Kapitel 2: Die AGM-Schranke und andere obere Schranken für die Größe von Anfrageergebnissen

Der Begriff "AGM-Schranke" ist benannt nach den Autoren Atserias, Grohe und Marx des Artikels "Size bounds and query plans for relational joins" In Proc. FoCS 2008, pp. 739-748.

2.1 Warm-Up: Die Δ-Anfrage

Betrachte die Anfrage Q_Δ aus Beispiel 1.2

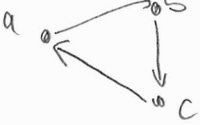
$$Q_\Delta(A,B,C) \leftarrow E(A,B), E(B,C), E(C,A).$$

Sei $\sigma := \sigma_E = \{E\}$. Jede σ-DB $\mathcal{D}$ entspricht einem gerichteten Graphen $G^{\mathcal{D}} = (V^{\mathcal{D}}, E^{\mathcal{D}})$ mit $V^{\mathcal{D}} := adom(\mathcal{D})$.

Klar: Für jede σ-DB $\mathcal{D}$ ist

$$Q_\Delta(\mathcal{D}) = \{ (a,b,c) : (a,b) \in E^{\mathcal{D}} \text{ und } (b,c) \in E^{\mathcal{D}} \text{ und } (c,a) \in E^{\mathcal{D}} \}$$

Skizze:



D.h: $Q_\Delta(\mathcal{D})$ besteht aus allen gerichteten Dreiecken in $G^{\mathcal{D}}$.

Ziel: Gib eine obere Schranke für die Anzahl $|Q_\Delta(\mathcal{D})|$ von Tupeln in $Q_\Delta(\mathcal{D})$ an und finde einen möglichst effizienten Algorithmus, der $Q_\Delta(\mathcal{D})$ berechnet.

Notation: $n^D := |adom(D)|$

$N^D := |E^D|$

Offensichtliche obere Schranken für $|Q_\Delta(D)|$:

1) $|Q_\Delta(D)| \leq (n^D)^3$

2) $|Q_\Delta(D)| \leq (N^D)^2$

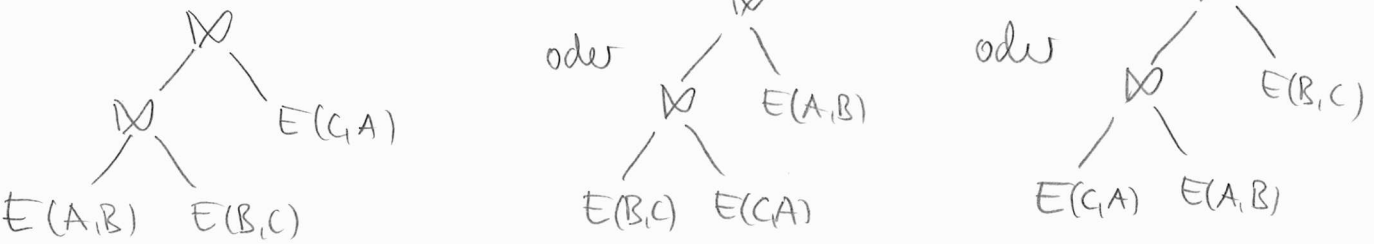
Frage: Geht das besser?

Antwort, die der nächste Satz (Satz Δ) gibt:

Ja: $|Q_\Delta(D)| \leq 2 \cdot (N^D)^{1,5}$ — und es gibt auch einen Algorithmus, der $Q_\Delta(D)$ mit Laufzeit $\approx (N^D)^{1,5}$ löst.

Bevor wir uns den Satz (inkl. Beweis) anschauen, diskutieren wir aber erstmal, wie ein herkömmliches Datenbanksystem die Anfrage Q_Δ auswertet (siehe Beispiel 1.1 für eine Formulierung der Anfrage in SQL).

Um die Anfrage Q_Δ auszuwerten wird ein Datenbanksystem i.d.R. QEPs (query evaluation plans) der folgenden Art betrachten



Bei jedem dieser QEPs wird als Zwischenergebnis $Q_2(D)$ berechnet, für

$$Q_2(X,Y,Z) \leftarrow E(X,Y), E(Y,Z)$$

und dann wird

$$Q_\Delta(X,Y,Z) \leftarrow Q_2(X,Y,Z), E(Z,X)$$

auf D ausgewertet.

Beispiel 2.1

Für jede Zahl $m \geq 1$ gibt es eine 0-DB D_m mit

$N^{D_m} = \Theta(m)$, $|Q_\Delta(D_m)| = \Theta(m)$, aber

$|Q_2(D_m)| = \Omega(m^2)$ für $Q_2(X,Y,Z) \leftarrow E(X,Y), E(Y,Z)$.

Die "üblichen" QEPs (query evaluation plans), die in Datenbanksystemen zur Auswertung der Anfrage Q_Δ erzeugt werden, benötigen zum Auswerten von Q_Δ auf D_m daher Zeit $\Omega((N^{D_m})^2)$, während der Algorithmus aus dem folgenden Satz Δ nur Zeit $O((N^{D_m})^{1,5})$ benötigt.

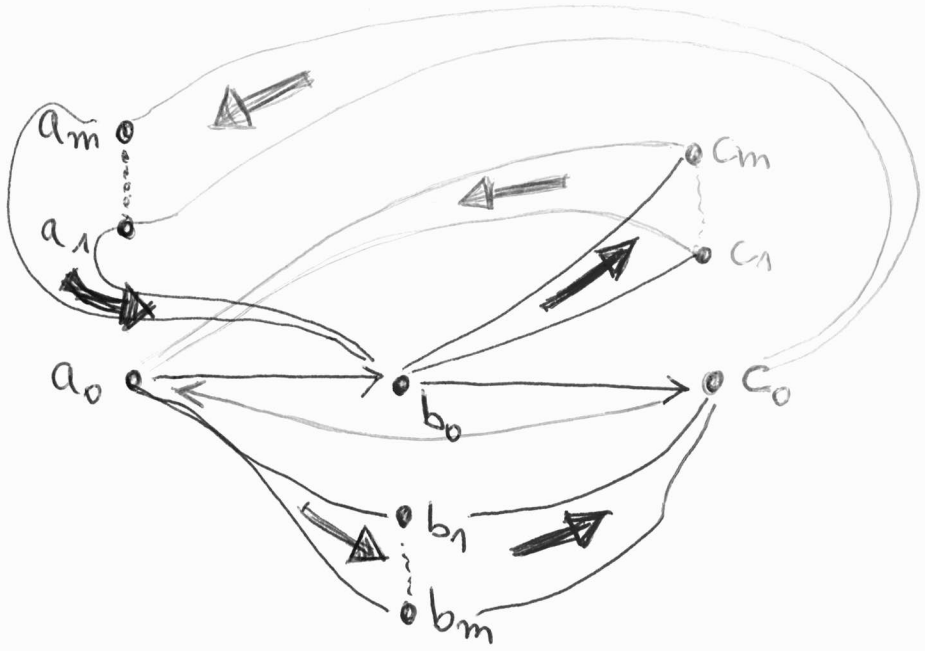
Konstruktion von D_m:

Seien $a_0, a_1, \ldots, a_m$, $b_0, b_1, \ldots, b_m$, $c_0, c_1, \ldots, c_m$ $3(m+1)$ verschiedene Elemente aus dom. Sei

$$
\begin{aligned}
E^{D_m} := \; & \{ (a_0, b_i) : i \in [0,m] \} \cup \\
& \{ (a_i, b_0) : i \in [1,m] \} \cup \\
& \{ (b_0, c_i) : i \in [0,m] \} \cup \\
& \{ (b_i, c_0) : i \in [1,m] \} \cup \\
& \{ (c_0, a_i) : i \in [0,m] \} \cup \\
& \{ (c_i, a_0) : i \in [1,m] \}
\end{aligned}
$$

Klar: $N^{D_m} := |E^{D_m}| = 3 \cdot (m+1+m) = 6m+3$

Skizze: D_m



Man kann sich leicht davon überzeugen, dass gilt:

$$Q_\Delta(D_m) = X \cup \pi_{2,3,1}(X) \cup \pi_{3,1,2}(X) \text{ für}$$

$$X = \{(a_0, b_0, c_i) : i \in [0,m]\} \cup \{(a_0, b_i, c_0) : i \in [1,m]\} \cup \{(a_i, b_0, c_0) : i \in [1,m]\}.$$

Es gilt: $|X| = m+1+2m = 3m+1$ und $|Q_\Delta(D_m)| = 3|X| = 9m+3$

Außerdem ist $(a_i, b_0, c_j) \in Q_2(D_m)$, f.a. $i,j \in [0,m]$,

also $|Q_2(D_m)| > m^2$

Satz Δ: Für jede σ-DB D gilt:

$$|Q_\Delta(D)| \leq 2 \cdot (N^D)^{3/2}$$

und es gibt einen Algorithmus, der bei Eingabe von D die Menge $Q_\Delta(D)$ in Zeit $O((N^D)^{3/2} \cdot t)$ berechnet, wobei t die Zeit ist, die wir benötigen, um bei Eingabe von (v,w) zu testen, ob $(v,w) \in E^D$ ist.

Beweis: Sei D eine beliebige σ-DB.

Für jedes $v \in adom(D)$ sei

$$aus\text{-}Grad^D(v) := |\{w : (v,w) \in E^D\}|$$
$$= |E^D(v,*)|$$

für $E^D(v,*) := \{w : (v,w) \in E^D\}|$.

Ein Knoten $v \in adom(D)$ heißt

- **schwer**, wenn $aus\text{-}Grad^D(v) \geq \sqrt{N^D}$ ist
- **leicht**, wenn v nicht schwer ist.

Ein Tupel $t = (a,b,c) \in Q_\Delta(D)$ heißt
schwer, wenn a schwer ist, und es heißt
leicht, wenn a leicht ist.

Für $X \in \{\text{leicht}, \text{schwer}\}$ sei

$$Q_{\Delta,X}(\mathbb{D}) := \{ t \in Q_\Delta(\mathbb{D}) : t \text{ ist } X \}.$$

Klar: $Q_\Delta(\mathbb{D}) = Q_{\Delta,\text{leicht}}(\mathbb{D}) \cup Q_{\Delta,\text{schwer}}(\mathbb{D})$

und $|Q_\Delta(\mathbb{D})| = |Q_{\Delta,\text{leicht}}(\mathbb{D})| + |Q_{\Delta,\text{schwer}}(\mathbb{D})|$.

