Factorization through the Lens of Information Theory

Dan Suciu
University of Washington

Based on some joint work with Batya Kenig

Motivation

- F-representations that use an F-tree are special kinds of acyclic join dependencies.
- Original motivation of F-trees: derive them from the query that produced the relation.
- This talk: discover an acyclic schema from the instance. Based loosely on [Kenig and Suciu, 2020, Kenig et al., 2020]
- We used information theory to both simplify the schema discovery and allow for noise in the data.

Definitions

Simplified from [Olteanu and Závodný, 2012]

F-Representation and its Schema:

Definition

- $\mathtt{Scm}(\emptyset) = \mathtt{Scm}(\{()\}) = \emptyset$
- $\mathtt{Scm}(\{< A : a >\}) = \{A\}$
- $\mathtt{Scm}(R_1 \times R_2) = \mathtt{Scm}(R_1) \cup \mathtt{Scm}(R_2)$
- $\mathtt{Scm}(R_1 \cup R_2) = = \mathtt{Scm}(R_1) = \mathtt{Scm}(R_2)$

F-Tree of a representation:

Definition

- $\mathtt{FTree}(\emptyset) = \mathtt{FTree}(\{()\}) : \emptyset$
- $\mathtt{FTree}(\{< A : a >\}) = \mathtt{node}(A)$
- $\mathtt{FTree}(R_1 \times R_2) = \mathtt{FTree}(R_1) \cup \mathtt{FTree}(R_2)$
- $\mathtt{FTree}(\bigcup_{a \in \mathrm{Dom}} \{< A : a >\} \times R_a) = \mathtt{node}(A) \cup \mathtt{FTree}(R_a)$

Definitions

Simplified from [Olteanu and Závodný, 2012]

F-Representation and its Schema:

Definition

- $\texttt{Scm}(\emptyset) = \texttt{Scm}(\{()\}) = \emptyset$
- $\texttt{Scm}(\{< A : a >\}) = \{A\}$
- $\texttt{Scm}(R_1 \times R_2) = \texttt{Scm}(R_1) \cup \texttt{Scm}(R_2)$
- $\texttt{Scm}(R_1 \cup R_2) = = \texttt{Scm}(R_1) = \texttt{Scm}(R_2)$

F-Tree of a representation:

Definition

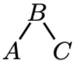
- $\texttt{FTree}(\emptyset) = \texttt{FTree}(\{()\}) : \emptyset$
- $\texttt{FTree}(\{< A : a >\}) = \texttt{node}(A)$
- $\texttt{FTree}(R_1 \times R_2) = \texttt{FTree}(R_1) \cup \texttt{FTree}(R_2)$
- $\texttt{FTree}(\bigcup_{a \in \mathrm{Dom}} \{< A : a >\} \times R_a) = \texttt{node}(A) \cup \texttt{FTree}(R_a)$

Example

Example 2. Consider a relation over schema $\{A, B, C\}$ and domain $\mathcal{D} = \{1, \ldots, 5\}$ that represents the inequalities $A < B < C$. An f-representation of this relation is

$$
\begin{array}{ll}
\langle B:2\rangle \times \langle A:1\rangle & \times\, (\langle C:3\rangle \cup \langle C:4\rangle \cup \langle C:5\rangle) \cup \\
\langle B:3\rangle \times (\langle A:1\rangle \cup \langle A:2\rangle) & \times\, (\langle C:4\rangle \cup \langle C:5\rangle) \cup \\
\langle B:4\rangle \times (\langle A:1\rangle \cup \langle A:2\rangle \cup \langle A:3\rangle) & \times\, \langle C:5\rangle.
\end{array}
$$

over the f-tree



$\square$

Example 3. The relation $\{\langle 1, 1, 1\rangle, \langle 2, 1, 2\rangle\}$ over schema $\{A, B, C\}$ does not admit an f-representation over the f-tree from Example 2, since any such f-representation must essentially be of the form $\langle B:1\rangle \times E_A \times E_C$, where E_A is a union of A-values and E_C is a union of C-values. $\square$

The Factorization Problem

- Motivation in factorized databases: given a conjunctive query Q, compute a factorization for its answer. [Olteanu and Závodný, 2012, Olteanu and Závodný, 2015, Olteanu and Schleich, 2016]:

- Motivation in this talk: given instance R, discover a factorization.

