Boolean Tensor Decomposition for Conjunctive Queries with Negation

Mahmoud Abo Khamis    Hung Q. Ngo    Dan Olteanu    Dan Suciu
RelationalAI (USA) & U. Oxford (UK) & U. Washington (USA)

International Conference on Database Theory
Lisbon, March 2019
Conjunctive Queries with Negated Bounded-Degree Relations

\[ Q(X_F) \leftarrow \text{body} \land \bigwedge_{S \in \overline{E}} \neg R_S(X_S), \]

- **body** is the body of an arbitrary (positive) conjunctive query
- \( X_F = (X_i)_{i \in F} \) denotes a tuple of variables indexed by \( F \subset \mathbb{N} \)
- \( \overline{E} \) is the set of hyperedges of a multi-hypergraph \( \overline{H} = (\overline{V}, \overline{E}) \)
- Each \( S \in \overline{E} \) corresponds to a **bounded-degree relation** \( R_S \)
Directed graph $G = ([n], E)$ with $n$ nodes and $N = |E|$ edges.

$W() \leftarrow E(X_1, X_2) \land E(X_2, X_3) \land \cdots \land E(X_k, X_{k+1})$

Hypergraph $\overline{H}$ is empty since $W$ has no negated relations.
Query Example 1/3: $k$-walk

Directed graph $G = ([n], E)$ with $n$ nodes and $N = |E|$ edges.

$W() \leftarrow E(X_1, X_2) \land E(X_2, X_3) \land \cdots \land E(X_k, X_{k+1})$

Hypergraph $\overline{H}$ is empty since $W$ has no negated relations.

Time complexity:

- $O(kN \log N)$ [Yannakakis’81]
Directed graph \( G = ([n], E) \) with \( n \) nodes and \( N = |E| \) edges.

\[
P() \leftarrow E(X_1, X_2) \land E(X_2, X_3) \land \cdots \land E(X_k, X_{k+1}) \land \\
\bigwedge_{i,j \in [k+1], i+1 < j} X_i \neq X_j
\]

Disequality is negation of bounded-degree equality relation:

\[
X_i \neq X_j \equiv \lnot (X_i = X_j)
\]

Hypergraph \( \overline{\mathcal{H}} = ([k + 1], \{(i, j) \mid i, j \in [k + 1], i + 1 < j\}) \)
Query Example 2/3: $k$-path

Directed graph $G = ([n], E)$ with $n$ nodes and $N = |E|$ edges.

$$P() \leftarrow E(X_1, X_2) \land E(X_2, X_3) \land \cdots \land E(X_k, X_{k+1}) \land \bigwedge_{i,j \in [k+1], i+1 < j} X_i \neq X_j$$

Disequality is negation of bounded-degree equality relation:

$$X_i \neq X_j \equiv \neg(X_i = X_j)$$

Hypergraph $\overline{H} = ([k+1], \{(i, j) \mid i, j \in [k+1], i + 1 < j\})$

Time complexity:

- $O(k^k N \log N)$ [Plehn, Voigt'90]
- $2^{O(k)} N \log N$ using color-coding [Alon, Yuster, Zwick'95]
Directed graph $G = ([n], E)$ with $n$ nodes and $N = |E|$ edges.

$I() \leftarrow E(X_1, X_2) \land E(X_2, X_3) \land \cdots \land E(X_k, X_{k+1}) \land \bigwedge_{i,j \in [k+1], i+1 < j} (\neg E(X_i, X_j) \land X_i \neq X_j))$

Each edge twice in $\overline{H}$ due to negated edge relation and disequality
Directed graph $G = ([n], E)$ with $n$ nodes and $N = |E|$ edges.