Behauptung 1: $|Q_{\Delta,\text{leicht}}(\mathbb{D})| \leq (N^{\mathbb{D}})^{3/2}$

Beweis: Für jeden leichten Knoten $a \in \text{adom}(\mathbb{D})$ ist $\text{aus-Grad}^{\mathbb{D}}(a) < \sqrt{N^{\mathbb{D}}}$, und es gilt:

$$\begin{aligned}
&|\{(b,c) : (a,b,c) \in Q_\Delta(\mathbb{D})\}| \\
&\leq |\{(b,c) : (a,b) \in E^{\mathbb{D}} \text{ und } (c,a) \in E^{\mathbb{D}}\}| \\
&\leq \text{aus-Grad}^{\mathbb{D}}(a) \cdot \text{ein-Grad}^{\mathbb{D}}(a), \quad \text{wobei} \\
&\text{ein-Grad}^{\mathbb{D}}(a) := |\{w : (w,a) \in E^{\mathbb{D}}\}| = |E^{\mathbb{D}}(*,a)| \\
&\text{für } E^{\mathbb{D}}(*,a) := \{w : (w,a) \in E^{\mathbb{D}}\}.
\end{aligned}$$

Insgesamt gilt:

$$\begin{aligned}
|Q_{\Delta,\text{leicht}}(\mathbb{D})| &\leq \sum_{\substack{a \in \text{adom}(\mathbb{D}),\\ a \text{ leicht}}} |\{(b,c) : (a,b,c) \in Q_\Delta(\mathbb{D})\}| \\
&\leq \sum_{\substack{a \in \text{adom}(\mathbb{D}),\\ a \text{ leicht}}} \text{aus-Grad}^{\mathbb{D}}(a) \cdot \text{ein-Grad}^{\mathbb{D}}(a) \\
&\leq \sqrt{N^{\mathbb{D}}} \cdot \sum_{a \in \text{adom}(\mathbb{D})} \text{ein-Grad}^{\mathbb{D}}(a) \\
&\leq \sqrt{N^{\mathbb{D}}} \cdot |E^{\mathbb{D}}| \\
&= \sqrt{N^{\mathbb{D}}} \cdot N^{\mathbb{D}} = (N^{\mathbb{D}})^{3/2}
\end{aligned}$$

□ Beh 1.

Sei S^D die Menge aller schweren Knoten von D,
d.h. $S^D := \{ v \in adom(D) : v \text{ ist schwer} \}$.

Behauptung 2: $|Q_{\Delta,\text{schwer}}(D)| \leq |S^D| \cdot N^D$

Beweis:

$$\begin{aligned} Q_{\Delta,\text{schwer}}(D) &= \{ (a,b,c) : a \in S^D \text{ und } (a,b,c) \in Q_\Delta(D) \} \\ &\subseteq \{ (a,b,c) : a \in S^D \text{ und } (b,c) \in E^D \} \\ &= S^D * E^D \end{aligned}$$

Somit: $|Q_{\Delta,\text{schwer}}(D)| \leq |S^D * E^D| = |S^D| \cdot |E^D| = |S^D| \cdot N^D$

□ Beh 2

Behauptung 3: $|S^D| \leq \sqrt{N^D}$

Beweis: Für jedes $v \in S^D$ ist $|E^D(v,*)| = \text{aus-Grad}^D(v) \geq \sqrt{N^D}$.

Es gilt:

$$\begin{aligned} N^D = |E^D| &\geq \Big| \bigcup_{v \in S^D} \{ (v,w) : w \in E^D(v,*) \} \Big| \\ &= \sum_{v \in S^D} |\{ (v,w) : w \in E^D(v,*) \}| \\ &= \sum_{v \in S^D} |E^D(v,*)| \\ &\geq \sum_{v \in S^D} \sqrt{N^D} \\ &= |S^D| \cdot \sqrt{N^D} \end{aligned}$$

Somit ist $|S^D| \leq \frac{N^D}{\sqrt{N^D}} = \sqrt{N^D}$.

□ Beh 3

Insgesamt gilt:

$$|Q_\Delta(D)| = |Q_{\Delta,\text{leicht}}(D)| + |Q_{\Delta,\text{schwer}}(D)|$$

$$\overset{\text{Beh 1\&2}}{\leq} (N^D)^{3/2} + |S^D| \cdot N^D$$

$$\overset{\text{Beh 3}}{\leq} (N^D)^{3/2} + \sqrt{N^D} \cdot N^D$$

$$= (N^D)^{3/2} + (N^D)^{3/2}$$

$$= 2 \cdot (N^D)^{3/2}.$$

Unser Algorithmus zur Berechnung von $Q_\Delta(D)$ geht bei Eingabe von D wie folgt vor:

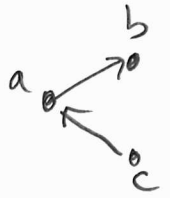
0) Initialisiere $Q_\Delta(D) := \emptyset$

1) Berechne für jedes $v \in \text{adom}(D)$ Folgendes:
die Mengen $E^D(v,*) = \{w : (v,w) \in E^D\}$ und
$E^D(*,v) = \{w : (w,v) \in E^D\}$ und
die Zahl $\text{aus-Grad}^D(v)$.
Durch geeignetes Sortieren von E^D geht das in Zeit $O(N^D \cdot \log N^D) \leq O((N^D)^{3/2})$

2) Berechne die Menge S^D aller schweren $v \in \text{adom}(D)$ und die Menge $L^D := \text{adom}(D) \setminus S^D$ aller leichten $v \in \text{adom}(D)$.
Unter Verwendung der in 1) gesammelten Infos geht das in Zeit $O(N^D)$.

3) Betrachte jedes leichte $a \in L^D$
Betrachte jedes $b \in E^D(a,*)$
Betrachte jedes $c \in E^D(*,a)$
und teste, ob $(b,c) \in E^D$.
Wenn ja, füge (a,b,c) in $Q_\Delta(D)$ ein

Skizze:



Auf die gleiche Art wie im Beweis von Beh 1 erhalten wir, dass das in Zeit $O((N^D)^{3/2} \cdot t)$ geht.

4) Betrachte jedes schwere $a \in S^D$
Betrachte jedes $(b,c) \in E^D$
und teste, ob $(a,b) \in E^D$
und
$(c,a) \in E^D$.

Skizze:

Wenn ja, füge (a,b,c) in $Q_\Delta(D)$ ein.

Das geht in Zeit $|S^D| \cdot N^D \cdot 2t = O((N^D)^{3/2} \cdot t)$.

Insgesamt berechnen wir so $Q_\Delta(D)$ in Zeit $O((N^D)^{3/2} \cdot t)$.

□ Satz Δ

Bemerkung Δ:

Der Algorithmus aus Satz Δ ist im folgenden Sinn Worst-case-optimal:

Es gibt eine Folge von σ-Datenbanken $D_1, D_2, D_3, \ldots$ so dass für $N_i := |E^{D_i}|$ für $i \in \mathbb{N}_{\geq 1}$ gilt:

$N_1 < N_2 < N_3 < \ldots$ und

$|Q_\Delta(D_i)| = (N_i)^{3/2}$.

Dazu wähle für $i \geq 1$ die DB D_i mit $E^{D_i} = [1,i] \times [1,i]$.

Dann ist $N_i = i^2$ und $Q_\Delta(D_i) = [1,i]^3$, also

$|Q_\Delta(D_i)| = i^3 = (N_i)^{3/2}$.

Literaturhinweis:

Der hier präsentierte Beweis von Satz △ ist aus der folgenden Arbeit entnommen:

"Worst-case optimal join algorithms" von Ngo, Porat, Ré, Rudra, In Proc. PODS 2012, pp. 37–48

Die Beweisidee wird dort Loomis und Whitney zugeschrieben ("An inequality related to the isoperimetric inequality" von Loomis, Whitney. In Bull. Amer. Math. Soc., 55: 961–962, 1949).

2.2 Die AGM-Schranke

Die AGM-Schranke wurde zunächst für sog. Join-Anfragen formuliert und später dann auf beliebige konjunktive Anfragen verallgemeinert.

Definition 2.2

Eine Join-Anfrage (vom Schema σ) ist eine CQ Q (vom Schema σ) der Form

$$Q(\overline{x}_0) \leftarrow R_1(\overline{x}_1), \ldots, R_m(\overline{x}_m) \qquad (*)$$

für die gilt:

1) $\text{cons}(Q) = \emptyset$,
2) im Kopf der Anfrage kommen alle Variablen vor, die im Rumpf der Anfrage vorkommen
3) $R_1, \ldots, R_m$ sind paarweise verschieden, und
4) für jedes $i \in [0,m]$ gilt:
   $\overline{x}_i$ ist ein Tupel von paarweise verschiedenen Variablen

Beispiel 2.3

- $Q_\Delta(A,B,C) \leftarrow E(A,B), E(B,C), E(C,A)$
  ist keine Join-Anfrage (da Bedingung 3) verletzt wird

- $Q'_\Delta(A,B,C) \leftarrow E_1(A,B), E_2(B,C), E_3(C,A)$
  ist eine Join-Anfrage über dem Schema
  $\sigma' := \{E_1, E_2, E_3\}$, das aus drei verschiedenen Relationssymbolen der Stelligkeit 2 besteht.

- $Q''_\Delta() \leftarrow E_1(A,B), E_2(B,C), E_3(C,A)$
  ist keine Join-Anfrage (da Bedingung 2) verletzt wird

- Die Anfrage
  $Q_{(a)}(A) \leftarrow E(A, \text{Sascha Lobo}), E(A, \text{Mesut Özil})$
  aus Beispiel 1.2(a) ist keine Join-Anfrage,
  (da sie die Bedingungen 1) und 3) verletzt).

Bemerkung 2.4

Die in Definition 2.2 eingeführten Join-Anfragen entsprechen in der "benannten Perspektive" der Situation, in der jedes R_i die Attribute $\overline{X_i}$ hat, und in der der "natürliche Join" $R_1[\overline{X_1}] \bowtie \cdots \bowtie R_m[\overline{X_m}]$ berechnet werden soll.

Die AGM-Schranke liefert eine Antwort auf die folgende Frage:

Gegeben sei eine Join-Anfrage Q der Form ⊛ und eine DB D mit $N_i := N_i^D := |R_i^D|$ f.a. $i \in [1,m]$.

Frage: Wie viele Tupel kann es im Anfrageergebnis $Q(D)$ höchstens geben?

Betrachten wir dazu zunächst einige Spezialfälle:

1) Wenn es ein $i \in [1,m]$ gibt, s.d. $\overline{x_0} = \overline{x_i}$ ist, dann ist

$$Q(D) \subseteq R_i^D, \text{ und somit } |Q(D)| \leq N_i$$

2) Wenn $\overline{x_0} = \overline{x_{i_1}} \cdot \overline{x_{i_2}}$ für $i_1, i_2 \in [1,m]$ ist, dann ist

$$Q(D) \subseteq R_{i_1}^D \times R_{i_2}^D \text{ und somit } |Q(D)| \leq N_{i_1} \cdot N_{i_2}$$

3) Wenn $\overline{x_0} = \overline{x_{i_1}}, \ldots, \overline{x_{i_k}}$ für $i_1, \ldots, i_k \in [1,m]$ ist, dann ist

$$Q(D) \subseteq R_{i_1}^D \times \ldots \times R_{i_k}^D \text{ und somit } |Q(D)| \leq \prod_{j=1}^{k} N_{i_j}$$

Letztendlich kommt es dabei nicht auf die Reihenfolge an, in der die einzelnen Variablen in einem Tupel $\overline{x_i}$ auftreten. So gilt an Stelle von 1), 2), 3) sogar

1') Wenn es ein $i \in [1,m]$ gibt, s.d. $\{\overline{x_0}\} = \{\overline{x_i}\}$, dann ist $|Q(D)| \leq N_i$.

Hierbei verwenden wir folgende Notation:
Für ein Tupel $\overline{t} = (t_1, \dots, t_\ell)$ ist
$\{\overline{t}\} := \{t_1, \dots, t_\ell\}$.
Für eine Liste von Tupeln $\overline{t}_1, \dots, \overline{t}_k$ ist
$\{\overline{t}_1, \dots, \overline{t}_k\} := \{\overline{t}_1\} \cup \dots \cup \{\overline{t}_k\}$.

