Optimal Upper and Lower Bounds for Boolean Expressions by Dissociation

Wolfgang Gatterbauer*, Dan Suciu
Computer Science and Engineering, University of Washington, Seattle, WA, USA

Abstract

This paper develops upper and lower bounds for the probability of Boolean expressions by treating multiple occurrences of variables as independent and assigning them new individual probabilities. Our technique generalizes and extends the underlying idea of a number of recent approaches which are variously called node splitting, variable renaming, variable splitting, or dissociation for probabilistic databases. We prove that the probabilities we assign to new variables are the best possible in some sense.

Keywords: Boolean expressions, approximation algorithms, node splitting, relaxation, probabilistic databases, partition function

1. Introduction

Several recent papers propose to approximate an intractable counting problem with a tractable relaxed version by treating multiple occurrences of variables, nodes, or tuples as independent, or ignoring some constraints. Choi et al. [1] approximate inference in Bayesian networks by node splitting, i.e. removing some dependencies from the original model, and show how the technique subsumes mini-bucket elimination [2]. Ramirez and Geffner [3] treat the problem of obtaining a minimum cost satisfying assignment of a CNF formula by variable renaming, i.e. replacing a variable that appears in many clauses by many new variables that appear in few. Pipatsrisawat and Darwiche [4] provide lower bounds for MaxSAT by variable splitting, i.e. compiling a relaxation of the original CNF. Andersen et al. [5] relax constraint satisfaction problems by refinement through node splitting, i.e. ignoring some interactions between variables. In our recent work [6], we develop a technique called dissociation to approximate the ranking of answers to intractable conjunctive queries.

In this paper, we study the probabilities of Boolean expression after treating some occurrences of variables as independent, and assigning them new individual probability values. We call this approach dissociation. It turns out that the unifying idea of the above papers is to provide either lower bounds for conjunctive expressions, or upper bounds for disjunctive expressions by assigning dissociated variables their original probabilities. We show that these bounds can be understood as duals and give new and non-obvious upper bounds for conjunctive and lower bounds for disjunctive expression (see Fig. 1). We further show that those are optimal in some sense.

We start with some necessary notations and definitions (Sect. 2) and provide our main results (Sect. 4), treat binary dissociation separately (Sect. 5), illustrate the results with examples (Sect. 6), and give the full proofs in the appendix.

2. Notational and Mathematical Background

We use \([k]\) as short notation for \(\{1, \ldots, k\}\), write \(x_i\) as short notation for \(x_i, i \in [k]\) if \(k\) is clear from the context or not relevant, use the bar sign for the complement of an event or probability (e.g., \(\bar{x} = \neg x\)), and use a bold notation for sets or vectors of variables (e.g., \(\mathbf{x} = (x_1, \ldots, x_k)\)) alike. Probabilities are always assumed to be between 0 and 1.

Our treatment of Boolean expressions is notably inspired by Crama and Hammer [7], and by Fuhr and Rölleke [8]. We assume a set \(\mathbf{x} = \{x_1, \ldots, x_k\}\) of independent Boolean random variables, and assign to each variable \(x_i\) a primitive event (we do not formally distinguish between the variable \(x_i\) and the event \(\mathbf{x}\) that it is true) which is true with probability \(\mathbf{P}[x_i] = p_i\). Thus, all primitive events are assumed to be independent (e.g., \(\mathbf{P}[x_1 x_2] = p_1 p_2\)). We are interested in computing the probabilities of composed events, i.e. event expressions \(\phi\) which are logically composed of primitive events.

In our formalism, we make further use of a set \(\mathbf{A}\) of complex events, which are composed from primitive events. It is known that arbitrary correlations can be represented with event expressions starting from independent Boolean random variables only (see Appendix A). The correlation \(\rho(A, B)\) between Boolean events \(A\) and \(B\) is defined as \(\rho(A, B) = \frac{\text{cov}(A, B)}{\text{var}(A) \cdot \text{var}(B)}\), with co-variance \(\text{cov}(A, B) = \mathbf{P}[AB] - \mathbf{P}[A] \mathbf{P}[B]\) and variance \(\text{var}(A) = \mathbf{P}[A] - (\mathbf{P}[A])^2\). Hence, complex events can be arbitrarily correlated (e.g., \(0 \leq \mathbf{P}[AB] \leq \mathbf{P}[A]\), or equally \(\rho(A, B) \leq 1\)).

