\).

**Complexity**

For strong duality, \([15] [10]\) points out how the result of Fredman and Khachiyan \([8]\) can be re-used to devise an incremental quasi-polynomial time algorithm, called Dualize and Advance, for some pattern mining problems. For weak duality, whenever the duality gap remains polynomial in the size of the problem and \((Q, \leq_Q)\) isomorphic to \((P(E), \subseteq)\) for some set \(E\), the Dualize and Advance algorithm can be re-used with the same complexity if the following assumptions hold:

1. The reflection \(f\) of \((P, \leq)\) to \((P(E), \subseteq)\) and its inverse, is computable in polynomial time.
2. Given two elements \(x, y \in P\), checking \(x \leq y\) is polynomial time.

**Applications**

The hypergraph dualization is a crucial step in many applications in logics, databases, artificial intelligence and pattern mining \([11] [3] [8] [4] [15]\), especially for hypergraphs, i.e. boolean lattices. The main application domain concerns pattern mining problems, i.e. the identification of maximal interesting patterns in database by asking membership queries (predicate) to a database. In the rest of this section, we give two examples of pattern mining problems related to Dual-Enum and weak duality.

**Frequent conjunctive queries**

We consider the problem statement defined in \([9]\). Let \(R = \{R_1, \ldots, R_n\}\) be a database schema, \(D\) the domain of \(R\) and \(sch(R) = \{R_i.A | R_i \in R, A \in R_i\}\). A (simple) conjunctive queries \(Q\) over \(R\) is of the form \(\pi_X(\sigma_F(R_1 \times \ldots \times R_n))\) (\(\pi_X(\sigma_F)\) for short) where \(X \subseteq sch(R)\) and \(F\) a conjunction of equalities of the form \(R_i.A = c\) with \(R_i.A, R_i.B \in sch(R)\) and \(c \in D\). Let \(Q_r\) be the set of all possible conjunctive queries over \(R\). For a given database \(d\) over \(R\), we note \(Adom(d) \subseteq D\) the active domain of \(d\) and \(Q(d)\) the result of the evaluation of \(Q\) against \(d\). We note \(F\) the finite set of all possible selection formula over \(R\) and \(Adom(d)\), i.e. \(F = \{\{A, B\} \mid A \neq B, A \in R, B \in R \cup Adom(d)\}\).

Let \(Q_1, Q_2\) be two conjunctive queries over \(R\). \(Q_1\) is contained in \(Q_2\), denoted \(Q_1 \subseteq Q_2\), if for every database \(d\) over \(R\), \(Q_1(d) \subseteq Q_2(d)\). \(Q_1\) is diagonally contained in \(Q_2\), denoted \(Q_1 \subseteq^\Delta Q_2\), if \(Q_1\) is contained in a projection of \(Q_2\), i.e. for instance \(Q_1 \subseteq \pi_X(Q_2)\). The frequency of \(\pi_X(\sigma_F)\) in \(d\) is defined by \(|\pi_X(\sigma_F)(d)|\). A query \(\pi_X(\sigma_F)\) is frequent in \(d\) with respect to a given threshold \(\epsilon\) if \(|\pi_X(\sigma_F)(d)| \geq \epsilon\). The frequency is anti-monotonic with respect to \(\subseteq^\Delta\) \([9]\).

**Proposition 2.** Let \(Q_1 = \pi_X(\sigma_{F_1})\) and \(Q_2 = \pi_X(\sigma_{F_2})\) be two queries of \(Q_r\). Then \(Q_1 \subseteq^\Delta Q_2\) iff \(X_1 \subseteq X_2\) and \(F_2 \subseteq F_1\). Equivalently, \(Q_1 \subseteq^\Delta Q_2\) iff \(X_1 \cup (F \setminus F_1) \subseteq X_2 \cup (F \setminus F_2)\).
From Proposition 2 \( f : Q_r \to \mathcal{P}(R \cup F) \) with \( f(\pi_X(\sigma_F)) = X \cup (F \setminus F) \) is a bijective embedding. Thus \( Q_r \) ordered under \( \subseteq^A \) is a boolean lattice and Theorem 1 can be applied. It is interesting to consider the subclass of \( Q_r \) restricted to consistent queries, i.e. queries for which there exists at least one database such that their evaluations return values different from zero. For instance, \( \sigma(B = 1 \land B = 2) \) or \( \sigma(A = B \land A = 1 \land B = 2) \) are not consistent. Let us consider the set \( Q_C \subset Q_r \) of all consistent queries.