2') Wenn $\{\overline{x_0}\} = \{\overline{x_{i_1}}, \overline{x_{i_2}}\}$ für $i_1, i_2 \in [1,m]$ ist, dann ist $|Q(D)| \leq N_{i_1} \cdot N_{i_2}$

3') Wenn $\{\overline{x_0}\} = \{\overline{x_{i_1}}, \dots, \overline{x_{i_k}}\}$ für $i_1, \dots, i_k \in [1,m]$ ist, dann ist $|Q(D)| \leq \prod_{j=1}^{k} N_{i_j}$

Wir formulieren nun 3') von einem anderen Blickwinkel aus:

Definition 2.4 (Anfrage-Graph G_Q)

Sei Q eine Join-Anfrage der Form

$$Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$$

Sei $n := |\{\bar{x}_1, \ldots, \bar{x}_m\}|$ die Anzahl der verschiedenen in Q vorkommenden Variablen und sei $A_1, \ldots, A_n$ eine Liste all dieser Variablen.

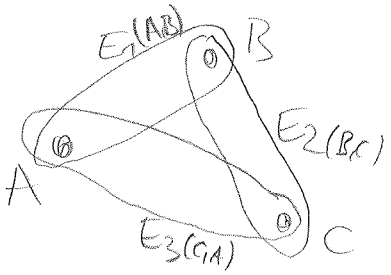
Der zu Q gehörende **Hypergraph** G_Q hat die Knotenmenge $V_Q := \{A_1, \ldots, A_n\} = vars(Q)$, und für jedes $i \in [1,m]$ hat G_Q eine Hyperkante (namens $R_i(\bar{x}_i)$), die aus den Knoten $\{\bar{x}_i\}$ besteht.

Beachte: Gemäß Bedingung 2) in Def. 2.2 gilt: $\{\bar{x}_0\} = \{A_1, \ldots, A_n\}$

Beispiel 2.5

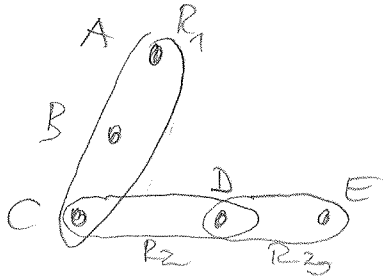
(a) $Q'_\Delta(A,B,C) \leftarrow E_1(A,B), E_2(B,C), E_3(C,A)$

$G_{Q'_\Delta}$:



(b) $Q_{(b)}(A,B,C,D,E) \leftarrow R_1(A,B,C), R_2(C,D), R_3(D,E)$

$G_{Q_{(b)}}$:



Definition 2.6 (Kantenüberdeckung)

Sei G ein beliebiger Hypergraph mit Knotenmenge V.

Eine Kantenüberdeckung (engl.: edge cover) von G ist eine Menge K von Hyperkanten von G, für die gilt: $\bigcup_{e \in K} e = V$ (d.h.: jeder Knoten von G ist in mindestens einer Hyperkante enthalten, die in K gehört).

Beispiel 2.7

$K := \{E_1(A,B), E_2(B,C)\}$ ist eine Kantenüberdeckung von $G_{Q'_\Delta}$ für die Anfrage Q'_Δ aus Bsp. 2.5(a)

$K := \{R_1(A,B,C), R_3(D,E)\}$ ist eine Kantenüberdeckung von $G_{Q_{(b)}}$ für die Anfrage $Q_{(b)}$ aus Bsp. 2.5(b)

Notation:

Wenn wir Join-Anfragen Q betrachten, sprechen wir von "Kantenüberdeckungen von Q" und meinen damit Kantenüberdeckungen von G_Q.

Die Beobachtung 3') lässt sich also wie folgt formulieren:

3'') Wenn $i_1, \dots, i_k \in [1,m]$ so gewählt sind, dass $K := \{ R_{i_1}(\overline{x}_{i_1}), \dots, R_{i_k}(\overline{x}_{i_k}) \}$ eine Kantenüberdeckung von Q ist, dann gilt: $|Q(D)| \leq \prod_{j=1}^{k} N_{i_j}$.

Eine solche Menge K können wir auch repräsentieren als eine Abbildung $x : [1,m] \to \{0,1\}$ mit

$$x(i) = \begin{cases} 1 & \text{falls } R_i(\overline{x}_i) \in K \\ 0 & \text{falls } R_i(\overline{x}_i) \notin K. \end{cases}$$

Dann repräsentiert eine Abbildung $x : [1,m] \to \{0,1\}$ genau dann eine Kantenüberdeckung von Q, wenn gilt:

Für alle $j \in [1,n]$ gilt: $\sum_{\substack{i \in [1,m] \text{ mit} \\ A_j \in \{\overline{x}_i\}}} x(i) \geq 1$ (**)

Beobachtung 3'') lässt sich also auch wie folgt formulieren:

3''') Für jede Abbildung $x : [1,m] \to \{0,1\}$, die (**) erfüllt, gilt: $|Q(D)| \leq \prod_{i=1}^{m} (N_i)^{x(i)}$

Beachte: F.a. reellen Zahlen $z \neq 0$ gilt: $z^0 = 1$ und es gilt: $0^0 = 0$.

Die AGM-Schranke besagt, dass die Ungleichung (***) sogar für jede Abbildung $x : [1,m] \longrightarrow \mathbb{Q}_{\geq 0}$, die (**) erfüllt, gilt.

Notation: $\mathbb{Q}_{\geq 0}$ bezeichnet die Menge aller nicht-negativen rationalen Zahlen.

Präzise Formulierung der AGM-Schranke:

Definition 2.8 (fraktionale Kantenüberdeckung)
Sei G ein beliebiger Hypergraph mit Knotenmenge V und Hyperkantenmenge E.
Eine fraktionale Kantenüberdeckung (kurz: fKü, engl.: fractional edge cover) von G ist eine Abbildung $x : E \longrightarrow \mathbb{Q}_{\geq 0}$, so dass für jeden Knoten $v \in V$ gilt:

$$\sum_{\substack{e \in E: \\ v \in e}} x(e) \geq 1.$$

Notation:
Wenn wir Join-Anfragen Q betrachten, sprechen wir von "fraktionalen Kantenüberdeckungen von Q" und meinen damit fraktionale Kantenüberdeckungen des Anfrage-Hypergraphen G_Q. Statt "$x(R_i(\bar{x}_i))$" schreiben wir auch kurz: "$x(i)$".

Satz 2.9 (Die AGM-Schranke)

Sei Q eine Join-Anfrage der Form
$Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$.

Für jede fraktionale Kantenüberdeckung x von Q und jede Datenbank D gilt:

$$|Q(D)| \leq \prod_{i=1}^{m} N_i^{x(i)},$$

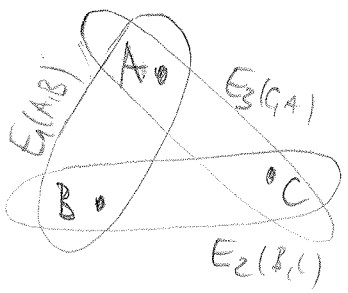
wobei $N_i := |R_i^D|$ für alle $i \in [1,m]$ ist.

Bevor wir die AGM-Schranke beweisen, schauen wir uns zunächst einige Beispiele an

Beispiel 2.10

(a) Betrachte die Anfrage $Q_\Delta'(A,B,C) \leftarrow E_1(A,B), E_2(B,C), E_3(C,A)$

Skizze von $G_{Q_\Delta'}$:



Die Abbildung x mit $x(1) = x(2) = x(3) = \frac{1}{2}$ ist eine f.Kü von Q_Δ'.
Laut AGM-Schranke gilt: $|Q_\Delta'(D)| \leq \prod_{i=1}^{3} N_i^{1/2} = \sqrt{N_1} \cdot \sqrt{N_2} \cdot \sqrt{N_3}$
f.a. DBen D und $N_i := E_i^D$ für $i \in [1,2,3]$.

Dies liefert auch eine leicht verbesserte Variante
von Satz Δ:

Für die Anfrage $Q_\Delta(A,B,C) \leftarrow E(A,B), E(B,C), E(C,A)$
und jede $\{E\}$-DB D gilt:

$$|Q_\Delta(D)| \leq (N^D)^{3/2},$$

wobei $N^D := |E^D|$ ist

Beweis:
Sei D eine $\{E\}$-DB und sei D' die $\{E_1, E_2, E_3\}$-DB
mit $E_i^{D'} := E^D$ f.a. $i \in [1,3]$. Dann gilt:

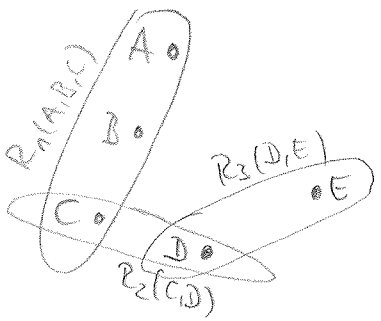
$Q_\Delta(D) = Q'_\Delta(D')$. Gemäß AGM-Schranke folgt:

$$|Q_\Delta(D)| = |Q'_\Delta(D')| \leq \sqrt{N^D} \cdot \sqrt{N^D} \cdot \sqrt{N^D} = (N^D)^{3/2}.$$

(b) Betrachte die Anfrage $Q_{(b)}$ aus Beispiel 2.5:

$$Q_{(b)}(A,B,C,D,E) \leftarrow R_1(A,B,C), R_2(C,D), R_3(D,E)$$

$\mathcal{H}_{Q_{(b)}}$:

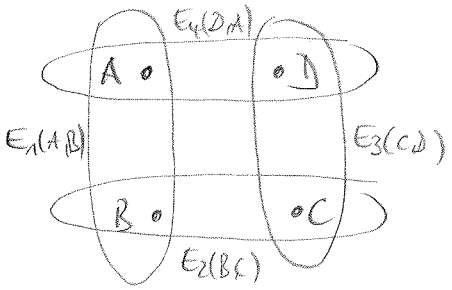


Für jede f.Kü x von $Q_{(b)}$ gilt: $x(1) \geq 1$ und $x(3) \geq 1$,
da der Knoten A (bzw. E) nur in einer einzigen
Hyperkante, nämlich $R_1(A,B,C)$ (bzw. $R_3(D,E)$) vorkommt.

(c) Betrachte die Anfrage

$$Q_4(A,B,C,D) \leftarrow E_1(A,B), E_2(B,C), E_3(C,D), E_4(D,A)$$

G_{Q_4}:



Einige fKü von Q_4:

	1	2	3	4
$x^{(1)}$	1	0	1	0
$x^{(2)}$	0	1	0	1
$x^{(3)}$	1/2	1/2	1/2	1/2
$x^{(4)}$	1/3	2/3	1/3	2/3

(d) Sei $k \in \mathbb{N}$ mit $k \geq 2$. Betrachte die Loomis-Whitney Join-Anfrage LW_k mit

$$LW_k(X_1, \dots, X_k) \leftarrow \underbrace{R_1(X_2, \dots, X_k), R_2(X_1, X_3, \dots, X_k), \dots, R_k(X_1, \dots, X_{k-1})}$$

d.h. für jedes $i \in [1,k]$ ein Atom der Form $R_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k)$

Eine fKü von LW_k ist die Abbildung $x: [1,k] \to \mathbb{Q}_{\geq 0}$ mit $x(i) := 1/(k-1)$ f.a. $i \in [1,k]$.