We write \(\phi(\mathbf{x})\) to indicate that \(\mathbf{x}\) is a set of primitive events appearing in the expression \(\phi\). Whenever we write \(\phi(\mathbf{x}, \mathbf{A})\),
we imply that a set \( x \) of primitive events and a set \( A \) of complex events are both appearing in \( \varphi \), and that \( x \) are independent of \( A \), i.e. the complex events \( A \) are composed of primitive events different from \( x \). For example, \( \varphi(x, A) \) may be defined as \( \varphi = \{x_1, [A, B]\} = xA \land xB \), and the complex events \( A \) and \( B \) as \( A := y_1 y_2 \) and \( B := \bar{y}_1 y_2 \) over \( y = (y_1, y_2) \).

The dual of a Boolean expression is obtained by exchanging the operators \( \lor \) and \( \land \), as well as the constants 0 and 1 [7 Def. 1.8]. The duality principle states that if a Boolean expression is valid, then so is its dual [7 Th. 4.4]. It plays an important role in our paper. The two dual De Morgan’s laws [7 Th. 1.1.10] state that \( \lnot(A \lor B) = \bar{A} \land \bar{B} \) and \( \lnot(A \land B) = A \lor B \). With absorption, we refer to the two dual identities [7 Th. 1.1.11]:

\[
A \lor B = A \bar{B} \land B = (A \lor B) \bar{B} .
\]

The (disjunctive) inclusion-exclusion principle [10 R. 11.8.1] and its less-known conjunctive dual statement:

\[
\begin{align*}
P[A \lor B] &= P[A] + P[B] - P[AB] \quad \text{(if \( P[AB] = 0 \))} \\
P[AB] &= P[A] + P[B] - 1 \quad \text{(if \( P[A \lor B] = 1 \))}
\end{align*}
\]

From absorption and two special inclusion-exclusion cases:

\[
\begin{align*}
P[A \lor B] &= P[A] + P[B] \\
P[AB] &= P[A] + P[B] - 1 \quad \text{(if \( P[A \lor B] = 1 \))}
\end{align*}
\]

we get the following dual rules, which we call event splitting:

\[
\begin{align*}
P[A \lor B] &= P[A \bar{B}] + P[B] \\
P[AB] &= P[A \lor B] + P[B] - 1 .
\end{align*}
\]

3. Dissociation and Statically-tight Bounds

In this paper, we are interested in statically-tight bounds for dissociated expressions. In this section, we define and illustrate these two concepts. Intuitively, a dissociation \( \varphi' \) of an expression \( \varphi \) is derived by treating multiple appearances of the same variable as independent, and assigning them individual new probabilities. We are then interested in assigning probabilities to these new variables so that the probability of the dissociated expression \( P[\varphi'] \) is always either an upper or lower bound for \( P[\varphi] \). Furthermore, we want to assign such probabilities which (i) can become tight, and which (ii) guarantee the best bounds possible when ignoring all the other variables. We call such bounds statically-tight.

**Definition 3.1 (Dissociation).** A dissociation of a Boolean expression \( \varphi(x, A) \) is a new expression \( \varphi'(x', A) \) so that there exists a substitution \( \varphi : x' \rightarrow x \) that transforms the new into the original expression: \( \varphi'(\varphi(x', A)) = \varphi(x, A) \). The probability \( P[\varphi'] \) is evaluated by assigning each new variable \( x'_i \in x' \) independently a new probability \( p'_i \).

**Example 3.2 (Dissociation).** Take the two DNF expressions:

\[
\begin{align*}
\varphi(x) &= \varphi(x_1, x_2, x_3, x_4) = x_1 x_3 \lor x_1 x_4 \lor x_2 x_4 \\
\varphi'(x') &= \varphi'(x_1, x_2, x_3, x'_4, x''_4) = x_1 x_3 \lor x_1 x'_4 \lor x_2 x''_4 .
\end{align*}
\]

Then \( \varphi'(x') \) is a dissociation of \( \varphi(x) \), as \( \varphi(x) = \varphi'(\varphi(x')) \) for the substitution \( \theta = [(x_1, x_2), (x_3, x_4), (x'_4, x_4), (x''_4, x_4)] \). Furthermore, assigning \( x'_4 \) and \( x''_4 \) the same probability as \( x_4 \) (i.e. \( p'_4 = p''_4 = p_4 \)) makes \( P[\varphi'(x')] \) an upper bound to \( P[\varphi(x)] \).