$I() \leftarrow E(X_1, X_2) \land E(X_2, X_3) \land \cdots \land E(X_k, X_{k+1}) \land \bigwedge_{i,j \in [k+1], \ i+1 < j} (\neg E(X_i, X_j) \land X_i \neq X_j)\) \bigwedge

Each edge twice in $\overline{H}$ due to negated edge relation and disequality

Time complexity:

- $O(f(k, d)N \log N)$ if $G$ has maximum degree $d$;
  $f$ depends exponentially on $k$ and $d$ [Plehn, Voigt’90]
Main Result: Time Complexity for Query Evaluation

Database with relations of size $O(N)$

Query $Q$ with positive body and negation hypergraph $\overline{H}$
Main Result: Time Complexity for Query Evaluation

Database with relations of size $\mathcal{O}(N)$

Query $Q$ with positive body and negation hypergraph $\overline{\mathcal{H}}$

Using a reduction to InsideOut $[\text{Abo Khamis et al’16}]$

\[
\mathcal{O}(F_{\text{InsideOut}}(Q) \cdot \log N \cdot (N^{fhtw_F(\text{body})} + |\text{output}|))
\]

depends on structure of $\overline{\mathcal{H}}$
degree of relations and InsideOut

same as for body
Main Result: Time Complexity for Query Evaluation

Database with relations of size $O(N)$

Query $Q$ with positive body and negation hypergraph $\overline{H}$

Using a reduction to InsideOut

$$\mathcal{O}(F_{\text{InsideOut}}(Q) \cdot \log N \cdot (N^{\text{fhtw}_F(\text{body})} + |\text{output}|))$$

depends on structure of $\overline{H}$
degree of relations
and InsideOut

same as for body

Using a reduction to PANDA

$$\mathcal{O}(F_{\text{PANDA}}(Q) \cdot (\text{poly}(\log N) \cdot N^{\text{subw}_F(\text{body})} + \log N \cdot |\text{output}|))$$

depends on structure of $\overline{H}$
degree of relations
and PANDA

same as for body
Our Query Evaluation Approach

1. Untangling negated bounded-degree relations

   Rewrite negated subquery into not-all-equal conjunction

   Not-all-equal (NAE) is multi-dimensional analog of $\neq$

2. Boolean tensor decomposition for NAE conjunction

   Probabilistic construction with efficient derandomization

   Generalization of color-coding from cliques of $\neq$ to NAE conjunctions

3. Use existing algorithms InsideOut and PANDA

   Decomposition preserves fhtw and subw of positive body
Untangling Bounded-Degree Relations
The Untangling Step via an Example

Given: Database with relations \( R, S, T \) with sizes \( \mathcal{O}(N) \)

Task: Compute the Boolean query

\[
Q() \leftarrow R(A, B) \land S(B, C) \land \neg T(A, C)
\]

What is the time complexity for computing \( Q \)?

- \( \mathcal{O}(N^2) \) trivially: First join \( R \) and \( S \) and then filter with \( T \)
The Untangling Step via an Example

Given: Database with relations $R, S, T$ with sizes $O(N)$

Task: Compute the Boolean query

$$Q() \leftarrow R(A, B) \land S(B, C) \land \neg T(A, C)$$

What is the time complexity for computing $Q$?

- $O(N^2)$ trivially: First join $R$ and $S$ and then filter with $T$
- Subquadratic if $T$ has degree bounded by a constant
Intermezzo: Bounded-degree Relations

Classical notion of degree $\Delta(T)$ of relation $T(A, C)$:

Maximum number of tuples with the same value for $A$ or $C$
Classical notion of degree $\Delta(T)$ of relation $T(A, C)$:

Maximum number of tuples with the same value for $A$ or $C$

Our notion of degree $\text{deg}(T)$ accounts for the arity of $T$:

Smallest number $d$ such that $T$ is a disjoint union of $d$ matchings

If $T$ has schema $S$: $\Delta(T) \leq \text{deg}(T) \leq |S| \cdot (\Delta(T) - 1) + 1$
Intermezzo: Bounded-degree Relations

Classical notion of degree $\Delta(T)$ of relation $T(A, C)$:

Maximum number of tuples with the same value for $A$ or $C$

Our notion of degree $\text{deg}(T)$ accounts for the arity of $T$:

Smallest number $d$ such that $T$ is a disjoint union of $d$ matchings

If $T$ has schema $S$: $\Delta(T) \leq \text{deg}(T) \leq |S| \cdot (\Delta(T) - 1) + 1$

Assumption in our example: $T$ has degree 2, that is,

$\exists$ matchings $M_1$ and $M_2$: $T(A, C) \equiv M_1(A, C) \lor M_2(A, C)$
Intermezzo: What is a Matching?