Die AGM-Schranke liefert: $|LW_k(D)| \leq \prod_{i=1}^{k} \sqrt[k-1]{N_i}$,

wobei $N_i := |R_i^D|$ f.a. $i \in [1,k]$ ist.

Insbes: Falls $N_1 = \dots = N_k = N$, so ist $|LW_k(D)| \leq N^{\frac{k}{k-1}} = N \cdot \sqrt[k-1]{N}$.

Betrachten wir nun folgendes Szenario:

Gegeben sei eine Join-Anfrage

$$Q(\bar{X}_0) \leftarrow R_1(\bar{X}_1), \ldots, R_m(\bar{X}_m)$$

mit $vars(Q) = \{A_1, \ldots, A_n\}$.

Außerdem seien Zahlen $N_1, \ldots, N_m$ bekannt, von denen wir wissen, dass die Datenbank D, auf der Q ausgewertet wird in der Relation R_i genau N_i Tupel enthält (d.h. $N_i = |R_i^D|$), f.a. $i \in [1,m]$

Die AGM-Schranke besagt, dass für jede fkü x von Q gilt: $|Q(D)| \leq \prod_{i=1}^{m} N_i^{x(i)}$

Frage: Wie können wir eine fkü x finden, für die $\prod_{i=1}^{m} N_i^{x(i)}$ möglichst klein ist?

Dazu lösen wir folgendes Optimierungsproblem:
Statt $x(i)$ schreibe x_i für $i \in [1,m]$ und setze $x := \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$.

Ziel: minimiere $\prod_{i=1}^{m} N_i^{x_i}$

unter der Bedingung, dass Folgendes gilt:

(1). f.a. $j \in [1,n]$ ist $\sum_{\substack{i \in [1,m]: \\ A_j \in \{\bar{X}_i\}}} x_i \geq 1$

und

(2). f.a. $i \in [1,m]$ ist $x_i \geq 0$, $x_i \in \mathbb{Q}$

Beachte: • $\prod_{i=1}^{m} N_i^{x_i}$ ist minimal

$\Leftrightarrow$ $\log\left(\prod_{i=1}^{m} N_i^{x_i}\right)$ ist minimal

$$= \sum_{i=1}^{m} \log\left(N_i^{x_i}\right) = \sum_{i=1}^{m} x_i \cdot \log N_i = \sum_{i=1}^{m} x_i \cdot c_i$$

mit $c_i := \log N_i$ f.a. $i \in [1,m]$.

Als Zielfunktion für unser Minimierungsproblem können wir statt $\prod_{i=1}^{m} N_i^{x_i}$ also auch $c^T x$ verwenden.

• Die Bedingung (1) lässt sich darstellen als

"$Ax \geq b$" für die Matrix $A_Q := A := (a_{ji})_{\substack{j \in [1,m], \\ i \in [1,m]}}$ mit

$$a_{ji} := \begin{cases} 1 & \text{falls } A_j \in \{\overline{X}_i\} \\ 0 & \text{sonst} \end{cases}$$

und den Vektor $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

Unser Ziel ist also:

(OPT:)

minimiere $c^T x$ (mit $c_i = \log N_i$ f.a. $i \in [1,m]$)

unter der Bedingung, dass

$Ax \geq b$, (mit $A = (a_{ji})$ und $a_{ji} := \begin{cases} 1 \text{ falls } A_j \in \{\overline{X}_i\} \\ 0 \text{ sonst} \end{cases}$)

$x \geq 0$, $x \in \mathbb{Q}^m$

Dies ist ein lineares Optimierungsproblem, für das wir eine rationale Lösung suchen.

Dazu stehen verschiedene Lösungsverfahren zur Verfügung:

- der Simplex-Algorithmus (exponentiell im worst-case, aber in der Praxis i.d.R. effizient)
- das Innere-Punkte-Verfahren
  (Karmarkar-Verfahren (1984) — läuft in Polynomialzeit)
- die Ellipsoid-Methode
  (Khachiyan (1979) — läuft in Polynomialzeit)

<u>Beachte:</u> gemäß Abth-Schranke

Ist x eine Lösung des Optimierungsproblems OPT, so gilt für jede DB D mit $N_i = |R_i^D|$ f.a. $i \in [1,m]$, dass $|Q(D)| \leq 2^{c^T x}$ ist.

Denn: $2^{c^T x} = 2^{\sum_{i=1}^{m} c_i x_i} = 2^{\sum_{i=1}^{m} x_i \cdot \log N_i} = 2^{\sum_{i=1}^{m} \log(N_i^{x_i})}$

$= 2^{\log(\prod_{i=1}^{m} N_i^{x_i})} = \prod_{i=1}^{m} N_i^{x_i}$

Beweis der AGM-Schranke

Zum Beweis benötigen wir einige Grundlagen aus der Wahrscheinlichkeitsrechnung und der Informationstheorie, die wir hier zunächst bereitstellen.

Wahrscheinlichkeitsräume

Definition: Ein endlicher **Wahrscheinlichkeitsraum** (Ω, P) besteht aus einer endlichen, nicht-leeren Menge Ω von **Ergebnissen** bzw. **Elementarereignissen**, denen Wahrscheinlichkeiten $P(\omega) = p_\omega \in \mathbb{R}$ für jedes $\omega \in \Omega$ zugeordnet sind, so dass gilt:

$0 \leq p_\omega \leq 1$, für jedes $\omega \in \Omega$, und $\sum_{\omega \in \Omega} p_\omega = 1$.

Ereignisse

Definition: Ein **Ereignis** ist eine Menge von Ergebnissen, d.h. eine Teilmenge von Ω.
Die **Wahrscheinlichkeit** eines Ereignisses $A \subseteq \Omega$ ist definiert als $P(A) := \sum_{\omega \in A} P(\omega)$.

Wir schreiben $\overline{A}$ um das Komplement von A zu bezeichnen, d.h. $\overline{A} = \Omega \setminus A$

Regeln zum Rechnen mit Wahrscheinlichkeiten:

Für jeden endlichen Wahrscheinlichkeitsraum (Ω, P) und für alle Ereignisse A und B gilt:

- $P(\overline{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Für $s \in \mathbb{N}_{\geq 1}$ und beliebige Ereignisse $A_1, \dots, A_s \subseteq \Omega$ gilt: $P(A_1 \cup \dots \cup A_s) \leq \sum_{i=1}^{s} P(A_i)$

Falls $A_i \cap A_j = \emptyset$ für alle $i, j \in [1,s]$ mit $i \neq j$, so gilt sogar: $P(A_1 \cup \dots \cup A_s) = \sum_{i=1}^{s} P(A_i)$.

Zufallsvariablen

Definition:

Sei (Ω, P) ein endlicher Wahrscheinlichkeitsraum.
Eine **Zufallsvariable** ist eine Funktion $Y: \Omega \to M$, für eine beliebige nicht-leere Menge M.
Für jedes $a \in M$ definieren wir die Wahrscheinlichkeit dafür, dass die Zufallsvariable Y den Wert a annimmt, durch

$P(Y = a) := P(A)$ für das Ereignis $A := \{\omega \in \Omega : Y(\omega) = a\}$

Beispiel 2.11

Betrachte eine Join-Anfrage Q der Form

$$Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$$

mit $\bar{x}_0 = A_1, \ldots, A_n$.

Sei D eine Datenbank vom Schema $\{R_1, \ldots, R_m\}$.

Wir betrachten das "Zufallsexperiment", bei dem wir zufällig, gleichverteilt ein beliebiges Tupel aus dem Anfrageergebnis $Q(D)$ ziehen.

Dies entspricht dem endlichen Wahrscheinlichkeitsraum (Ω, P) mit $\Omega := Q(D)$ und $P(\omega) := \frac{1}{|Q(D)|}$ f.a. $\omega \in Q(D)$.

Sei $M := \text{adom}(D)$.

Für jedes $j \in [1,n]$ sei $Y_j : \Omega \to M$ die Zufallsvariable, die jedem Tupel $t = (t_1, \ldots, t_n) \in Q(D)$ den Wert t_j zuordnet — d.h. $Y_j(t) = \pi_j(t)$ f.a. $t \in Q(D)$.

Für jedes feste Element $a \in \text{adom}(D)$ ist dann

$P(Y_j = a)$ die Wahrscheinlichkeit dafür, dass ein zufällig gleichverteilt aus $Q(D)$ gezogenes Tupel in der j-ten Spalte den Eintrag a hat. D.h.

$$P(Y_j = a) = \frac{|\{t \in Q(D) : \pi_j(t) = a\}|}{|Q(D)|}.$$

Informationsgehalt und Entropie

Definition:
Sei (Ω, P) ein endlicher Wahrscheinlichkeitsraum.
Der **Informationsgehalt** eines Elementarereignisses $\omega \in \Omega$ ist definiert als

$$\log\left(\frac{1}{p_\omega}\right) \qquad \left(= -\log(p_\omega) \right)$$

Die **Entropie** von (Ω, P) ist definiert als

$$H((\Omega,P)) := \sum_{\omega \in \Omega} p_\omega \cdot \log\left(\frac{1}{p_\omega}\right) \qquad \left(= -\sum_{\omega \in \Omega} p_\omega \cdot \log(p_\omega) \right)$$

Konvention für den Fall, dass $p_\omega = 0$ ist:

$$0 \cdot \log\left(\frac{1}{0}\right) := -0 \cdot \log(0) := 0$$

Bemerkung:
Man kann beweisen, dass Folgendes gilt:

$$H((\Omega,P)) \leq \left[\begin{array}{l}\text{durchschnittliche Länge des Bitstrings,}\\ \text{der ein gemäß } P \text{ zufällig aus } \Omega\\ \text{gewähltes Elementarereignis kodiert}\\ \text{(bei optimaler Kodierung)}\end{array}\right] \leq H((\Omega,P))+1$$

Wegen $H((\Omega,P)) \overset{\text{Def}}{=} \sum_{\omega \in \Omega} p_\omega \cdot \log\left(\frac{1}{p_\omega}\right)$ gilt daher

$$\left[\begin{array}{l}\text{Informationsgehalt}\\ \text{von } \omega \in \Omega\end{array}\right] \overset{\text{Def}}{=} \log\left(\frac{1}{p_\omega}\right) \approx \left[\begin{array}{l}\text{Länge des Bitstrings,}\\ \text{der } \omega \text{ kodiert}\\ \text{(bei optimaler Kodierung)}\end{array}\right]$$

Definition (Entropie einer Zufallsvariablen)

Sei (Ω, P) ein endlicher Wahrscheinlichkeitsraum, sei M eine beliebige nicht-leere Menge und sei $Y: \Omega \to M$ eine Zufallsvariable.

Die **Entropie von Y** ist definiert als

$$H(Y) := \sum_{a \in M} P(Y=a) \cdot \log\left(\frac{1}{P(Y=a)}\right).$$

Bemerkung:

Wir können uns $H(Y)$ vorstellen als ein Maß für die "Unsicherheit" bzw. "Unbestimmtheit" von Y.