This follows from \( P[\varphi(x)] = p_1 p_2 + p_1 p_4 + p_2 p_4 - p_1 p_2 p_4 - p_1 p_2 p_4 \) and \( p_1 p_2 p_4 \), whereas \( p_1 p_2 p_4 \) is in the sum \( P[\varphi'(x')] \). Thus \( P[\varphi'(x')] = p_1 p_2 p_4 - p_1 p_2 p_4 \) and thus \( P[\varphi(x)] = (P[\varphi(x)] - P[\varphi(x)]) | (p_1 p_2 p_4 - p_1 p_2 p_4)(p_4 - 1) \geq 0 .

Next consider the two DNF expressions

\[
\psi(y) = \psi(x_1, x_3, x_4) = x_1 x_3 \lor x_1 x_4 \lor x_1 x_4 \\
\psi'(y') = \psi'(x_1, x_3, x'_4, x''_4) = x_1 x_3 \lor x_1 x'_4 \lor x_1 x''_4 .
\]

Then \( \psi'(y') \) is a dissociation of \( \psi(y) \), as \( \psi(y) = \psi'(\varphi(y')) \) for the substitution \( \theta = [(x_1, x_1), (x_3, x_3), (x'_4, x_4), (x''_4, x_4)] \). Assigning again \( p'_4 = p''_4 = p_4 \), gives both expressions the same probability \( P[\psi'(y')] \).

Both of above dissociations follow a more general template

\[
\omega(z, A) = \omega(x_1, [A_0, A_1, A_2]) = A_0 \lor A_1 x_4 \lor A_2 x_4
\]

\[
\omega'(z', A) = \omega'(x'_4, [A_0, A_1, A_2]) = A_0 \lor A_1 x'_4 \lor A_2 x''_4
\]

with \( A_0 \) representing the following composed events: \( A_0 = x_1 \) and \( A_1 = x_1 \) for both, \( A_2 = x_3 \lor x_4 \), and \( A_2 = \bar{x}_1 \) for \( \varphi \). \( \omega'(y') \) is a dissociation of \( \psi(y) \), as \( \omega'(y') \) is the substitution \( \theta = [(x'_4, x_1), (x''_4, x_1)] \). As we show in this paper, the probability of the dissociation \( P[\omega'(z', A)] \) is always an upper bound to \( P[\omega(z, A)] \) irrespective of what expressions are substituted for \( A \), and as long as they are independent of \( x_4 \), and as long as \( p'_4 = p''_4 = p_4 \). Also, for some expressions \( A \), those bounds actually become tight (whenever \( A_1 \) and \( A_2 \) are identical). Furthermore, we cannot find values for \( p'_4 \) and \( p''_4 \) which give better bounds for all possible \( A \). We call such bounds statically-tight.

**Example 3.2** informally introduces the idea of statically-tight bounds for dissociated expressions. Intuitively, we are interested in bounding the probability of an event expression \( \psi(x, A) \) with another event expression \( \psi'(x', A) \), where \( \varphi \) and \( \psi \) use the same complex events \( A \) with unknown probabilities and correlations, but different primitive events \( x \) and \( x' \) with specified probabilities. In particular, we are interested in “the best” probability assignments \( p' \) to \( x' \) of \( \varphi \) (i.e. those values that give the tightest bounds) without knowing the probabilities of and correlations between events \( A \). We call such bounds statically-tight and define them as follows for the upper case:

**Definition 3.3 (Upper bound).** Given an event expression \( \psi(x, A) \) with \( p = P[x] \). Another event expression \( \psi'(x', A) \) with \( p' = P[x'] \) is an upper bound of \( \psi \) if \( \forall A : P[\psi(x, A)] \leq P[\psi(x', A)] \).

**Definition 3.4 (Statically-tight upper bound).** An upper bound becomes statically-tight if:

(i) \( \exists A : P[\psi(x, A)] = P[\psi(x', A)] \), where \( A \) must not be trivial, i.e. \( P[A] \neq 0 \) and \( P[A] \neq 1 \).
Conjunctive Dissociation

Theorem 4.2

The disjunctive expression $P\phi_1 \geq P\phi$ can be upper and lower bounded by the probability of $P\phi_1$.

Proof. By splitting on event $C$ (Eq. 3), we get

$P\phi_1 = P\phi_1 C \vee P\phi_1 \bar{C} + P\bar{C}$

$P\phi_1 = P\phi_1 A \vee P\phi_1 B \vee P\phi_1 C$.

Figure 1: Symmetric probability assignments for a disjunctive or a conjunctive dissociation to become a statically-tight upper or lower bound.

4. Statically-tight Bounds for Dissociated Expressions

This section states the main results of this paper.