- More generally: discover an acyclic schema.

Acyclic Schemas

R satisfies the *join dependency* $\bowtie\{A_1, \ldots, A_k\}$ if $R = \bowtie_i R[A_i]$

Fact 1 (folklore)

R admits an F-representation over an F-tree T iff it satisfies the acyclic join dependency over the root-to-leave sets of attributes.

$$R[ABCDE] = R[ABC] \bowtie R[ABD] \bowtie R[AE]$$

Fact 2 (folklore)

If R satisfies an acyclic join dependency, then it admits an F-representation over an F-tree derived from the acyclic join.

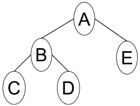
The F-tree is not unique. E.g. $R[AB] \bowtie R[BC] \bowtie R[BD] \bowtie R[AE]$
(Proof: use the recursive def. of acyclicity [Beeri et al., 1983])

Acyclic Schemas

R satisfies the *join dependency* $\bowtie\{A_1, \ldots, A_k\}$ if $R = \bowtie_i R[A_i]$

Fact 1 (folklore)

R admits an F-representation over an F-tree T iff it satisfies the acyclic join dependency over the root-to-leave sets of attributes.



$$R[ABCDE] = R[ABC] \bowtie R[ABD] \bowtie R[AE]$$

Fact 2 (folklore)

If R satisfies an acyclic join dependency, then it admits an F-representation over an F-tree derived from the acyclic join.

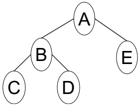
The F-tree is not unique. E.g. $R[AB] \bowtie R[BC] \bowtie R[BD] \bowtie R[AE]$
(Proof: use the recursive def. of acyclicity [Beeri et al., 1983])

Acyclic Schemas

R satisfies the *join dependency* $\bowtie\{A_1, \dots, A_k\}$ if $R = \bowtie_i R[A_i]$

Fact 1 (folklore)

R admits an F-representation over an F-tree T iff it satisfies the acyclic join dependency over the root-to-leave sets of attributes.



$$R[ABCDE] = R[ABC] \bowtie R[ABD] \bowtie R[AE]$$

Fact 2 (folklore)

If R satisfies an acyclic join dependency, then it admits an F-representation over an F-tree derived from the acyclic join.

The F-tree is not unique. E.g. $R[AB] \bowtie R[BC] \bowtie R[BD] \bowtie R[AE]$
(Proof: use the recursive def. of acyclicity [Beeri et al., 1983])

Acyclic Schemas

On the Desirability of Acyclic Database Schemes [Beeri et al., 1983]:

12 equivalent conditions for an acyclic schema $R = \bowtie\{R_1, \ldots, R_k\}$

Condition 3.5. **The join dependency $\bowtie\mathbf{R}$ is equivalent to a set of multivalued dependencies.**

Condition 3.6. **The join dependency $\bowtie\mathbf{R}$ is equivalent to a conflict-free set of multivalued dependencies.**

Multivalued Dependencies

Usual notation $X \twoheadrightarrow Y$.
Better notation $X \twoheadrightarrow Y|Z$ where $XYZ = \texttt{Scm}(R)$

Definition

R satisfies $X \twoheadrightarrow Y|Z$ if $(x, y_1, z_1), (x, y_2, z_2) \in R$ implies $(x, y_1, z_2) \in R$.

Equivalent to $R = R[XY] \bowtie R[XZ]$ when X disjoint from Y, Z.

R has F-tree iff $A \twoheadrightarrow BCD|E$, $AB \twoheadrightarrow C|DE$, and $AB \twoheadrightarrow CE|D$.

