# Spanner Evaluation over SLP-Compressed Documents

**Markus L. Schmid**, Nicole Schweikardt

HU Berlin, Germany

PODS 2021

# Document Spanners

# Document Spanners

- $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \ldots\}$ is an alphabet.
- $\mathcal{X} = \{\mathsf{x}, \mathsf{y}, \mathsf{z}, \ldots\}$ is a set of variables.
- D is a document over $\Sigma$.

# Document Spanners

- $\Sigma = \{a, b, c, \ldots\}$ is an alphabet.
- $\mathcal{X} = \{x, y, z, \ldots\}$ is a set of variables.
- D is a document over $\Sigma$.

$D = \texttt{a b b a b c c a b c}$

# Document Spanners

- $\Sigma = \{a, b, c, \ldots\}$ is an alphabet.
- $\mathcal{X} = \{x, y, z, \ldots\}$ is a set of variables.
- D is a document over $\Sigma$.

$D = \texttt{a b b a b c c a b c} \quad \Longrightarrow$

| x | y | z |
|---|---|---|
| $[2,5\rangle$ | $[4,7\rangle$ | $[1,10\rangle$ |
| $[3,5\rangle$ | $[5,8\rangle$ | $[4,7\rangle$ |
| $[1,3\rangle$ | $[3,10\rangle$ | $[2,4\rangle$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

# Document Spanners

- $\Sigma = \{a, b, c, \ldots\}$ is an alphabet.
- $\mathcal{X} = \{x, y, z, \ldots\}$ is a set of variables.
- D is a document over $\Sigma$.

$D = a\,b\,b\,a\,b\,c\,c\,a\,b\,c \quad \Longrightarrow$

| x | y | z |
|---|---|---|
| $[2, 5\rangle$ | $[4, 7\rangle$ | $[1, 10\rangle$ |
| $[3, 5\rangle$ | $[5, 8\rangle$ | $[4, 7\rangle$ |
| $[1, 3\rangle$ | $[3, 10\rangle$ | $[2, 4\rangle$ |
| ⋮ | ⋮ | ⋮ |

# Document Spanners

- $\Sigma = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \ldots\}$ is an alphabet.
- $\mathcal{X} = \{\mathsf{x}, \mathsf{y}, \mathsf{z}, \ldots\}$ is a set of variables.
- D is a document over $\Sigma$.

$$D = \mathtt{a\,b\,b\,a\,b\,c\,c\,a\,b\,c} \quad \Longrightarrow$$

| x | y | z |
|---|---|---|
| $[2, 5\rangle$ | $[4, 7\rangle$ | $[1, 10\rangle$ |
| $[3, 5\rangle$ | $[5, 8\rangle$ | $[4, 7\rangle$ |
| $[1, 3\rangle$ | $[3, 10\rangle$ | $[2, 4\rangle$ |
| ⋮ | ⋮ | ⋮ |

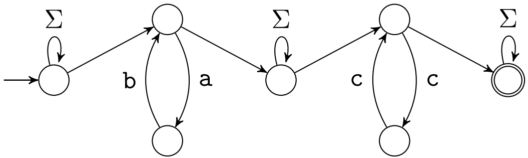
# Document Spanners

- $\Sigma = \{a, b, c, \ldots\}$ is an alphabet.
- $\mathcal{X} = \{x, y, z, \ldots\}$ is a set of variables.
- D is a document over $\Sigma$.

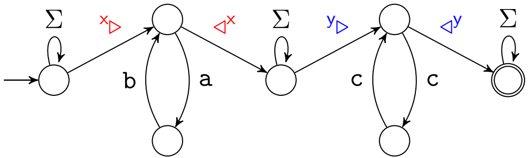
$D = \texttt{a b b a b c c a b c} \quad \Longrightarrow$

| x | y | z |
|---|---|---|
| $[2, 5\rangle$ | $[4, 7\rangle$ | $[1, 10\rangle$ |
| $[3, 5\rangle$ | $[5, 8\rangle$ | $[4, 7\rangle$ |
| $[1, 3\rangle$ | $[3, 10\rangle$ | $[2, 4\rangle$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

# Regular (Document) Spanners



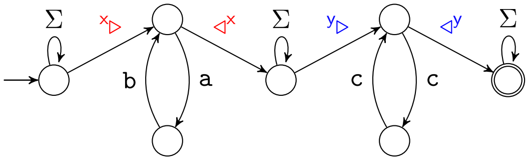
# Regular (Document) Spanners




Meta-symbols for variables $x, y, \ldots \in \mathcal{X}$:

${}^{x}\triangleright \ldots \triangleleft^{x}$ (start and end position of span extracted by x),
${}^{y}\triangleright \ldots \triangleleft^{y}$ (start and end position of span extracted by y).