Falls es ein $b \in M$ gibt, so das $P(Y=b) = 1$, so wissen wir, dass die Zufallsvariable, unabhängig vom Ausgang des Zufallsexperiments, stets den Wert b annimmt — die "Unsicherheit" bzw. "Unbestimmtheit" von Y ist also 0.

Die spiegelt sich in der Entropie wider, da

$$H(Y) \overset{\text{Def}}{=} \sum_{a \in M} P(Y=a) \cdot \log\left(\frac{1}{P(Y=a)}\right) = P(Y=b) \cdot \log\left(\frac{1}{P(Y=b)}\right)$$
$$= 1 \cdot \log\left(\frac{1}{1}\right) = 0$$

ist.

Falls jedoch für jedes $b \in M$ gilt, dass $P(Y=b) = \frac{1}{|M|}$, so ist die "Unsicherheit / Unbestimmtheit" von Y größtmöglich und es gilt: $H(Y) = \sum_{a \in M} \frac{1}{|M|} \cdot \log(|M|) = \log(|M|)$

Wichtige Eigenschaften der Entropie:

Sei (Ω, P) ein endlicher Wahrscheinlichkeitsraum

(1) Für jede Zufallsvariable $Y: \Omega \to M$ mit endlichem M gilt:

$$0 \leq H(Y) \leq \log(|M|)$$

Außerdem gilt:

- $H(Y) = 0 \iff$ es $b \in M$ s.d. $P(Y=b) = 1$
- $H(Y) = \log(|M|) \iff$ f.a. $b \in M$ ist $P(Y=b) = \frac{1}{|M|}$.

(2) Seien $Y_1: \Omega \to M_1$ und $Y_2: \Omega \to M_2$ zwei Zufallsvariablen. Die zusammengesetzte Zufallsvariable (Y_1, Y_2) ist die Zufallsvariable $Y: \Omega \to M_1 \times M_2$ mit $Y(\omega) := (Y_1(\omega), Y_2(\omega))$ f.a. $\omega \in \Omega$.

- Die <u>gemeinsame Entropie von Y_1 und Y_2</u> ist die Entropie von (Y_1, Y_2). D.h.:

$$H(Y_1, Y_2) = \sum_{\substack{a_1 \in M_1, \\ a_2 \in M_2}} P(Y_1 = a_1, Y_2 = a_2) \cdot \log\left(\frac{1}{P(Y_1 = a_1, Y_2 = a_2)}\right)$$

(Komma $\hat{=}$ "und")

- Für $a_1 \in M_1$ mit $P(Y_1 = a_1) \neq 0$ ist die <u>bedingte Entropie von Y_2 unter der Voraussetzung, dass $Y_1 = a_1$ ist</u>,

$$H(Y_2 \mid Y_1 = a_1) := \sum_{a_2 \in M_2} P(Y_2 = a_2 \mid Y_1 = a_1) \cdot \log\left(\frac{1}{P(Y_2 = a_2 \mid Y_1 = a_1)}\right)$$

Hierbei ist $P(Y_2 = a_2 \mid Y_1 = a_1)$ die bedingte Wahrscheinlichkeit dafür, dass Y_2 den Wert a_2 annimmt unter der Voraussetzung, dass Y_1 den Wert a_1 angenommen hat.

D.h.: $P(Y_2 = a_2 \mid Y_1 = a_1) := \dfrac{P(Y_2 = a_2, Y_1 = a_1)}{P(Y_1 = a_1)}$.

- Die <u>bedingte Entropie von Y_2 bei gegebenem Y_1</u> ist

$$H(Y_2 \mid Y_1) := \sum_{a_1 \in M_1} P(Y_1 = a_1) \cdot H(Y_2 \mid Y_1 = a_1)$$

$$= \sum_{a_1 \in M_1} \left(P(Y_1 = a_1) \cdot \sum_{a_2 \in M_2} P(Y_2 = a_2 \mid Y_1 = a_1) \cdot \log\left(\frac{1}{P(Y_2 = a_2 \mid Y_1 = a_1)} \right) \right.$$

- Es gilt: $\boxed{H(Y_1, Y_2) = H(Y_1) + H(Y_2 \mid Y_1)}$ (1)

Beweis:

$$H(Y_1, Y_2) \overset{\text{Def}}{=} \sum_{\substack{a_1 \in M_1 \\ a_2 \in M_2}} \overbrace{P(Y_1 = a_1) \cdot P(Y_2 = a_2 \mid Y_1 = a_1)}^{= P(Y_1 = a_1, Y_2 = a_2)} \cdot \log\left(\frac{1}{P(Y_1 = a_1, Y_2 = a_2)} \right)$$

$$= \sum_{a_1 \in M_1} \left(P(Y_1 = a_1) \cdot \sum_{a_2 \in M_2} P(Y_2 = a_2 \mid Y_1 = a_1) \cdot \left(\log\left(\frac{1}{P(Y_1 = a_1)} \right) + \log\left(\frac{1}{P(Y_2 = a_2 \mid Y_1 = a_1)} \right) \right) \right.$$

$$= \sum_{a_1 \in M_1} P(Y_1 = a_1) \cdot \log\left(\frac{1}{P(Y_1 = a_1)} \right) \cdot \underbrace{\left(\sum_{a_2 \in M_2} P(Y_2 = a_2 \mid Y_1 = a_1) \right)}_{= 1}$$

$$+ \sum_{a_1 \in M_1} P(Y_1 = a_1) \cdot \left(\sum_{a_2 \in M_2} P(Y_2 = a_2 \mid Y_1 = a_1) \cdot \log\left(\frac{1}{P(Y_2 = a_2 \mid Y_1 = a_1)} \right) \right)$$

$$= H(Y_1) + H(Y_2 \mid Y_1) \qquad \square$$

- Es gilt: $\boxed{H(Y_2 \mid Y_1) \leq H(Y_2)}$ (2)

Intuition dazu: Die "Unsicherheit von Y_2" kann durch die zusätzliche Information hinsichtlich Y_1 höchstens abnehmen, aber nicht größer werden.

Beweis von (2): Übung!

Hinweis: Verwende die so genannte Jensensche Ungleichung, die Folgendes besagt:

Für jede **konkave** Funktion f und für Zahlen $\lambda_1, \dots, \lambda_n \in \mathbb{R}_{\geq 0}$ mit $\sum_{i=1}^{n} \lambda_i = 1$ gilt

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i f(x_i) \qquad \text{für alle } x_1, \dots, x_n \in \mathbb{R}$$

Eine Funktion $f: \mathbb{R} \to \mathbb{R}$ heißt **konkav**, wenn für alle $x, y \in \mathbb{R}$ und alle $\lambda \in \mathbb{R}$ mit $0 \leq \lambda \leq 1$ gilt:

$$f(\lambda x + (1-\lambda) y) \geq \lambda f(x) + (1-\lambda) f(y)$$

Insbesondere ist die Funktion $f(x) := \log x$ konkav

(3) Die Begriffe aus (2) können für zusammengesetzte Zufallsvariablen $(Y_1, \dots, Y_n)$ verallgemeinert werden:

Sei $n \geq 2$. Für jedes $j \in [n]$ sei $Y_j: \Omega \to M_j$ eine Zufallsvariable.

Die zusammengesetzte Zufallsvariable $(Y_1, \dots, Y_n)$ ist die Zufallsvariable $Y: \Omega \to M_1 \times \dots \times M_n$ mit

$$Y(\omega) := (Y_1(\omega), \dots, Y_n(\omega)) \qquad \text{f.a. } \omega \in \Omega.$$

- Die gemeinsame Entropie von $Y_1, \dots, Y_n$ ist

$$H(Y_1, \dots, Y_n) := \sum_{(a_1, \dots, a_n) \in M_1 \times \dots \times M_n} P\left(\bigwedge_{j=1}^{n} Y_j = a_j\right) \cdot \log\left(\frac{1}{P\left(\bigwedge_{j=1}^{n} Y_j = a_j\right)}\right)$$

- Analog zu (1) gilt

$$H(Y_1, \dots, Y_n) = H(Y_1) + H(Y_2 | Y_1) + H(Y_3 | Y_1, Y_2) + \dots + H(Y_n | Y_1, \dots, Y_{n-1}) \quad (1')$$

- Analog zu (2) gilt f.a. $J \subseteq J' \subseteq [n]$ und f.a. $k \in [n]$:

$$H(Y_k | (Y_{j'})_{j' \in J'}) \leq H(Y_k | (Y_j)_{j \in J}) \quad (2')$$

Das zentrale Werkzeug, das wir zum Beweis der AGM-Schranke verwenden werden, ist

Shearer's Lemma: Sei (Ω, P) ein endliches Wahrscheinlichkeitsraum. Sei $n \in \mathbb{N}_{\geq 1}$ und sei $I := [1, \dots, n]$. Für jedes $j \in I$ sei $Y_j : \Omega \to M_j$ eine Zufallsvariable.
Für jedes $J \subseteq I$ sei $Y_J := (Y_j)_{j \in J}$.
Sei $\ell \in \mathbb{N}_{\geq 1}$, und seien $J_1, \dots, J_\ell$ Teilmengen von I, so dass für eine Zahl $q \in \mathbb{N}$ gilt: Jedes $j \in I$ kommt in mindestens q der Mengen $J_1, \dots, J_\ell$ vor.
Dann gilt: $H(Y_1, \dots, Y_n) \leq \frac{1}{q} \cdot \sum_{k=1}^{\ell} H(Y_{J_k})$

Beweis:

Gemäß (1) gilt für jedes $J \subseteq I$, dass

$$H(Y_J) = \sum_{j \in J} H\left(Y_j \mid (Y_i)_{\substack{i \in J \text{ mit} \\ i<j}}\right). \quad (*)$$

Somit gilt:

$$\sum_{k=1}^{\ell} H(Y_{J_k})$$

$$\overset{(*)}{=} \sum_{k=1}^{\ell} \sum_{j \in J_k} H\left(Y_j \mid (Y_i)_{\substack{i \in J_k \text{ mit} \\ i<j}}\right)$$

$$\overset{(2)}{\geq} \sum_{k=1}^{\ell} \sum_{j \in J_k} H\left(Y_j \mid (Y_i)_{\substack{i \in I \text{ mit} \\ i<j}}\right)$$

$$\geq q \cdot \sum_{j \in I} H\left(Y_j \mid (Y_i)_{\substack{i \in I \text{ mit} \\ i<j}}\right) \overset{(*)}{=} q \cdot H(\underbrace{Y_I}_{= (Y_1, \ldots, Y_n)})$$

denn laut Voraussetzung gilt:
jedes $j \in I$ kommt in
mindestens q der Mengen $J_1, \ldots, J_\ell$ vor.

□

Beweis der AGM-Schranke, d.h. Beweis von Satz 2.9:

Sei Q eine Join-Anfrage der Form
$Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \dots, R_m(\bar{x}_m)$ und sei $\bar{x}_0 = A_1 \cdots A_n$.
Sei x eine fraktionale Kantenüberdeckung von Q.
Sei D eine Datenbank vom Schema $\{R_1, \dots, R_m\}$, und
für jedes $i \in [1,m]$ sei $N_i := |R_i^D|$.

Zu zeigen: $|Q(D)| \leq \prod_{i=1}^{m} N_i^{x(i)}$.