Multivalued Dependencies

Usual notation $X \twoheadrightarrow Y$.
Better notation $X \twoheadrightarrow Y|Z$ where $XYZ = \texttt{Scm}(R)$

Definition

R satisfies $X \twoheadrightarrow Y|Z$ if $(x, y_1, z_1), (x, y_2, z_2) \in R$ implies $(x, y_1, z_2) \in R$.

Equivalent to $R = R[XY] \bowtie R[XZ]$ when X disjoint from Y, Z.

R has F-tree iff $A \twoheadrightarrow BCD|E$, $AB \twoheadrightarrow C|DE$, and $AB \twoheadrightarrow CE|D$.

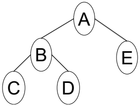
Multivalued Dependencies

Usual notation $X \twoheadrightarrow Y$.
Better notation $X \twoheadrightarrow Y|Z$ where $XYZ = \mathtt{Scm}(R)$

Definition

R satisfies $X \twoheadrightarrow Y|Z$ if $(x, y_1, z_1), (x, y_2, z_2) \in R$ implies $(x, y_1, z_2) \in R$.

Equivalent to $R = R[XY] \bowtie R[XZ]$ when X disjoint from Y, Z.

R has F-tree 

iff $A \twoheadrightarrow BCD|E$, $AB \twoheadrightarrow C|DE$, and $AB \twoheadrightarrow CE|D$.

The MVD Discovery Problem

Given an instance R, discover all MVDs that R satisfies.

- Lots of work on discovering Functional Dependencies and Unique Column Combinations; see references in [Kenig et al., 2020]
- They use *subset property*: if an FD holds in R, then it holds in all subsets; e.g. FastFD [Wyss et al., 2001].
- Subset property fails for MVDs: need new approach.

Information Theory!

Information Theory

Definition

Entropy of a random variable X with n outcomes: $H(X) \stackrel{\text{def}}{=} -\sum_i p_i \log p_i$.

Entropy of joint random variables: $H(XY), H(XYZ), H(YW), \ldots$

Shannon Inequalities:

$$H(Y|X) \stackrel{\text{def}}{=} H(XY) - H(X) \geq 0$$

$$I(Y;Z|X) \stackrel{\text{def}}{=} H(XY) + H(XZ) - H(X) - H(XYZ) \geq 0$$

Called conditional entropy and conditional mutual information

Information Theory

Definition

Entropy of a random variable X with n outcomes: $H(X) \stackrel{\text{def}}{=} -\sum_i p_i \log p_i$.

Entropy of joint random variables: $H(XY), H(XYZ), H(YW), \ldots$

Shannon Inequalities:

$$H(Y|X) \stackrel{\text{def}}{=} H(XY) - H(X) \geq 0$$

$$I(Y;Z|X) \stackrel{\text{def}}{=} H(XY) + H(XZ) - H(X) - H(XYZ) \geq 0$$

Called conditional entropy and conditional mutual information

Information Theory

Definition

Entropy of a random variable X with n outcomes: $H(X) \stackrel{\text{def}}{=} -\sum_i p_i \log p_i$.

Entropy of joint random variables: $H(XY), H(XYZ), H(YW), \ldots$

Shannon Inequalities:

$$H(Y|X) \stackrel{\text{def}}{=} H(XY) - H(X) \geq 0$$
$$I(Y;Z|X) \stackrel{\text{def}}{=} H(XY) + H(XZ) - H(X) - H(XYZ) \geq 0$$

Called conditional entropy and conditional mutual information

The Empirical Probability Distribution

$$p : R \to [0,1] \qquad p(t) = \frac{1}{|R|}, \forall t \in R \qquad H(X_1 \cdots X_n) = \log |R|$$

Random variables $X_1, \ldots, X_n$ correspond to its columns.