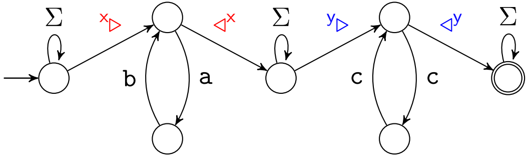
# Regular (Document) Spanners




Meta-symbols for variables $\mathsf{x}, \mathsf{y}, \ldots \in \mathcal{X}$:

$^{\mathsf{x}}\!\triangleright \ldots \triangleleft^{\mathsf{x}}$ (start and end position of span extracted by x),
$^{\mathsf{y}}\!\triangleright \ldots \triangleleft^{\mathsf{y}}$ (start and end position of span extracted by y).

# Regular (Document) Spanners

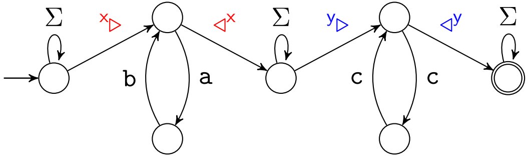



Meta-symbols for variables $\mathsf{x}, \mathsf{y}, \ldots \in \mathcal{X}$:

$^{\mathsf{x}}\triangleright \ldots \triangleleft^{\mathsf{x}}$ (start and end position of span extracted by $\mathsf{x}$),
$^{\mathsf{y}}\triangleright \ldots \triangleleft^{\mathsf{y}}$ (start and end position of span extracted by $\mathsf{y}$).

$\mathsf{D} = \mathtt{abbababccca}$

# Regular (Document) Spanners



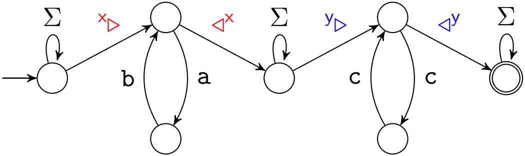

Meta-symbols for variables $\mathsf{x}, \mathsf{y}, \ldots \in \mathcal{X}$:

$^{\mathsf{x}}\triangleright \ldots \triangleleft^{\mathsf{x}}$ (start and end position of span extracted by x),
$^{\mathsf{y}}\triangleright \ldots \triangleleft^{\mathsf{y}}$ (start and end position of span extracted by y).

$\mathsf{D} = \mathtt{abbababccca}$

$\mathtt{abb}\,{}^{\mathsf{x}}\triangleright\,\mathtt{ab}\triangleleft^{\mathsf{x}}\,\mathtt{ab}\,{}^{\mathsf{y}}\triangleright\,\mathtt{cc}\triangleleft^{\mathsf{y}}\,\mathtt{ca}$

# Regular (Document) Spanners



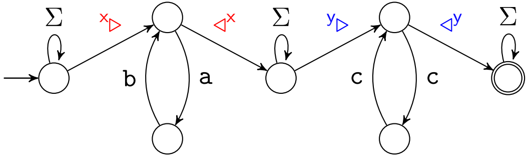

Meta-symbols for variables $x, y, \ldots \in \mathcal{X}$:

${}^{x}\triangleright \ldots \triangleleft^{x}$ (start and end position of span extracted by x),
${}^{y}\triangleright \ldots \triangleleft^{y}$ (start and end position of span extracted by y).

$D = \texttt{abbababcccca}$

$\texttt{abb}\,{}^{x}\triangleright\,\texttt{ab}\triangleleft^{x}\texttt{ab}\,{}^{y}\triangleright\,\texttt{cc}\triangleleft^{y}\texttt{ca} \implies ([4,6\rangle, [8,10\rangle)$

# Regular (Document) Spanners




Meta-symbols for variables $x, y, \ldots \in \mathcal{X}$:

${}^{x}\triangleright \ldots \triangleleft^{x}$ (start and end position of span extracted by x),
${}^{y}\triangleright \ldots \triangleleft^{y}$ (start and end position of span extracted by y).