Falls $Q(D) = \emptyset$ ist, so gilt dies offensichtlicherweise.
Falls $Q(D) \neq \emptyset$ ist, so betrachte das Zufallsexperiment aus Beispiel 2.11, bei dem zufällig, gleichverteilt ein beliebiges Tupel aus $Q(D)$ gewählt wird. D.h: Wir betrachten den endlichen Wahrscheinlichkeitsraum (Ω, P) mit $\Omega := Q(D)$ und $P(t) := \frac{1}{|Q(D)|}$ für jedes $t \in Q(D)$.
Sei $M := \text{adom}(D)$, und für jedes $j \in [1,n]$ sei $Y_j : \Omega \to M$ die Zufallsvariable mit $Y_j(t) := \pi_j(t)$ für jedes $t \in Q(D)$ (d.h: Y_j ordnet jedem Tupel in $Q(D)$ seinen Eintrag in der j-ten Spalte zu).
Beachte: Dann ist $Y := (Y_1, \dots, Y_n)$ die Zufallsvariable $Y : Q(D) \to \text{adom}(D)^n$ mit $Y(t) = t$ f.a. $t \in Q(D)$.
Insbes. gilt für jedes $t \in \text{adom}(D)^n$, dass

$$P(Y = t) = \begin{cases} 0 & \text{falls } t \notin Q(D) \\ \frac{1}{|Q(D)|} & \text{falls } t \in Q(D). \end{cases}$$

Daher ist

$$H(Y) = \sum_{t \in \text{adom}(D)^n} P(Y=t) \cdot \log\left(\frac{1}{P(Y=t)}\right) = \sum_{t \in Q(D)} P(Y=t) \cdot \log\left(\frac{1}{P(Y=t)}\right) = \log(|Q(D)|) \quad (*)$$

Laut Voraussetzung ist x eine fraktionale Kantenüberdeckung von Q, d.h.: $x(1), \dots, x(m) \in \mathbb{Q}_{\geq 0}$ und für jedes $j \in [1,n]$ gilt

$$\sum_{\substack{i \in [1,m] \text{ mit} \\ A_j \in \{\overline{X}_i\}}} x(i) \geq 1 \qquad (2)$$

Wegen $x(1), \dots, x(m) \in \mathbb{Q}_{\geq 0}$ gibt es ein $q \in \mathbb{N}_{\geq 1}$ und $p_1, \dots, p_m \in \mathbb{N}$, s.d. $x(i) = \frac{p_i}{q}$ f.a. $i \in [1,m]$ ist.

Sei $I := \{1, \dots, n\}$. Für jedes $i \in [1,m]$ sei

$J^{(i)} := \{ j \in [1,n] : A_j \in \overline{X}_i \}$, und sei

$J^{(i)}_1, \dots, J^{(i)}_{p_i}$ eine Liste, die aus p_i Kopien von $J^{(i)}$ besteht.

Dann ist $J^{(1)}_1, \dots, J^{(1)}_{p_1}, \dots, J^{(m)}_1, \dots, J^{(m)}_{p_m}$ eine Liste von $\ell = p_1 + \dots + p_m$ Teilmengen von I, für die gilt:

Jedes $j \in I$ kommt in mindestens q dieser Mengen vor,

denn: Für jedes $j \in I = [1,n]$ gilt:

Anzahl der Mengen in $\left(J^{(i)}_1, \dots, J^{(i)}_{p_i}\right)_{i \in [1,m]}$, in denen j vorkommt

$$= \sum_{\substack{i \in [1,m] \text{ mit} \\ j \in J^{(i)}}} p_i$$

$$= \sum_{\substack{i \in [1,m] \text{ mit} \\ A_j \in \{\overline{X}_i\}}} q \cdot \underbrace{\frac{p_i}{q}}_{= x(i)} = q \cdot \sum_{\substack{i \in [1,m] \text{ mit} \\ A_j \in \{\overline{X}_i\}}} x(i) \overset{(2)}{\geq} q.$$

Shearer's Lemma liefert: $H(Y) \leq \frac{1}{q} \cdot \sum_{\substack{i \in [1,m], \\ k \in [1,p_i]}} H\left(Y_{J^{(i)}_k}\right)$, (3)

wobei $Y_{j^{(i)}_k} = (Y_j)_{j \in J^{(i)}_k} = (Y_j)_{A_j \in \{\bar{x}_i\}}$ ist

Beachte: $Y_{j^{(i)}} := (Y_j)_{A_j \in \{\bar{x}_i\}}$ ist eine Zufallsvariable

$Y_{j^{(i)}} : Q(D) \to \text{adom}(D)^{\text{ar}(R_i)}$, für die für jedes $t \in Q(D)$ gilt: $Y_{j^{(i)}}(t) \in R_i^D$

(da Q von der Form $Q(A_1,\dots,A_n) \leftarrow R_1(\bar{x}_1), \dots, R_m(\bar{x}_m)$ ist).

Daher ist

$$H(Y_{j^{(i)}}) \leq \log(|R_i^D|) = \log N_i. \quad (4)$$

(da $H(Y) \leq \log(|M|)$ für jede Zufallsvariable $Y: \Omega \to M$ gilt).

Insgesamt erhalten wir:

$$\log(|Q(D)|) \overset{(1)}{=} H(Y) \overset{(3)}{\leq} \frac{1}{q} \cdot \sum_{i \in [1,m]} p_i \cdot H(Y_{j^{(i)}})$$

$$\overset{x(i) = \frac{p_i}{q}}{=} \sum_{i \in [1,m]} x(i) \cdot H(Y_{j^{(i)}})$$

$$\overset{(4)}{\leq} \sum_{i \in [1,m]} x(i) \cdot \log N_i$$

Somit ist

$$|Q(D)| = 2^{\log(|Q(D)|)} \leq 2^{\sum_{i=1}^{m} x(i) \cdot \log N_i}$$

$$= \prod_{i=1}^{m} 2^{x(i) \cdot \log N_i} = \prod_{i=1}^{m} N_i^{x(i)}$$

□ Satz 2.9

Die AGM-Schranke ist im folgenden Sinn optimal:

Satz 2.10 (Optimalität der AGM-Schranke)

Sei Q eine Join-Anfrage der Form

$$Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m).$$

Für jedes $N \in \mathbb{N}$ gibt es eine Datenbank D vom Schema $\{R_1, \ldots, R_m\}$ mit $N_i := |R_i^D| \geq N$ für alle $i \in [1,m]$ und eine fraktionale Kantenüberdeckung x von Q, so dass

$$|Q(D)| = \prod_{i=1}^{m} N_i^{x(i)} \quad \text{ist.}$$

Beweis:

Zum Beweis nutzen wir das aus der linearen Optimierung bekannte Dualitätsprinzip:

Zur Erinnerung:

Primales LP: (P)

minimiere $c^T x$

unter der Bedingung:

$Ax \geq b$, $x \geq 0$

mit $c = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} \in \mathbb{R}^m$, $x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$

$A = (a_{ji})_{\substack{j \in [1,n], \\ i \in [1,m]}}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$

Duales LP: (D)

maximiere $y^T b$

unter der Bedingung

$y^T A \leq c^T$, $y \geq 0$

mit $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

Für jedes $x \in \mathbb{R}^m$ mit $Ax \geq b$, $x \geq 0$
und jedes $y \in \mathbb{R}^n$ mit $y^T A \leq c^T$, $y \geq 0$ gilt:

$$c^T x \geq y^T A x \geq y^T b.$$

Insbes. gilt für jede optimale Lösung x^* von (P)
und jede optimale Lösung y^* von (D), dass

$$c^T x^* \geq y^{*T} b$$

Der <u>starke Dualitätssatz</u> besagt, dass sogar gilt:

$$c^T x^* = y^{*T} b.$$

Um Satz 2.10 zu beweisen, nutzen wir den starken Dualitätssatz wie folgt:

Gegeben seien eine Join-Anfrage $Q(A_1, \dots, A_n) \leftarrow R_1(\overline{x}_1), \dots, R_m(\overline{x}_m)$
und Zahlen $N_1, \dots, N_m \in \mathbb{N}_{\geq 1}$.
Betrachte das Primale LP $P(Q, N_1, \dots, N_m)$:

minimiere $c^T x$ — $\sum_{i=1}^{n} x_i \cdot \log N_i$
unter der Bedingung
$Ax \geq b$, $x \geq 0$ — f.a. $j \in [1,m]$ ist $\sum_{i \in [1,n]: A_j \in \{\overline{x}_i\}} x_i \geq 1$ und f.a. $i \in [1,n]$ ist $x_i \geq 0$

mit $c = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix}$ und $c_i := \log N_i$ f.a. $i \in [1,m]$,

$A = (a_{ji})_{\substack{j \in [1,m], \\ i \in [1,n]}}$ mit $a_{ji} = \begin{cases} 1 & \text{falls } A_j \in \{\overline{x}_i\} \\ 0 & \text{sonst} \end{cases}$

und $b = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$.

Das zugehörige Duale LP $\boxed{D(Q, N_1, \dots, N_m)}$ ist dann:

maximiere $y^T b$
unter der Bedingung
$y^T A \leq c^T, \quad y \geq 0$

$\sum_{j=1}^{n} y_j$

f.a. $i \in [1,m]$ ist $\sum_{\substack{j \in [1,n]: \\ A_{ji} \in \{\overline{x}_i\}}} y_j \leq \log N_i$
und
f.a. $j \in [1,n]$ ist $y_j \geq 0$

Der Starke Dualitätssatz besagt, dass für jede optimale Lösung x^* von $P(Q, N_1, \dots, N_m)$ und jede optimale Lösung y^* von $D(Q, N_1, \dots, N_m)$ gilt.

$$c^T x^* = y^{*T} b$$

$$\sum_{i=1}^{m} x_i^* \cdot \log N_i \qquad \sum_{j=1}^{n} y_j^*$$

Außerdem ist laut Aufg. Schranke (Satz 2.9)

$$|Q(D)| \leq \prod_{i=1}^{m} N_i^{x_i^*} = 2^{\sum_{i=1}^{m} x_i^* \cdot \log N_i} = 2^{\sum_{j=1}^{n} y_j^*}$$

für jede DB D mit $|R_i^D| = N_i$ f.a. $i \in [1,m]$.

Wir betrachten dies hier für den Fall, dass alle Zahlen N_i Zweierpotenzen sind, d.h. $N_i = 2^{L_i}$ mit $L_i \in \mathbb{N}$ f.a. $i \in [1,m]$. Dann ist $c_i = \log N_i = L_i \in \mathbb{N}$ f.a. $i \in [1,m]$, und aus der Linearen Optimierung ist bekannt, dass das Optimierungsproblem $D(Q, N_1, \dots, N_m)$ eine optimale Lösung $y^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_m^* \end{pmatrix}$ besitzt mit $y_j \in \mathbb{Q}$ f.a. $j \in [1,n]$.

Seien $p_1, \dots, p_n, q \in \mathbb{N}$ mit $q \neq 0$, s.d. $y_j^* = \frac{p_j}{q}$ f.a. $j \in [1,n]$.