Theorem 4.1 (Disjunctive Dissociation). Let $x$ be a Boolean random variable with probability $p$ and $A_0, \ldots, A_n$ arbitrary Boolean events independent of $x$. Then the probability $P\phi_d$ of the disjunctive expression

$\phi_d = A_0 \vee xA_1 \vee xA_2 \vee \ldots \vee xA_n$

can be upper and lower bounded by the probability $P\phi_d$ of its dissociation

$\phi_d^c = A_0 \vee x_1A_1 \vee x_2A_2 \vee \ldots \vee x_nA_n$

with $x_i$ as new independent random variables, and

(a) $p_i \geq p$ for upper bounds; and
(b) $p_i \leq p$, s.t. $\prod_i(1-p_i) \geq 1 - p$ for lower bounds.

Furthermore, $P\phi_d^c$ becomes a statically-tight upper bound for $p_i \geq p$, and a statically-tight lower bound for $p_i \leq p$, s.t. $\prod_i(1-p_i) = 1 - p$. Requiring all $p_i$ to be the same, the symmetric statically-tight lower bound results from $p_i = 1 - \sqrt{1-p}$.

Theorem 4.2 (Conjunctive Dissociation). Let $x$ be a Boolean random variable with probability $p$ and $A_0, \ldots, A_n$ arbitrary

1Note that this distinction is analogous to query-centric vs. data-centric rewriting of queries in databases. Query-centric algorithms only consider the syntactic query expression, whereas data-centric algorithms can also make use of particularities in the data instance. The query-centric technique that motivates this paper is dissociation for probabilistic databases [6], which approximates the probability scores of queries over probabilistic without looking at the data-instance. Analogously, statically-tight bounds give the best bounds that can be given without evaluating the actual expressions represented by $A_i$.  

Boolean events independent of $x$. Then the probability $P\phi_d$ of the conjunctive expression

$\phi_c = A_0 \land (x \vee A_1) \land (x \vee A_2) \land \ldots \land (x \vee A_n)$

can be upper and lower bounded by the probability $P\phi_c$ of its dissociation

$\phi_c = A_0 \land x_1 \land A_1 \land x_2 \land A_2 \land \ldots \land x_n \land A_n$

with $x_i$ as new independent random variables, and

(a) $p_i \geq p$, s.t. $\prod_i(1-p_i) \geq 1 - p$ for upper bounds; and
(b) $p_i \leq p$, s.t. $\prod_i(1-p_i) = 1 - p$ for lower bounds.

Furthermore, $P\phi_c$ becomes a statically-tight upper bound for $p_i \geq p$, s.t. $\prod_i(1-p_i) \geq 1 - p$ and a statically-tight lower bound for $p_i = p$. Requiring all $p_i$ to be the same, the symmetric statically-tight upper bound results from $p_i = \sqrt{1-p}$.
Hence, the proof follows from comparing the probabilities of
\[\psi_a = xD \lor xE \]
\[\psi_a' = x_1D \lor x_2E \]
with \(D := AC\) and \(E := BC\) as new events. From disjunctive inclusion-exclusion (Eq. 1), we get
\[
P[\psi_a] = P[xD] + P[xE] - P[xDE]
\]
\[
P[\psi_a'] = pP[D] + pP[E] - pP[DE]
\]
\[
P[\psi_a''] = P[x_1D] + P[x_2E] - P[x_1x_2DE]
\]
\[
\Delta = (p_1 - p_1p_2)P[D] - p_1p_2P[DE] \geq 0.
\]
This is guaranteed to hold if \(p_1 \geq p\) and \(p_1p_2 \geq p\). Since \(\Delta\) is monotone in \(p_1\), the smallest such bound is given for \(p_1 = p\).
Furthermore, \(\Delta\) is bound in tight that events \(A\) and \(B\) are disjoint and thus \(P[DE] = P[ABC] = 0\). Since we assume lack of knowledge of the probabilities of and the correlations between \(A\) and \(B\), the latter bound is statically-tight.

Proposition 5.2 (Upper Conjunctive dissociation). Let \(x\) be a random variable with probability \(p\), and \(A, B, C\) be Boolean events independent of \(x\). The probability of
\[\phi_c = (x \lor A) \land (x \lor B) \land C\]
can then be lower bounded by the probability of
\[\phi_c' = (x_1 \lor A) \land (x_2 \lor B) \land C\]
with \(x_i\) as new random variables and choosing \(p_i \geq p\), s.t. \(p_1p_2 \geq p\). Furthermore, choosing \(p_1 \geq p\), s.t. \(p_1p_2 = p\) results in a statistically-tight upper bound (e.g. by setting \(p_1 = \sqrt{p}\)).