$R =$

X	Y	Z	prob
a	b	b	1/6
b	c	c	1/6
b	c	d	1/6
b	d	c	1/6
b	d	d	1/6
c	a	a	1/6

$$H(X) = \frac{1}{6}\left(2\log 6 + \log\frac{6}{4}\right)$$
$$= \frac{\log 2}{6} + \frac{\log 3}{2}$$
$$H(Y) = \ldots$$
$$\ldots$$
$$H(XYZ) = \log 6$$

The Empirical Probability Distribution

$$p : R \to [0,1] \qquad p(t) = \frac{1}{|R|}, \forall t \in R \qquad H(X_1 \cdots X_n) = \log |R|$$

Random variables $X_1, \ldots, X_n$ correspond to its columns.

$R =$

X	Y	Z	prob
a	b	b	1/6
b	c	c	1/6
b	c	d	1/6
b	d	c	1/6
b	d	d	1/6
c	a	a	1/6

$$H(X) = \frac{1}{6}\left(2\log 6 + \log\frac{6}{4}\right) = \frac{\log 2}{6} + \frac{\log 3}{2}$$

$$H(Y) = \ldots$$

$$\ldots$$

$$H(XYZ) = \log 6$$

Data Dependencies through Information Theory

[Lee, 1987]:

Theorem 2: Let $\boldsymbol{R}[\Omega]$ be a relation and $X, Y \subseteq \Omega$. Then any one of the following is equivalent to the FD: $\boxed{X \rightarrow Y}$ in $\boldsymbol{R}[\Omega]$:

$$\text{(i)}\quad H(X) = H(XY), \tag{26}$$

$$\text{(ii)}\quad \boxed{H(Y \mid X) = 0,} \tag{27}$$

$$\text{(iii)}\quad I(X;\, Y) = H(Y). \tag{28}$$

■

Theorem 3: Let $\boldsymbol{R}[\Omega]$ be a relation and $X, Y \subseteq \Omega$, $Z = \Omega - XY$. Then any one of the following is equivalent to the MVD: $\boxed{X \rightarrow\rightarrow Y}$ in $\boldsymbol{R}[\Omega]$.

$$\text{(i)}\quad \boxed{I(Y;\, Z \mid X) = 0.} \tag{29}$$

$$\text{(ii)}\quad H(XYZ) = H(XY) + H(XZ) - H(X). \tag{30}$$

$$\text{(iii)}\quad H(YZ \mid X) = H(Y \mid X) + H(Z \mid X). \tag{31}$$

Approximate Acyclic Schema

Exact MVD $I(Y;Z|X) = 0$: brittle in the presence of noisy data.
Approximate MVD $I(Y;Z|X) \leq \varepsilon$: more robust.
What is an Approximate Acyclic Schema?

Let $\mathbf{A} \stackrel{\text{def}}{=} \{A_1, \ldots, A_n\}$ be an acyclic schema.

$$\mathcal{I}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_i H(A_i | A_i \cap A_{\text{parent}(i)}) - H(\mathtt{Scm}(R))$$

Theorem

[Lee, 1987] R satisfies the acyclic schema $\mathbf{A}$ exactly iff $\mathcal{I}(\mathbf{A}) = 0$.

Definition

[Kenig et al., 2020] R satisfies approximate acyclic schema $\mathbf{A}$ if $\mathcal{I}(\mathbf{A}) \leq \varepsilon$.

Approximate Acyclic Schema

Exact MVD $I(Y;Z|X) = 0$: brittle in the presence of noisy data.
Approximate MVD $I(Y;Z|X) \leq \varepsilon$: more robust.
What is an Approximate Acyclic Schema?
Let $\mathbf{A} \stackrel{\text{def}}{=} \{A_1, \ldots, A_n\}$ be an acyclic schema.

$$\mathcal{I}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_i H(A_i | A_i \cap A_{\text{parent}(i)}) - H(\mathtt{Scm}(R))$$

Theorem

[Lee, 1987] R satisfies the acyclic schema $\mathbf{A}$ exactly iff $\mathcal{I}(\mathbf{A}) = 0$.