Behauptung: $\hat{y} := \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$ ist eine optimale Lösung von $D(Q, N_1^q, \dots, N_m^q)$

Beweis: Wir wissen, dass $y^* = \begin{pmatrix} p_1/q \\ \vdots \\ p_n/q \end{pmatrix}$ eine optimale Lösung von $D(Q, N_1, \dots, N_m)$ ist.

maximiere $\sum_{j=1}^{n} y_j$ unter der Bedingung $y \geq 0$, $y^T A \leq c^T$

mit $c = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix}$ und $c_i = \log N_i$ f.a. $i \in [1,m]$.

Außerdem ist $D(Q, N_1^q, \dots, N_m^q)$ das Optimierungsproblem

maximiere $\sum_{j=1}^{n} y'_j$ unter der Bedingung $y' \geq 0$, $y'^T A \leq c'^T$

mit $c' = \begin{pmatrix} c'_1 \\ \vdots \\ c'_m \end{pmatrix}$ und $c'_i = \log(N_i^q) = q \cdot \log N_i = q \cdot c_i$

f.a. $i \in [1,m]$.

Daher gilt Folgendes f.a. $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ und $y' := \begin{pmatrix} qy_1 \\ \vdots \\ qy_n \end{pmatrix}$

$$y \geq 0 \text{ und } y^T A \leq c^T \iff y' \geq 0 \text{ und } y'^T A \leq c'^T$$

Da y^* eine optimale Lösung von $D(Q, N_1, \dots, N_m)$ ist, ist also $\hat{y}$ eine optimale Lösung von $D(Q, N_1^q, \dots, N_m^q)$. □Beh

Gemäß starkem Dualitätssatz und AGM-Schranke gilt für jede DB D mit $|R_i^D| = N_i^q$ f.a. $i \in [1,m]$, dass

$$|Q(D)| \leq 2^{\sum_{j=1}^{n} \hat{y}_j} = 2^{\sum_{j=1}^{n} p_j} \text{ ist.}$$

Wir konstruieren nun eine konkrete DB D, für die gilt:

$|R_i^D| = N_i^q$ f.a. $i \in [1,m]$ und $|Q(D)| = 2^{\sum_{j=1}^{n} p_j} = \prod_{j=1}^{n} 2^{p_j}$

Klar: Dies beweist dann insbes die Aussage von Satz 2.10.

Zur Konstruktion von $\mathcal{D}$ gehen wir wie folgt vor:

Für jedes $i \in [1,m]$ sei $r_i := ar(R_i)$, und seien $j(i,1), \dots, j(i,r_i) \in [1,m]$ so dass $\overline{x}_i = A_{j(i,1)}, \dots, A_{j(i,r_i)}$.

D.h. das i-te Atom $R_i(\overline{x}_i)$ im Rumpf von Q ist von der Form $R_i(A_{j(i,1)}, \dots, A_{j(i,r_i)})$.

Sei $J^{(i)} := \{j(i,1), \dots, j(i,r_i)\} = \{j \in [1,m] : A_j \in \{\overline{x}_i\}\}$.

Da $\hat{y} = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix}$ eine Lösung von $\mathcal{D}(Q, N_1^q, \dots, N_m^q)$ ist, gilt insbes.: $\hat{y}^T A \leq c^T$, dh f.a. $i \in [1,m]$ ist

$$\left.\begin{aligned} \sum_{\substack{j \in [1,m]: \\ A_j \in \{\overline{x}_i\}}} \hat{y}_j &\leq \log(N_i^q) \\ \sum_{k=1}^{r_i} \hat{y}_{j(i,k)} &= \sum_{k=1}^{r_i} p_{j(i,k)} \end{aligned}\right\} \quad \square$$

Wähle die DB $\mathcal{D}'$ mit $R_i^{\mathcal{D}'} := [1, 2^{p_{j(i,1)}}] \times \dots \times [1, 2^{p_{j(i,r_i)}}]$.

Dann ist $Q(\mathcal{D}') = \{t = (t_1, \dots, t_m) : (t_{j(i,1)}, \dots, t_{j(i,r_i)}) \in R_i^{\mathcal{D}'} \text{ f.a. } i \in [1,m]\}$

$$= [1, 2^{p_1}] \times \dots \times [1, 2^{p_m}],$$

und somit ist $|Q(\mathcal{D}')| = \prod_{j=1}^{m} 2^{p_j}$.

Außerdem gilt für jedes $i \in [1,m]$, dass

$$|R_i^{D'}| = \prod_{k=1}^{r_i} 2^{p_{j(i,k)}} = 2^{\sum_{k=1}^{r_i} p_{j(i,k)}} \overset{\text{(1)}}{\leq} 2^{\log(N_i^q)} = N_i^q.$$

Durch Hinzufügen von weiteren Tupeln erhalten wir eine DB D mit $R_i^D \supseteq R_i^{D'}$ und $|R_i^D| = N_i^q$ f.a. $i \in [1,m]$.

Es gilt: $Q(D) \supseteq Q(D')$, also

$|Q(D)| \geq \prod_{j=1}^{n} 2^{p_j}$ — und auf Grund der AGM-Schranke (Satz 2.9) wissen wir auch, dass $|Q(D)| \leq \prod_{j=1}^{n} 2^{p_j}$ ist. Dies beendet den Beweis von Satz 2.10

□

2.3 Verallgemeinerung der AGM-Schranke für beliebige konjunktive Anfragen

Ziel dieses Abschnitts ist, eine Verallgemeinerung der AGM-Schranke zu finden, die nicht nur für Join-Anfragen, sondern für beliebige konjunktive Anfragen gilt. Wir gehen dazu in mehreren Schritten vor, in denen wir immer allgemeinere Varianten von Anfragen betrachten.

Zunächst betrachten wir:

Projektionen von Join-Anfragen

Definition 2.11:

(a) Eine Projektion einer Join-Anfrage ist eine konjunktive Anfrage Q der Form

$$Q(A_1, \dots, A_k) \leftarrow R_1(\bar{x}_1), \dots, R_m(\bar{x}_m),$$

für die es ein $n \geq k$ und Variablen $A_{k+1}, \dots, A_n$ gibt, so dass die Anfrage $\tilde{Q}$ mit

$$\tilde{Q}(A_1, \dots, A_n) \leftarrow R_1(\bar{x}_1), \dots, R_m(\bar{x}_m)$$

eine Join-Anfrage ist.

(b) Sei $Q(A_1,\dots,A_k) \leftarrow R_1(\overline{x}_1),\dots,R_m(\overline{x}_m)$ eine Projektion eines Join-Anfrage.
Eine fraktionale Kantenüberdeckung von Q ist eine Abbildung $x: [m] \to \mathbb{R}_{\geq 0}$, s.d. f.a. $j \in [k]$ gilt:

$$\sum_{i \in [m] \text{ mit } A_j \in \{\overline{x}_i\}} x(i) \geq 1$$

Unter Verwendung der AGM-Schranke (Satz 2.9 & 2.10) können wir leicht folgern, dass Folgendes gilt:

Satz 2.12:

Sei Q eine Projektion eines Join-Anfrage.

(a) Für jede fraktionale Kantenüberdeckung x von Q und jede Datenbank D vom Schema $\{R_1,\dots,R_m\}$ gilt

$$|Q(D)| \leq \prod_{i=1}^{m} N_i^{x(i)},$$

wobei $N_i := |R_i^D|$ f.a. $i \in [m]$ ist.

(b) Für jedes $N \in \mathbb{N}$ gibt es eine $\{R_1,\dots,R_m\}$-DB D mit $N_i := |R_i^D| \geq N$ f.a. $i \in [m]$ und eine fraktionale Kantenüberdeckung x von Q, s.d.

$$|Q(D)| = \prod_{i=1}^{m} N_i^{x(i)}.$$

Beweis:

(a) Sei $\tilde{Q}(A_1,\dots,A_n) \leftarrow R_1(\bar{x}_1),\dots,R_m(\bar{x}_m)$ eine zu Q gehörige Join-Anfrage (i.s.v. Def 2.11 (a)).

Für jedes $i \in [m]$ sei $\bar{x}_i'$ die Variablenliste, die aus $\bar{x}_i$ entsteht, indem man alle Variablen aus $\{A_{k+1},\dots,A_n\}$ weglässt, sei r_i' die Länge der Liste $\bar{x}_i'$ und sei R_i' ein neues Relationssymbol der Stelligkeit r_i'.
Wir betrachten die Anfrage Q' vom Schema $\{R_1',\dots,R_m'\}$ mit

$$Q'(A_1,\dots,A_k) \leftarrow R_1'(\bar{x}_1'),\dots,R_m'(\bar{x}_m').$$

Offensichtlicherweise ist Q' eine Join-Anfrage, und für jede Abbildung $x: [m] \to \mathbb{Q}_{\geq 0}$ gilt:

(1) x ist eine fraktionale Kantenüberdeckung von Q' $\iff$ x ist eine fraktionelle Kantenüberd. von Q'.

Sei nun D eine beliebige DB vom Schema $\{R_1,\dots,R_m\}$.
Sei D' die DB vom Schema $\{R_1',\dots,R_m'\}$ mit

$$R_i'^{D'} := Q_i(D)$$

wobei Q_i die Anfrage $Q_i(\bar{x}_i') \leftarrow R_i(\bar{x}_i)$ ist, f.a. $i \in [m]$ (d.h.: $R_i'^{D'}$ ist die Projektion von R_i^D auf die Komponenten, die zu $A_1,\dots,A_k$ gehören).

Klar: $N_i' := |R_i'^{D'}| \leq |R_i^D| = N_i$

Man kann sich leicht davon überzeugen, dass

$$Q(D) \cong Q'(D')$$

ist. Sei x eine fraktionale Kantenüberdeckung von Q (und Q', wegen ▷). Gemäß AGM-Schranke (Satz 2.9) gilt

$$|Q'(D')| \leq \prod_{i=1}^{m} N_i^{x(i)}$$

$$|Q(D)| \leq \prod_{i=1}^{m} N_i^{x(i)}$$

$\square$(a)

(b) Wir nutzen die gleiche Notation wie im Beweis von (a). Sei $N \in \mathbb{N}$ beliebig. Gemäß Optimalität der AGM-Schranke (Satz 2.10) gibt es eine DB D' vom Schema $\{R_1', \dots, R_m'\}$ mit

$N_i := |R_i^{D'}| \geq N$ f.a. $i \in [m]$

und eine fraktionale Kantenüberdeckung x von Q' (und Q, wegen ▷), so dass

$$|Q'(D')| = \prod_{i=1}^{m} N_i^{x(i)}.$$

Sei a ein beliebiges, fest gewähltes Element in dom und sei D die DB vom Schema $\{R_1, \dots, R_m\}$, die man aus D' erhält, indem man jedes Tupel $t' \in R_i^{D'}$ an geeigneten Stellen um weitere

Komponenten anreichert, deren Eintrag jeweils den Wert a bekommt

Beispiel: Falls das i-te Atom von Q die Form $R_i(A_2, A_{k+1}, A_1)$ hat (und $k \geq 3$), so hat das i-te Atom von Q' die Form $R_i'(A_2, A_1)$. Für jedes Tupel $t' = (x,y) \in R_i'^{D'}$ enthält R_i^D dann das Tupel $\hat{t} := (x, a, y)$.