Proof. By splitting on event \(C\) (Eq. 4), we get
\[
P[\phi_c] = P[(x \lor A) \land (x \lor B) \land C] + P[C] - 1
\]
\[
P[\phi_c'] = P[(x_1 \lor A) \land (x_2 \lor B) \land C] + P[C] - 1.
\]
Hence, the proof follows from comparing the probabilities of
\[\psi_c = (x \lor D) \land (x \lor E) \]
\[\psi_c' = (x_1 \lor D) \land (x_2 \lor E) \]
with \(D := A \lor C\) and \(E := B \lor C\) as new events. From disjunctive inclusion-exclusion (Eq. 2) and subsequent event splitting (Eq. 3) on \(D, E\), and \(D \lor E\), we get
\[
P[\psi_c] = P[xD] + P[x \lor E] - P[x \lor D \lor E]
\]
\[
P[\psi_c'] = P[x_1D] + P[x_2E] - P[x_1 \lor x_2 \lor D \lor E]
\]
\[
\Delta = (p_1 - p_1p_2)(P[D] - P[DE]) + (p_1p_2 - p)P[DE] \geq 0.
\]
Hence, \(P[\psi_c'] \geq P[\psi_c]\), and also \(P[\phi_c''] \geq P[\phi_c']\) iff

Proposition 5.3 (Lower Disjunctive Dissociation). Let \(x\) be a random variable with probability \(p\), and \(A, B, C\) be Boolean events independent of \(x\). The probability of
\[\phi_d = xA \lor xB \lor C\]
can then be lower bounded by the probability of
\[\phi_d' = (x_1 \lor A) \lor (x_2 \lor B) \lor C\]
with \(x_i\) as new random variables and choosing \(p_i \leq p\), s.t. \((1 - p_1)(1 - p_2) \geq 1 - p\). Furthermore, choosing \(p_i \leq p\), s.t. \((1 - p_1)(1 - p_2) = 1 - p\) results in a statically-tight lower bound (e.g. by setting \(p_1 = 1 - \sqrt{1 - p}\)).

Proof. Using De Morgan’s laws, we can write the complements of \(\phi_d\) and \(\phi_d'\) as:
\[\tilde{\phi}_d = (\bar{x} \land \bar{A}) \land (\bar{x} \land \bar{B}) \land \bar{C}\]
\[\tilde{\phi}_d' = (\bar{x}_1 \land \bar{A}) \land (\bar{x}_2 \land \bar{B}) \land \bar{C}\]
Treating the complements \(\tilde{x}, \tilde{x}_1, \tilde{x}_2, \bar{A}, \bar{B}, \bar{C}, \bar{D} = \bar{A} \lor C, \bar{E} = \bar{B} \lor C\) as new events, and applying [Prop. 5.2] we know that \(P[\tilde{\phi}_d] \geq P[\tilde{\phi}_d']\) if \(\tilde{p}_i \geq \tilde{p}\) and \(\tilde{p}_1\tilde{p}_2 \geq \tilde{p}\). Hence, it follows that \(P[\phi_d''] \leq P[\phi_d]\) if \(p_1 \leq p\) and \((1 - p_1)(1 - p_2) \geq 1 - p\), and that the largest such bounds result from setting \(p_1 \leq p\), s.t. \((1 - p_1)(1 - p_2) = 1 - p\). From the proof of [Prop. 5.2] it further follows that these latter bounds become tight if \(A\) and \(B\) are identical, and hence \(P[D] = P[E] = P[DE]\). Since we assume lack of knowledge of the probabilities of and the correlations between \(A\) and \(B\), these latter bounds are statically-tight. The symmetric statically-tight bound results from setting \(p_1 = p_2\), and hence \(p_1 = \sqrt{p}\).