$D = \texttt{abbababccca}$

$\texttt{abb}{}^{x}\triangleright\texttt{ab}\triangleleft^{x}\texttt{ab}{}^{y}\triangleright\texttt{cc}\triangleleft^{y}\texttt{ca} \implies ([4,6\rangle, [8,10\rangle)$
$\texttt{abb}{}^{x}\triangleright\texttt{abab}\triangleleft^{x}\texttt{c}{}^{y}\triangleright\texttt{cc}\triangleleft^{y}\texttt{a} \implies ([4,8\rangle, [9,11\rangle)$

$\vdots$

# Regular Spanners – Notations

$[\![M]\!]$ denotes the spanner represented by an NFA $M$.

# Regular Spanners – Notations

$[\![M]\!]$ denotes the spanner represented by an NFA $M$.

$[\![M]\!](\mathrm{D})$ denotes the span-relation extracted from a document D.

# Regular Spanners – Notations

$[\![M]\!]$ denotes the spanner represented by an NFA $M$.

$[\![M]\!](\mathrm{D})$ denotes the span-relation extracted from a document D.

A spanner $S$ is a *regular spanner* if $S = [\![M]\!]$ for some NFA $M$.

# Results About Regular Spanners

Introduced by Fagin et al. PODS 2013, JACM 2015.

# Results About Regular Spanners

Introduced by Fagin et al. PODS 2013, JACM 2015.

Since then intensely studied; many positive results.

# Results About Regular Spanners

Introduced by Fagin et al. PODS 2013, JACM 2015.

Since then intensely studied; many positive results.

A major result: linear preprocessing and constant delay enumeration (Florenzano et al. PODS 2018, Amarilli et al. ICDT 2019).

# Approach of this Paper

| Spanner Evaluation over Compressed Documents | |
| --- | --- |
| Input: | A spanner represented by an NFA $M$,<br>a document D given in a **compressed form*** $\mathcal{S}$. |
| Task: | Evaluate $M$ on D (e. g., model checking, computing or enumerating $[\![M]\!](\mathrm{D})$)... but **without** decompressing $\mathcal{S}$. |

*Compression Scheme: Straight-Line Programs (SLPs).

# Straight-Line Programs

# Straight-Line Programs

## Straight-Line Program

A straight-line program for document D is a context-free grammar $\mathcal{S}$ that describes the language $\{D\}$.

# Straight-Line Programs

## Example

Let $\mathcal{S}$ have rules

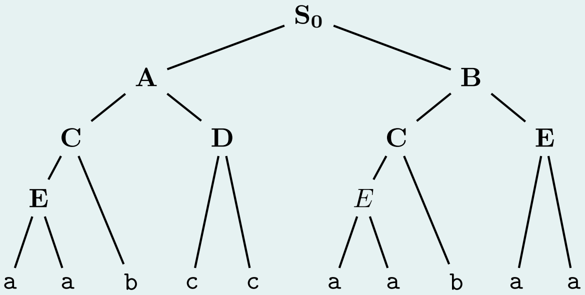
$$S_0 \to AB, \qquad A \to CD, \qquad B \to CE\,,$$
$$C \to E\mathtt{b}, \qquad D \to \mathtt{cc}, \qquad E \to \mathtt{aa}$$

# Straight-Line Programs

## Example

Let $\mathcal{S}$ have rules

$$S_0 \to AB, \quad A \to CD, \quad B \to CE\,,$$
$$C \to E\texttt{b}, \quad D \to \texttt{cc}, \quad E \to \texttt{aa}$$

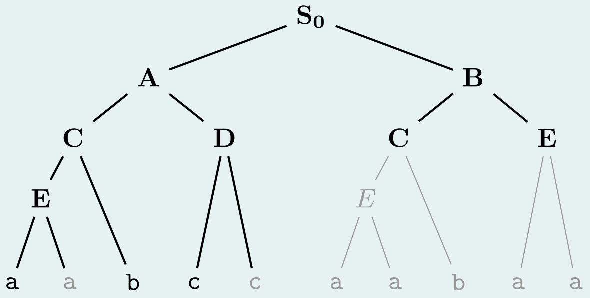


# Straight-Line Programs

## Example

Let $\mathcal{S}$ have rules

$$S_0 \to AB, \quad A \to CD, \quad B \to CE\,,$$
$$C \to E\texttt{b}, \quad D \to \texttt{cc}, \quad E \to \texttt{aa}$$



# Good Properties of SLPs

SLPs are intensely researched in TCS and many things are known:

- Exponential compression rates.

# Good Properties of SLPs

SLPs are intensely researched in TCS and many things are known:

- Exponential compression rates.
- SLPs are mathematically easy to handle ($\Rightarrow$ good for theoretical considerations).