**Lemma 1.** Let \( Q_1 = \pi_{X_1}(\sigma_{F_1}) \) and \( Q_2 = \pi_{X_2}(\sigma_{F_2}) \) be two queries of \( Q_r \) such that \( Q_1 \subseteq^A Q_2 \). Then \( Q_2 \) is consistent implies \( Q_1 \) is consistent.

Notice that the restriction of \( f \) to \( Q_C \) is still a convex embedding, but no longer bijective. More interestingly, the associated duality gap is not polynomial. Indeed, \( B_0^+ = \emptyset \) but \( B_0^- \) has a size exponential in the size of \( R \cup \text{Adom}(d) \) since the number of selections of the form \( \sigma(A_1 = A_2 \land \ldots \land A_{n-1} = A_n \land A_1 = v \land A_n = v') \) is exponential in the number of attributes.

### Rigid sequences

Let us consider sequences with or without wildcard (denoted \( * \)), see e.g. [1]. Let \( \Sigma \) be an alphabet and \( * \notin \Sigma \). A rigid sequence \( s[n] \) is a word of size \( n \) of \( (\Sigma \cup \{ * \})^\ast \) such that \( s[1] \neq * \) and \( s[n] \neq * \). The set of all rigid sequences of size at most \( n \) are denoted by \( \Sigma_R \) and the empty sequence by \( \epsilon \). Let \( s[i], t[k] \in \Sigma_R \). We consider the following classical (prefix and factor) partial orders on rigid sequences.

- \( s \sqsubseteq f t \), if there exists \( j \in [1..k] \) such that for every \( i \in [1..l] \), either \( s[i] = t[j] + i - 1 \) or \( s[i] = * \) (prefix).
- \( s \sqsubseteq p t \), if for every \( i \in [1..l] \), either \( s[i] = t[i] \) or \( s[i] = * \) (factor).

The following theorem shows that the duality gap between the dualization in prefix posets of rigid sequences and TRANS-ENUM is bounded by a polynomial in \( n \) and \( |\Sigma| \).

**Theorem 2.** [16] Let \( f : (\Sigma_R \setminus \{ \epsilon \}, \subseteq_p) \to (\mathcal{P}\{1, \ldots, n\} \times \Sigma, \subseteq) \) be a function defined by \( f(s) = \{(i, s[i]) \mid s[i] \neq *, i \leq n\} \). Then \( f \) is a convex embedding with \( B_0^+ = \{(i, x) \mid x \in \Sigma, i \in [2..n]\} \) and \( B_0^- = \{(1, x), (1, y) \mid x, y \in \Sigma, x \neq y\} \cup \{(1, x), (i, y), (i, z) \mid x, y, z \in \Sigma, y \neq z, i \in [2..n]\} \).

**Proposition 3.** [16] There is a poset reflection \( f : (\Sigma_R, \subseteq_f) \to (\Sigma_R, \subseteq_p) \) with a duality gap bounded by a polynomial in \( n \).

Using Theorem 2 and Proposition 3 we conclude that the duality gap between the dualization in factor posets of rigid sequences and TRANS-ENUM is bounded by a polynomial the size of \( \Sigma \) and \( n \) [16].

### Open Problems

1. The challenging question is to find an output-polynomial time algorithm for TRANS-ENUM.
2. Lattices are a particular class of posets. For example, the dualization over product of chains can be done with the same complexity as TRANS-ENUM which is equivalent to dualization in boolean lattices. For distributive lattices class which contains boolean lattice and the product of chains, the dualization is open.
3. Many connections have to be done between TRANS-ENUM and graph theory problems, such as minimal dominating sets [13].

4. Many problems in data mining can be formulated as dualization in posets, e.g. frequent subgraphs or frequent subtrees. An interesting direction is to identify posets for which the dualization is equivalent to TRANS-ENUM.

URLs to Code and Data Sets
Program Codes and Instances for Hypergraph Dualization can be found on the Takeaki Uno’s webpage at [http://research.nii.ac.jp/~uno/dualization.html](http://research.nii.ac.jp/~uno/dualization.html). Some pattern mining problems, reducible to TRANS-ENUM with strong duality, can be found on the iZi webpage at [http://liris.cnrs.fr/izi/](http://liris.cnrs.fr/izi/).

Cross-References

- Minimal Dominating Set Enumeration
- Polynomial Time Algorithms for Dualization; Solvable Cases
- Frequent Itemset Mining

Recommended Reading