Definition

[Kenig et al., 2020] R satisfies approximate acyclic schema $\mathbf{A}$ if $\mathcal{I}(\mathbf{A}) \leq \varepsilon$.

Approximate Acyclic Schema

Exact MVD $I(Y;Z|X) = 0$: brittle in the presence of noisy data.
Approximate MVD $I(Y;Z|X) \leq \varepsilon$: more robust.
What is an Approximate Acyclic Schema?
Let $\mathbf{A} \stackrel{\text{def}}{=} \{A_1, \ldots, A_n\}$ be an acyclic schema.

$$\mathcal{I}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_i H(A_i | A_i \cap A_{\text{parent}(i)}) - H(\mathtt{Scm}(R))$$

Theorem

[Lee, 1987] R satisfies the acyclic schema $\mathbf{A}$ exactly iff $\mathcal{I}(\mathbf{A}) = 0$.

Definition

[Kenig et al., 2020] R satisfies approximate acyclic schema $\mathbf{A}$ if $\mathcal{I}(\mathbf{A}) \leq \varepsilon$.

Approximate Acyclic Schema

Exact MVD $I(Y;Z|X) = 0$: brittle in the presence of noisy data.
Approximate MVD $I(Y;Z|X) \leq \varepsilon$: more robust.
What is an Approximate Acyclic Schema?
Let $\mathbf{A} \stackrel{\text{def}}{=} \{A_1, \ldots, A_n\}$ be an acyclic schema.

$$\mathcal{I}(\mathbf{A}) \stackrel{\text{def}}{=} \sum_i H(A_i|A_i \cap A_{\text{parent}(i)}) - H(\mathtt{Scm}(R))$$

Theorem

[Lee, 1987] R satisfies the acyclic schema $\mathbf{A}$ *exactly iff* $\mathcal{I}(\mathbf{A}) = 0$*.*

Definition

[Kenig et al., 2020] R satisfies approximate acyclic schema $\mathbf{A}$ if $\mathcal{I}(\mathbf{A}) \leq \varepsilon$.

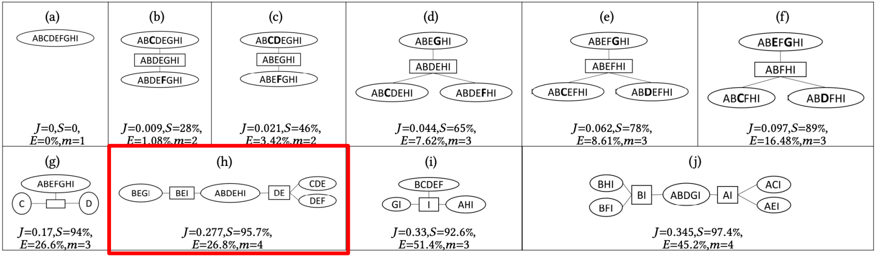


Fig. 10 *The Nursery use case, showing the 10 pareto optimal schemes (out of 415). We encode the 9 attributes as A, B, $\cdots$, I (top). The data does not admit a exact decomposition (a), but we obtain increasingly better schemes (b)-(j) as we increase the J-measure, with increased space savings S, at the cost of increased rate of spurious tuples E; for example, for J = 0.277 the data decomposes into 4 relations, S = 95.7% (see text for the explanation of why it is so high) and E = 26.8%.*

[Kenig et al., 2020]

Example: $\mathcal{I} = 0.277$
Acyclic schema with 4 relations.
Compression $S = 97\%$
Spurious tuples $E = 26.8\%$