$M$ is matching iff $\forall x_S, x'_S \in M$ either $x_S = x'_S$ or $\forall i \in S : x_i \neq x'_i$

Linear-time decomposition of relation $R$ into $|S| \cdot \Delta(R)$ matchings
Assume matching $M_i(A, C)$. When is $(\bullet, \blacksquare) \in \neg M_i$?

1. \blacksquare is in the domain of $C$ but not in $M_i$

$$W_i(C) = \text{Dom}(C) \wedge \neg (\exists X M_i(X, C))$$
Intermezzo: Negating a Binary Matching

Assume matching $M_i(A, C)$. When is $(\bigcirc, \blacksquare) \in \neg M_i$?

1. $\blacksquare$ is in the domain of $C$ but not in $M_i$

$$W_i(C) = \text{Dom}(C) \land \neg (\exists X M_i(X, C))$$

2. or $\blacksquare$ is paired with $\bigcirc \neq \bigcirc$ in $M_i$

$$\exists A_i (M(A_i, C) \land A_i \neq A)$$
Intermezzo: Negating a Binary Matching

Assume matching $M_i(A, C)$. When is $(\bullet, \blacksquare) \in \neg M_i$?

1. $\blacksquare$ is in the domain of $C$ but not in $M_i$
   
   $W_i(C) = \text{Dom}(C) \land \neg(\exists X M_i(X, C))$

2. or $\blacksquare$ is paired with $\bullet \neq \bullet$ in $M_i$
   
   $\exists A_i (M(A_i, C) \land A_i \neq A)$

$\neg M_i(A, C) \equiv W_i(C) \lor \exists A_i (M(A_i, C) \land A_i \neq A)$
Negating a Bounded-degree Relation

Recall: \( T(A, C) \equiv M_1(A, C) \lor M_2(A, C) \), \( M_1 \) and \( M_2 \) matchings

\[
\neg T(A, C) \equiv \neg M_1(A, C) \land \neg M_2(A, C)
\]

\[
\begin{align*}
W_1(C) \land \exists A_1 (M_1(A_1, C) \land A_1 \neq A) & \quad \text{and} \quad W_2(C) \land \exists A_2 (M_2(A_2, C) \land A_2 \neq A) \\
\end{align*}
\]

Flatten out \( \neg T(A, C) \) into a disjunction of four conjunctions:

\[
\begin{align*}
W_1(C) \land W_2(C) \\
W_1(C) \land M_2(A_2, C) \land A \neq A_2 \\
W_2(C) \land M_1(A_1, C) \land A \neq A_1 \\
M_1(A_1, C) \land M_2(A_2, C) \land A \neq A_1 \land A \neq A_2 \\
\end{align*}
\]

The negative subqueries are now disequalities on variables
The Untangling Step

The query $Q$ becomes $Q_1 \lor Q_2 \lor Q_3 \lor Q_4$:

$Q_1() \leftarrow R(A, B) \land S(B, C) \land W_1(C) \land W_2(C)$

$Q_2() \leftarrow R(A, B) \land S(B, C) \land W_1(C) \land M_2(A_2, C) \land A \neq A_2$

$Q_3() \leftarrow R(A, B) \land S(B, C) \land W_2(C) \land M_1(A_1, C) \land A \neq A_1$

$Q_4() \leftarrow R(A, B) \land S(B, C) \land M_1(A_1, C) \land M_2(A_2, C) \land A \neq A_1 \land A \neq A_2$

Our rewriting

- **extends** the positive body of $Q$
  - Replaced $T$ by (conjunctions of some of) its matchings
  - Added unary relations

- **preserves** the data complexity (fhtw and subw) of body

- **blows up** the query size exponentially in the degree
Boolean Tensor Decomposition
∀i ∈ [\log N], f_i : \text{Dom}(A) \rightarrow \{0, 1\} gives the i-th bit of A

\[ A \neq A_2 \equiv \bigvee_{x \in \{0,1\}} \bigvee_{i \in [\log N]} f_i(A) = x \land f_i(A_2) \neq x \]

This is a Boolean decomposition of \( A \neq A_2 \):

- **Rank** \( r \) is the number \( 2 \log N \) of conjuncts
- Each conjunct is a conjunction of positive unary relations

Analogy: Each function \( f_i \) is a “coloring”:

It assigns a \( \{0, 1\} \) color to each element of \text{Dom}(A)
How to Evaluate Conjunctions of Disequalities Efficiently?

