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

□