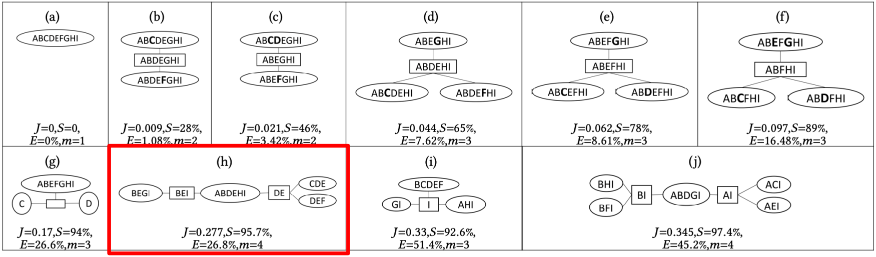


Fig. 10 *The Nursery use case, showing the 10 pareto optimal schemes (out of 415). We encode the 9 attributes as A, B, $\cdots$, I (top). The data does not admit a exact decomposition (a), but we obtain increasingly better schemes (b)-(j) as we increase the J-measure, with increased space savings S, at the cost of increased rate of spurious tuples E; for example, for J = 0.277 the data decomposes into 4 relations, S = 95.7% (see text for the explanation of why it is so high) and E = 26.8%.*

[Kenig et al., 2020]

Example: $\mathcal{I} = 0.277$
Acyclic schema with 4 relations.
Compression $S = 97\%$
Spurious tuples $E = 26.8\%$

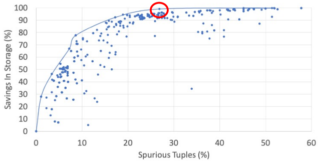


Fig. 11 *All 415 schemes discovered for Nursery. The plot shows the savings S v.s. the spurious tuples E. The line connects the ten pareto-optimal schemes further detailed in Fig. 10. .*

Summary

- F-representations that use an F-tree are special kinds of acyclic join dependencies.
- Original motivation of F-trees: derive them from the query that produced the relation.
- This talk: discover an acyclic schema from the instance.
- We used information theory to both simplify the schema discovery and allow for noise in the data.

THANK YOU!

Summary

- F-representations that use an F-tree are special kinds of acyclic join dependencies.
- Original motivation of F-trees: derive them from the query that produced the relation.
- This talk: discover an acyclic schema from the instance.
- We used information theory to both simplify the schema discovery and allow for noise in the data.

THANK YOU!

Beeri, C., Fagin, R., Maier, D., and Yannakakis, M. (1983).
On the desirability of acyclic database schemes.
J. ACM, 30(3):479–513.

Kenig, B., Mundra, P., Prasaad, G., Salimi, B., and Suciu, D. (2020).
Mining approximate acyclic schemes from relations.
In Maier, D., Pottinger, R., Doan, A., Tan, W., Alawini, A., and Ngo, H. Q., editors, *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 297–312. ACM.

Kenig, B. and Suciu, D. (2020).
Integrity constraints revisited: From exact to approximate implication.
In Lutz, C. and Jung, J. C., editors, *23rd International Conference on Database Theory, ICDT 2020, March 30-April 2, 2020, Copenhagen, Denmark*, volume 155 of *LIPIcs*, pages 18:1–18:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Lee, T. T. (1987).
An information-theoretic analysis of relational databases - part II: information structures of database schemas.
IEEE Trans. Software Eng., 13(10):1061–1072.

Olteanu, D. and Schleich, M. (2016).
Factorized databases.
SIGMOD Rec., 45(2):5–16.

Olteanu, D. and Závodný, J. (2012).
Factorised representations of query results: size bounds and readability.
In Deutsch, A., editor, *15th International Conference on Database Theory, ICDT '12, Berlin, Germany, March 26-29, 2012*, pages 285–298. ACM.

Olteanu, D. and Závodný, J. (2015).
Size bounds for factorised representations of query results.
ACM Trans. Database Syst., 40(1):2:1–2:44.

Wyss, C. M., Giannella, C., and Robertson, E. L. (2001).

Fastfds: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances - extended abstract.
In Kambayashi, Y., Winiwarter, W., and Arikawa, M., editors, *Data Warehousing and Knowledge Discovery, Third International Conference, DaWaK 2001, Munich, Germany, September 5-7, 2001, Proceedings*, volume 2114 of *Lecture Notes in Computer Science*, pages 101–110. Springer.