$Q_2$ becomes the disjunction of $2 \log N$ acyclic queries

$$Q_2^{x,i} \leftarrow R(A, B) \land S(B, C) \land W_1(C) \land M_2(A_2, C) \land f_i(A) = x \land f_i(A_2) \neq x$$

Time complexity:

- $Q_2^{x,i}$ can be answered in time $O(N \log N)$
- $Q_2$ can be answered in time $O(N \log^2 N)$
- Further shave off a $\log N$ factor (see paper)

Boolean semiring $\rightarrow$ Bit-vector semiring
How to Evaluate Conjunctions of Disequalities Efficiently?

$Q_2$ becomes the disjunction of $2 \log N$ acyclic queries

\[
Q_2^{x,i} \leftarrow R(A, B) \land S(B, C) \land W_1(C) \land M_2(A_2, C) \land f_i(A) = x \land f_i(A_2) \neq x
\]

Time complexity:

- $Q_2^{x,i}$ can be answered in time $O(N \log N)$
- $Q_2$ can be answered in time $O(N \log^2 N)$
- Further shave off a $\log N$ factor (see paper)

Boolean semiring $\rightarrow$ Bit-vector semiring

$Q_4$ is more involved: $A \neq A_1 \land A \neq A_2$

- Three-dimensional tensor of Boolean rank $\log^2 N$
- We can reduce the rank to $\log N$
Boolean Tensor Decomposition for $A \neq A_1 \land A \neq A_2$

$$A \neq A_1 \land A \neq A_2 \equiv \bigvee_{(c,c_1,c_2) \in \{0,1\}^3} \bigvee_{f \in F} f(A) = c \land f(A_1) = c_1 \land f(A_2) = c_2$$

There exists a family $\mathcal{F}$ of functions $f : \text{Dom}(A) \to \{0,1\}$:

- $\forall (a, a_1, a_2) \in \text{Dom}(A)^3$ st $a \neq a_1 \land a \neq a_2$:
  $$\exists f \in \mathcal{F} \quad \text{st} \quad f(a) \neq f(a_1) \land f(a) \neq f(a_2)$$

- $|\mathcal{F}| = \mathcal{O}(\log N)$

- $\mathcal{F}$ can be constructed in time $\mathcal{O}(N \log N)$
**Intermezzo: Disjunct Matrices**

A \( k \text{-disjunct} \ t \times N \) matrix \( X \):

\[
\forall j \in [N], S \subseteq [N] \st |S| \leq k, j \not\in S : \exists i \in [t] \st X_{i,j} = 1, (X_{i,j'})_{j' \in S} = 0
\]
**Intermezzo: Disjunct Matrices**

A **$k$-disjunct** $t \times N$ matrix $X$ is defined as follows:

\[ \forall j \in [N], S \subseteq [N] \text{ st } |S| \leq k, j \not\in S : \exists i \in [t] \text{ st } X_{i,j} = 1, (X_{i,j'})_{j' \in S} = 0 \]

We can construct a $k$-disjunct matrix $X$:

- with $t = O(k^2 \log N)$
- in time $O(k^2 N \log N)$

[Porat, Rothschild’11]
How to Use Disjunct Matrices for Our Problem?

Each row $i = \text{function } f_i$ in $\mathcal{F}$

$X_{i,j} = f_i(A)$

$X_{i,S} = [f_i(A_1), f_i(A_2)] \Rightarrow k = 2$

- $X$ has size $O(\log N) \times N$
- $X$ constructed in time $O(N \log N)$

$$
\begin{bmatrix}
0 & 0 & 0 \\
1 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\Rightarrow 
\mathcal{O}(\log N)
$$
Generalizing the Example
Negating a Ternary Matching

Matching $M(X_1, X_2, X_3)$. Single out (wlog) $X_3$.