Proposition 5.4 (Lower Conjunctive Dissociation). Let \(x\) be a random variable with probability \(p\), and \(A, B, C\) be Boolean events independent of \(x\). The probability of
\[\phi_c = (x \lor A) \land (x \lor B) \land C\]
can then be lower bounded by the probability of
\[\phi_c' = (x_1 \lor A) \land (x_2 \lor B) \land C\]
with \(x_i\) as new random variables with \(p_i \leq p\). Furthermore, \(p_i = p\) is a statically tight lower bound.
Proof. Using De Morgan’s laws, we can write the complements of \( \phi_c \) and \( \phi_d \) as:

\[
\bar{\phi_c} = \bar{x}A \lor \bar{x}B \lor \bar{C}
\]

\[
\bar{\phi_d} = \bar{x}C \lor \bar{x}\bar{A} \lor \bar{x}\bar{B}
\]

Treating the complements \( \bar{x}, \bar{x}_1, \bar{x}_2, \bar{A}, \bar{B}, \bar{C}, \bar{D} = \bar{AC} \), and \( \bar{E} = \bar{BC} \) as new events, and applying [Prop. 5.1], we know that \( P(\bar{\phi_c}) \geq P(\bar{\phi_d}) \) if \( p_1 \geq \bar{p} \). Hence, it follows that \( P(\phi_d) \leq P(\phi_c) \) if \( p_1 \leq p \), and that the largest such bound results from \( p_1 = \bar{p} \). From the proof of [Prop. 5.1], it further follows that this bound is tight if \( \bar{A} \) and \( B \) are disjoint (i.e. \( P[A \lor B = 1] \)) and hence \( P(\bar{D}E) = P(\bar{A}\bar{B}C) = 0 \). Since we assume lack of knowledge of the probabilities of and correlations between \( D \) and \( E \), this bound is statically-tight.

\[ \square \]

6. Illustration

We illustrate our symmetric statically-tight bounds for the disjunctive expression \( \phi_d = xA \lor xB \) and the conjunctive expression \( \phi_c = (x \lor A)(x \lor B) \). Let \( p = P[x], q = P[A] = P[B], \) and assume \( x \) to be independent of \( A \) and \( B \), which can be arbitrarily correlated: \( -1 \leq \rho(A, B) \leq 1 \). Further, let \( p_1 = P[x1] = P[x2] \) be the probabilities in the dissociated expressions.

In a few steps, one can calculate the probabilities of \( \phi_d, \phi_c \) and their dissociations as

\[
P(\phi_d) = 2pq - pP[AB]
\]

\[
P(\phi_c) = 2pq + p^2P[AB]
\]

\[
P(\phi_d) = p + (1 - p)P[AB]
\]

\[
P(\phi_c) = 2pq + p^2(1 - 2q) + (1 - p^2)P[AB]
\]

Figure 2 illustrates the probabilities of the expressions and their symmetric statically-tight upper and lower dissociations for various values of \( p, q \) and as function of the correlation \( \rho(A, B) \), and by setting \( p_1 \) according to Fig. 1. Remember that \( p(\phi_d, \phi_c) = P[AB]^{\frac{1}{2}} \) and, hence: \( P[AB] = \rho(A, B) \cdot (q - q^2) + q^2 \). Further, \( P(\bar{A}B) = 0 \) (i.e. disjointness between \( A \) and \( B \)) is not possible for \( q > 0.5 \), and from \( P[A \lor B = 1] \), one can derive \( P[AB] \geq 2p - 1 \). In turn, \( \rho = -1 \) is not possible for \( q < 0.5 \), and it must hold \( P[AB] \geq 0 \). From both together, one can derive the condition \( \rho_{\text{min}}(q) = \max(-\frac{q}{1 - q}, -\frac{1}{q^2 - 2q}) \). This marks the beginning of the graphs for different values of \( q \) in the figure.

Acknowledgements

We like to thank Adnan Darwiche for helpful discussions. This work was partially supported by NSF grant IIS-0915054, and was performed in the context of the LaPushDB project \( \text{http://LaPushDB.com/} \).

References

[1] A. Choi, M. Chavira, A. Darwiche, Node splitting: A scheme for generating upper bounds in bayesian networks, in: UAI (Annual Conference on Uncertainty in Artificial Intelligence), 2007, pp. 57–66.

[2] R. Dechter, I. Rish, Mini-buckets: A general scheme for bounded inference, J. ACM 50 (2) (2003) 107–135.

[3] M. Ramírez, H. Geffner, Structural relaxations by variable renaming and their compilation for solving MinCostSAT, in: CP (Principles and Practice of Constraint Programming), 2007, pp. 605–619.

[4] K. Pipatsrisawat, A. Darwiche, Clone: Solving weighted Max-SAT in a reduced search space, in: Australian Conference on Artificial Intelligence, 2007, pp. 223–233.

[5] H. R. Andersen, T. Hadzic, J. N. Hooker, P. Tiedemann, A constraint store based on multivalued decision diagrams, in: CP (Principles and Practice of Constraint Programming), 2007, pp. 118–132.