Präzise: Für $i \in [m]$ sei $R_i(\bar{x}_i)$ von der Form $R_i(A_{j(i,1)}, \dots, A_{j(i,r_i)})$ und sei $M_i := \{\ell \in \{1, \dots, r_i\} : j(i,\ell) > k\}$. Jedem Tupel $t' = (t'_1, \dots, t'_{r_i}) \in \mathrm{dom}^{r_i}$ ordnen wir das Tupel $\hat{t} = (\hat{t}'_1, \dots, \hat{t}'_{r_i}) \in \mathrm{dom}^{r_i}$ zu, für das gilt: 1) $\hat{t}'_\ell = a$ f.a. $\ell \in M_i$, und

2) t' ist das Tupel, das aus $\hat{t}$ entsteht, indem man für jedes $\ell \in M_i$ die ℓ-te Komponente $\hat{t}'_\ell$ löscht.

D ist dann die DB von Schema $\{R_1, \dots, R_m\}$ mit

$$R_i^D := \{\hat{t} : t' \in R_i'^{D'}\} \quad \text{f.a. } i \in [m].$$

Dann ist $|R_i^D| = |R_i'^{D'}| = N_i$ f.a. $i \in [m]$.

Außerdem kann man sich leicht davon überzeugen, dass $Q(D) = Q'(D')$ ist.

Insgesamt gilt also: $|Q(D)| = \prod_{i=1}^{m} N_i^{x(i)}$ und $N_i = |R_i^D|$ f.a. $i \in [m]$.

Beispiel 2.13:

Die Anfrage

$$Q(A,B) \leftarrow E_1(A,B), E_2(B,C), E_3(C,A)$$

ist eine Projektion einer Join-Anfrage — nämlich der Join-Anfrage

$$\tilde{Q}(A,B,C) \leftarrow E_1(A,B), E_2(B,C), E_3(C,A).$$

Natürlich ist $x_1: \{1,2,3\} \to \mathbb{Q}_{\geq 0}$ mit $x_1(1) = x_1(2) = x_1(3) = \frac{1}{2}$ eine fraktionale Kantenüberdeckung von $\tilde{Q}$ und von Q.

Aber gemäß Definition 2.11 ist auch $x_2: \{1,2,3\} \to \mathbb{Q}_{\geq 0}$ mit $x_2(1) = 1$ und $x_2(2) = x_2(3) = 0$ eine fraktionale Kantenüberdeckung von Q (aber nicht von $\tilde{Q}$).

Satz 2.12 liefert, dass für jede DB D gilt:

$$|Q(D)| \leq \prod_{i=1}^{3} N_i^{x_2(i)} = N_1,$$

für $N_i := |R_i^D|$ f.a. $i \in [3]$ ist.

(... was wir bei dieser einfachen Anfrage Q aber auch direkt, ohne Nutzung der AGM-Schranke, sehen können).

Konjunktive Anfragen ohne Konstanten

Definition 2.14

Sei $Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$ eine konjunktive Anfrage, in der keine Konstanten vorkommen (d.h. $cons(Q) = \emptyset$), sei $\{A_1, \ldots, A_n\} = vars(Q)$ und $\{\bar{x}_0\} = \{A_1, \ldots, A_k\}$.

Eine fraktionale Kantenüberdeckung von Q ist eine Abbildung $x : [m] \to \mathbb{Q}_{\geq 0}$, s.d. f.a. $j \in [k]$ gilt:

$$\sum_{\substack{i \in [m] \text{ mit} \\ A_j \in \{\bar{x}_i\}}} x(i) \geq 1$$

Satz 2.15 (Gottlob, Lee, Valiant, 2009)

Sei Q eine konjunktive Anfrage mit $cons(Q) = \emptyset$ der Form $Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$.

(a) Für jede fraktionale Kantenüberdeckung x von Q und jede Datenbank $\mathcal{D}$ und alle $N_1, \ldots, N_m \in \mathbb{N}$ mit $|R_i^{\mathcal{D}}| \leq N_i$ f.a. $i \in [m]$ gilt:

$$|Q(\mathcal{D})| \leq \prod_{i=1}^{m} N_i^{x(i)}.$$

(b) Für jedes $N \in \mathbb{N}$ gibt es natürliche Zahlen $N_1, \ldots, N_m \geq N$, eine Datenbank $\mathcal{D}$ mit $|R_i^{\mathcal{D}}| \leq \sum_{\substack{j \in [m] \text{ mit} \\ R_j = R_i}} N_j$ f.a. $i \in [m]$ und eine fraktionale Kantenüberdeckung x von Q, s.d.

$$|Q(\mathcal{D})| \geq \prod_{i=1}^{m} N_i^{x(i)}.$$

Beweis:

(I) Wir zeigen die Aussage zunächst für selbst-join-freie Anfragen, d.h. für Anfragen, bei denen die Relationssymbole $R_1, \dots, R_m$ paarweise verschieden sind (d.h. $R_i \neq R_j$ f.a. $i, j \in [m]$ mit $i \neq j$).

Für jedes $i \in [0, m]$ sei $\overline{x}_i^{versch}$ die Variablenliste, die man aus $\overline{x}_i$ erhält, indem man Mehrfach-Vorkommen derselben Variablen entfernt. ("versch" steht für "verschieden"). Insbes. ist $\{\overline{x}_i^{versch}\} = \{\overline{x}_i\}$.

Sei $k_i := |\{\overline{x}_i\}| = |\{\overline{x}_i^{versch}\}|$ und sei R_i^{versch} ein neues Relationssymbol der Stelligkeit k_i.

Betrachte die Anfrage

$$Q^{versch}(\overline{x}_0^{versch}) \leftarrow R_1^{versch}(\overline{x}_1^{versch}), \dots, R_m^{versch}(\overline{x}_m^{versch}).$$

Diese Anfrage ist eine Projektion einer Join-Anfrage. (i.S.v. Def. 2.11). Außerdem gilt für jede Abbildung $x: [m] \to \mathbb{Q}_{>0}$:

x ist eine fract. Kantenüberdeckung von Q^{versch} $\iff$ x^* ist eine fract. Kantenüberdeckung von Q

Betrachte eine beliebige DB D vom Schema $\{R_1, \dots, R_m\}$. Sei D' die DB vom Schema $\{R_1^{versch}, \dots, R_m^{versch}\}$ mit $(R_i^{versch})^{D'} := Q_i'(D)$ mit $Q_i'(\overline{x}_i^{versch}) \leftarrow R_i(\overline{x}_i)$.

Dann ist $N_i' := |(R_i^{versch})^{D'}| \leq |R^D|$.

Und es gilt: $|Q(D)| = |Q^{vesch}(D')|$ (*)

Gemäß Satz 2.12(a) gilt für jede fraktionale Kantenüberdeckung x von Q^{vesch} (und Q, wegen (V)), dass $|Q^{vesch}(D')| \leq \prod_{i=1}^{m} N_i^{x(i)}$.

Also gilt:

$$|Q(D)| \overset{(*)}{=} |Q^{vesch}(D')| \leq \prod_{i=1}^{m} N_i^{x(i)} \leq \prod_{i=1}^{m} |R_i^D|^{x(i)} \leq \prod_{i=1}^{m} N_i^{x(i)}$$

f.a. $N_1, \dots, N_m \in \mathbb{N}$ mit $|R_i^D| \leq N_i$ f.a. $i \in [m]$.

Also gilt Aussage (a) für selbst-join-freie Anfragen Q.

Zum Beweis von Aussage (b) für selbst-join-freie Anfragen Q betrachte ein beliebiges $N \in \mathbb{N}$.

Gemäß Satz 2.12(b) gibt es eine $\{R_1^{vesch}, \dots, R_m^{vesch}\}$-DB $\hat{D}$ mit $\hat{N}_i := |(R_i^{vesch})^{\hat{D}}| \geq N$ f.a. $i \in [m]$ und eine fraktionale Kantenüberdeckung x von Q^{vesch} (und Q wg. (V)), s.d. $|Q^{vesch}(\hat{D})| = \prod_{i=1}^{m} \hat{N}_i^{x(i)}$. (**)

Man kann leicht eine DB D vom Schema $\{R_1, \dots, R_m\}$ konstruieren, s.d. $D' = \hat{D}$ und $|R_i^D| = |(R_i^{vesch})^{D'}|$ f.a. $i \in [m]$ ist (Details: Übung). Insgesamt gilt:

$$|Q(D)| \overset{(*)}{=} |Q^{vesch}(D')| = \prod_{i=1}^{m} \hat{N}_i^{x(i)} = \prod_{i=1}^{m} |R_i^D|^{x(i)}.$$

Dies beweist Aussage (b) für selbst-join-freie Anfragen Q.

(II) Wir zeigen nun, dass die Aussage auch für Anfragen Q gilt, die nicht selbst-join-frei sind. Wir ordnen dazu der Anfrage

$$Q(\bar{x}_0) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$$

die selbst-join-freie Anfrage

$$Q'(\bar{x}_0) \leftarrow R'_1(\bar{x}_1), \ldots, R'_m(\bar{x}_m)$$

vom Schema $\{R'_1, \ldots, R'_m\}$ zu, wobei $R'_1, \ldots, R'_m$ paarweise verschiedene Relationssymbole mit $ar(R'_i) = ar(R_i)$ f.a. $i \in [m]$ sind.

Gemäß (I) wissen wir, dass die Aussagen (a) und (b) für die Anfrage Q' gelten. Außerdem gilt für jede Abbildung $x : [m] \to \mathbb{Q}_{\geq 0}$:

(▷) x ist eine fraktionale Kantenüberdeckung von Q $\iff$ x ist eine fraktionale Kantenüberdeckung von Q'

Zum Beweis von (a) ordnen wir jeder $\{R_1, \ldots, R_m\}$-DB D die $\{R'_1, \ldots, R'_m\}$-DB D' mit $R_i^{D'} := R_i^D$ f.a. $i \in [m]$ zu. Klar: $Q(D) = Q'(D')$.

Damit erhalten wir, dass Aussage (a) auch für Q gilt.

Zum Beweis von Aussage (b) für Q betrachte ein beliebiges $N \in \mathbb{N}$. Gemäß (I)(b) gibt es eine $\{R'_1, \ldots, R'_m\}$-DB D' mit $N_i := |R_i^{D'}| \geq N$ f.a. $i \in [m]$ und eine fraktionale Kantenüberdeckung x von Q' (und Q, wegen (▷)), s.d. $|Q'(D')| \geq \prod_{i=1}^{m} N_i^{x(i)}$ (Begründung: Übung!)

Sei $\tilde{D}$ die DB vom Schema $\{R_1, \dots, R_m\}$ mit

$$R_i^{\tilde{D}} := \bigcup_{\substack{j \in [m] \text{ mit} \\ R_j = R_i}} R_j^{D'}.$$

Man kann leicht nachprüfen, dass

$$Q(\tilde{D}) \supseteq Q'(D').$$

Also gilt:

$$|Q(\tilde{D})| \geq |Q'(D')| \geq \prod_{i=1}^{m} N_i^{\chi(i)}.$$

Außerdem gilt $\forall$ a. $i \in [m]$, dass

$$|R_i^{\tilde{D}}| \leq \sum_{\substack{j \in [m]: \\ R_j = R_i}} |R_j^{D'}| = \sum_{\substack{j \in [m] \\ R_j = R_i}} N_j.$$

Dies zeigt, dass Aussage (b) für Q gilt $\square$