Tuple $(x_1, x_2, x_3) \in \neg M$ iff

1. At least one of $x_1$ or $x_2$ is not in $M$ OR

2. $x_1$ and $x_2$ are in $M$, but at least one is paired with $x_3' \neq x_3$ OR they are paired with diff. $X_3$ values

See paper for extension to $k$-ary matchings.
Matching $M(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$. Single out (wlog) $\mathbf{x}_3$.

Tuple $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \in \neg M$ iff

1. At least one of $\mathbf{x}_1$ or $\mathbf{x}_2$ is not in $M$ OR
2. $\mathbf{x}_1$ and $\mathbf{x}_2$ are in $M$, but at least one is paired with $\mathbf{x}_3' \neq \mathbf{x}_3$ OR
   they are paired with diff. $\mathbf{x}_3$ values

\[
\neg M(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \equiv (W_1(\mathbf{x}_1) \lor W_2(\mathbf{x}_2)) \lor \\
\exists \mathbf{y}_1 \exists \mathbf{y}_2 [\text{NAE}(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}_3) \land M(\mathbf{x}_1, \_ \_ \mathbf{y}_1) \land M(\_ \_ \mathbf{x}_2, \mathbf{y}_2)]
\]

\[
\text{NAE}(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}_3) \overset{\text{def}}{=} \neg (\mathbf{y}_1 = \mathbf{y}_2 \land \mathbf{y}_1 = \mathbf{x}_3 \land \mathbf{y}_2 = \mathbf{x}_3) \\
= \quad Y_1 \neq Y_2 \lor Y_1 \neq X_3 \lor Y_2 \neq X_3
\]

See paper for extension to $k$-ary matchings.
General Untangling

Query \( Q \) rewritten into a disjunction of queries

\[
Q_i(X_F) \leftarrow \text{body}_i \land \bigwedge_{S \in A_i} \text{NAE}(Z_S).
\]

Data complexity (fhtw and subw) of \( \text{body}_i \); same as for \( \text{body} \)

Number of queries \( Q_i \); exponential in the degree
General Boolean Tensor Decomposition

\[ \bigwedge_{S} \text{NAE}(Z_S) \equiv \bigvee_{j \in [r]} \bigwedge_{i \in \bigcup S Z_S} f_i^{(j)}(Z_i) \]

- \( \bigwedge_{S} \text{NAE}(Z_S) \): rank-\( r \) tensor multivariate function
- \( \bigvee_{j \in [r]} \bigwedge_{i \in \bigcup S Z_S} f_i^{(j)}(Z_i) \): univariate function
- rank-1 tensor

Multi-hypergraph \( G = (\bigcup S Z_S, A) \) of \( \bigwedge_{S} \text{NAE}(Z_S) \)

Boolean rank \( r = P(G, c) \cdot |F| \) depends on:

- Chromatic polynomial of \( G \) using \( c \leq |\bigcup S Z_S| \) colors
- Size of a family of hash functions that represent proper \( c \)-colorings of homomorphic images of \( G \)
General Boolean Tensor Decomposition

$\bigwedge_{S} \text{NAE}(Z_{S}) \equiv \bigvee_{j \in [r]} \bigwedge_{i \in \bigcup S Z_{S}} f_{i}^{(j)}(Z_{i})$

Multi-hypergraph $\mathcal{G} = (\bigcup_{S} Z_{S}, \mathcal{A})$ of $\bigwedge_{S} \text{NAE}(Z_{S})$

Boolean rank $r = P(\mathcal{G}, c) \cdot |\mathcal{F}|$ depends on:

- Chromatic polynomial of $\mathcal{G}$ using $c \leq |\bigcup_{S} Z_{S}|$ colors
  
  $c =$ maximum chromatic number of a hypergraph defined by any homomorphic image of $\mathcal{G}$

- Size of a family of hash functions that represent proper $c$-colorings of homomorphic images of $\mathcal{G}$