[6] W. Gatterbauer, A. K. Jha, D.Suciu, Dissociation and propagation for efficient query evaluation over probabilistic databases, in: MUD (VLDB Workshop on Management of Uncertain Data), 2010.

[7] Y. Cráma, P. L. Hammer, Boolean Functions: Theory, Algorithms, and Applications, Cambridge University Press, 2011 (preprint 2009).

[8] N. Fuhr, T. Rölleke, A probabilistic relational algebra for the integration of information retrieval and database systems, ACM Trans. Inf. Syst. 15 (1) (1997) 32–66.

[9] W. Feller, An introduction to probability theory and its applications, 3rd Edition, Wiley, New York, 1968.

[10] E. Lehman, F. T. Leighton, A. R. Meyer, Mathematics for Computer Science, 2010.

[11] D. Poole, Probabilistic horn abduction and bayesian networks, Artif. Intell. 64 (1) (1993) 81–129.

Appendix A. Representing Complex Events

It is known from Poole’s independent choice logic [11] that arbitrary correlations between events can be composed from disjoint-independent events only. A disjoint-independent event is represented by a non-Boolean independent random variable \( x \) which takes either of \( k \) values \( v = (v_1, \ldots, v_k) \) with respective
probabilities $p = (p_1, \ldots, p_k)$ and $\sum_i p_i = 1$. Poole writes such a "disjoint declaration" as $x_0 = x_1 + \cdots + x_k$. 

In turn, any $k$ disjoint events can be represented starting from $k - 1$ independent Boolean variables $y = (y_1, \ldots, y_{k-1})$ and probabilities $P[y] = (p_1, p_2, \ldots, p_k)$, by assigning the disjoint-independent event variable $x$ its value $v_i$ whenever event $A_i$ is true with $A_i$ defined as:

$$(x = v_i) \iff A_i := x_1$$

$$(x = v_2) \iff A_2 := \bar{x}_1 x_2$$

$$\vdots$$

$$(x = v_{k-1}) \iff A_{k-1} := x_{1} \cdots \bar{x}_{k-2} x_{k-1}$$

$$(x = v_k) \iff A_k := \bar{x}_1 \cdots \bar{x}_{k-2} \bar{x}_{k-1}$$

For example, a primitive disjoint-independent event variable $x(v_1: \frac{1}{2}, v_1: \frac{1}{2}, v_1: \frac{1}{2}, v_1: \frac{1}{2})$ can be represented with three independent Boolean variables $y = (y_1, y_2, y_3)$ and $P[y] = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$.

It follows that arbitrary correlations between events can be modeled starting from independent Boolean random variables alone. For example, two complex events $A$ and $B$ with $P[A] = P[B] = q$ and varying correlation (see Sect. 6) can be represented as composed events $A := y_1 y_2 \lor y_3 \lor y_4$ and $B := y_1 y_2 \lor y_3 \lor y_5$ over the primitive events $y$ with varying probabilities $P[y]$. Events $A$ and $B$ become identical for $P[y] = (0, q, q, 0, q)$, independent for $P[y] = (0, q, q, q, 0)$, and disjoint for $P[y] = (0, 0, 0, q, q)$ with $q \leq 0.5$.

Appendix B. Proofs of Main Theorems

**Proof [Theorem 4.2 (a)]** The proof follows from [Prop. 5.2] and induction on the number of new variables for each original disjointed variable. Assume we know that $P[\phi_d'] \geq P[\phi_d]$ for

$$\phi_d = A_0 \lor x_1 A_1 \lor \cdots \lor x_k A_k = A_0 \lor x(A_1 \lor \cdots \lor A_k)$$

$$\phi_d' = A_0 \lor x_1 A_1 \lor \cdots \lor x_{k-1} A_{k-1} \lor x_0 A_k \lor \cdots \lor A_n$$

with $p_0 \geq p_i$, $i \in [k]$ and $k < n$. We need to show that then $P[\phi_d^{k+1}] \geq P[\phi_d]$, with $p_0 \geq p_i$, $i \in [k+1]$ for

$$\phi_d^{k+1} = A_0 \lor x_1 A_1 \lor \cdots \lor x_k A_k \lor x_{k+1} A_{k+1} \lor \cdots \lor A_n$$

Applying [Prop. 5.1] to $\phi_d'$, we know $P[\phi_d^{k+1}] \geq P[\phi_d]$, with $p_0' \geq p_i \geq p$ and $p_0'' \geq p \geq p_0$. Hence, we have shown $P[\phi_d^{k+1}] \geq P[\phi_d'] \geq P[\phi_d]$ after substituting $p_0 \leftarrow p_0'$ and $p_0+1 \leftarrow p_0''$. Since $\phi_d'$ is monotone in $p_1$, it follows that the smallest such upper bound results from choosing $p_1 = p$. Furthermore, this bound is tight in case the events $A_i$, $i \in [n]$ are disjoint, since then $P[\phi_d^{k+1}] = P[\phi_d] + pP[A_0A_1] + \cdots + pP[A_0A_n] = P[\phi_d]$. Since we assume lack of knowledge of the probabilities and of correlations between events $A_i$, the bound is statically tight.

**Proof [Theorem 4.1 (a)]** Using De Morgan’s laws, we can write the complements of $\phi_d$ and $\phi_d'$ as:

$$\bar{\phi}_d = \bar{A_0} \land (\bar{x} \lor \bar{A}_1) \land \cdots \land (\bar{x} \lor \bar{A}_n)$$

$$\bar{\phi}_d' = \bar{A}_0 \lor \bar{x}_1 \lor \bar{A}_1 \land \cdots \land \bar{x}_{k-2} \lor \bar{A}_{k-1} \lor \bar{x}_{k-1} \lor \bar{A}_k$$

From [Theorem 4.2 (a)] we know that $P[\bar{\phi}_d'] \geq P[\bar{\phi}_d]$ if $\bar{p}_i \geq \bar{p}$, s.t. $P[\bar{\phi}_d] \geq P[\bar{\phi}_d']$ if $\bar{p} \leq p$. Hence, it follows that $P[\bar{\phi}_d'] \leq P[\bar{\phi}_d]$ if $\bar{p} \leq \bar{p}$, s.t. $P[\bar{\phi}_d] \leq P[\bar{\phi}_d']$ if $\bar{p} \geq \bar{p}$. Furthermore, these upper bounds are tight in case $A_i, i \in [n]$ are identical, since then $P[\bar{\phi}_d'] = P[\bar{x}_1 \land \cdots \land \bar{x}_n A_1 \lor \bar{A}_1] + P[\bar{A}_0] = 1 = P[\bar{x}_1 \lor \bar{A}_1] + P[\bar{A}_0] = 1 = P[\bar{\phi}_d]$. Since we assume lack of knowledge of the probabilities of and correlations between events $A_i$, these latter bounds are statically-tight. The symmetric statically-tight lower bound results from additionally requiring $p_1 = p_j$, and thus setting $p_i = \sqrt{p}$. 

**Proof [Theorem 4.2 (b)]** Using De Morgan’s laws, we can write the complements of $\phi_d$ and $\phi_d'$ as:

$$\bar{\phi}_d' = \bar{A}_0 \lor \bar{x}_1 \land \cdots \land \bar{x}_k \land \bar{A}_k$$

From [Theorem 4.1 (a)] we know that $P[\bar{\phi}_d'] = P[\bar{\phi}_d]$ if $\bar{p}_i \geq \bar{p}$, s.t. $P[\bar{\phi}_d] \geq P[\bar{\phi}_d']$ if $\bar{p} \leq \bar{p}$. Furthermore, this latter bound is tight in case $A_i, i \in [n]$ are disjoint (this is true if $P[A_1 \land A_2] = 1$, since then $P[\bar{\phi}_d'] = P[\bar{A}_0] + (1 - \bar{p})P[\bar{A}_1] + \cdots + (1 - \bar{p})P[\bar{A}_n] = P[\bar{\phi}_d]$. Since we assume lack of knowledge of the probabilities of and correlations between events $A_i$, this bound is statically-tight.
**Appendix C. Nomenclature**

| Symbol(s) | Description |
|-----------|-------------|
| $x, y, z$ | primitive events: independent Boolean random variables |
| $\phi, \psi, \omega$ | composed events: Boolean event expressions |
| $A, B, \ldots$ | complex events: events composed from primitive events with unknown event expressions |
| $P[A], P[\phi]$ | probability of an event or expression |
| $p_i$ | probability $P[x_i]$ of random variable $x_i$ |
| $x, A, p$ | ordered sets $(x_1, \ldots, x_k)$ or unordered sets $\{x_1, \ldots, x_k\}$ of variables, events or probabilities |
| $\bar{x}, \bar{A}, \bar{\phi}, \bar{p}$ | complements: $\neg x, \neg A, \neg \phi, 1 - p$ |