# Good Properties of SLPs

SLPs are intensely researched in TCS and many things are known:

- Exponential compression rates.
- SLPs are mathematically easy to handle ($\Rightarrow$ good for theoretical considerations).
- High practical relevance (SLPs cover many practically applied dictionary-based compression schemes).

# Good Properties of SLPs

SLPs are intensely researched in TCS and many things are known:

- Exponential compression rates.
- SLPs are mathematically easy to handle ($\Rightarrow$ good for theoretical considerations).
- High practical relevance (SLPs cover many practically applied dictionary-based compression schemes).
- Many approximations and heuristics exist that efficiently compute small SLPs.

# Good Properties of SLPs

SLPs are intensely researched in TCS and many things are known:

- Exponential compression rates.
- SLPs are mathematically easy to handle ($\Rightarrow$ good for theoretical considerations).
- High practical relevance (SLPs cover many practically applied dictionary-based compression schemes).
- Many approximations and heuristics exist that efficiently compute small SLPs.
- SLPs are suitable for algorithmics on compressed strings: comparison, pattern matching, membership in a regular language, retrieving subwords, etc.

# Research Task

| Spanner Evaluation over SLP-Compressed Documents | |
| --- | --- |
| Input: | A spanner represented by an NFA $M$,<br>an SLP $\mathcal{S}$ for a document D. |
| Non-emptiness: | Check whether $[\![M]\!](\mathrm{D}) \neq \emptyset$. |
| Model Checking: | Check whether $t \in [\![M]\!](\mathrm{D})$ for a span-tuple $t$. |
| Computation: | Compute $[\![M]\!](\mathrm{D})$. |
| Enumeration: | Enumerate $[\![M]\!](\mathrm{D})$. |

# Results

# Results

**Theorem (Data Complexity)**

| | |
|---|---|
| Non-emptiness: | $O(\text{size}(\mathcal{S}))$ |
| Model Checking: | $O(\text{size}(\mathcal{S}))$ |
| Computation: | $O(\text{size}(\mathcal{S}) \cdot \text{size}([\![M]\!](D)))$ |
| Enumeration: | preprocessing time $O(\text{size}(\mathcal{S}))$ and delay $O(\log(\mid D \mid))$. |

# Results

**Theorem (Data Complexity)**

| | |
|---|---|
| Non-emptiness: | $O(\text{size}(\mathcal{S}))$ |
| Model Checking: | $O(\text{size}(\mathcal{S}))$ |
| Computation: | $O(\text{size}(\mathcal{S}) \cdot \text{size}([\![M]\!](D)))$ |
| Enumeration: | preprocessing time $O(\text{size}(\mathcal{S}))$ and delay $O(\log(\lvert D \rvert))$. |

Two remarks about combined-complexity:

- Sets of markers ("$\{{}^{x}\triangleright, \triangleleft^{y}, {}^{z}\triangleright\}$") as arc labels of the NFA (a.k.a. *extended variable-set automata*), which makes the NFA larger.

# Results

**Theorem (Data Complexity)**

| | |
|---|---|
| Non-emptiness: | $O(\text{size}(\mathcal{S}))$ |
| Model Checking: | $O(\text{size}(\mathcal{S}))$ |
| Computation: | $O(\text{size}(\mathcal{S}) \cdot \text{size}([\![M]\!](D)))$ |
| Enumeration: | preprocessing time $O(\text{size}(\mathcal{S}))$ and delay $O(\log(\mid D \mid))$. |

Two remarks about combined-complexity:

- Sets of markers ("$\{{}^{x}\triangleright, \triangleleft^{y}, {}^{z}\triangleright\}$") as arc labels of the NFA (a.k.a. *extended variable-set automata*), which makes the NFA larger.
- For the enumeration result, we require the NFA also to be deterministic.

# Proof Sketches

# Non-Emptiness, Model-Checking and Computation

Follows (non-trivial) from known results about the regular membership problem for SLP-compressed words:

| | |
|---|---|
| Non-emptiness: | $\mathrm{O}(\mathsf{size}(\mathcal{S}))$ |
| Model Checking: | $\mathrm{O}(\mathsf{size}(\mathcal{S}))$ |
| Computation: | $\mathrm{O}(\mathsf{size}(\mathcal{S}) \cdot \mathsf{size}(\llbracket M \rrbracket(\mathrm{D})))$ |

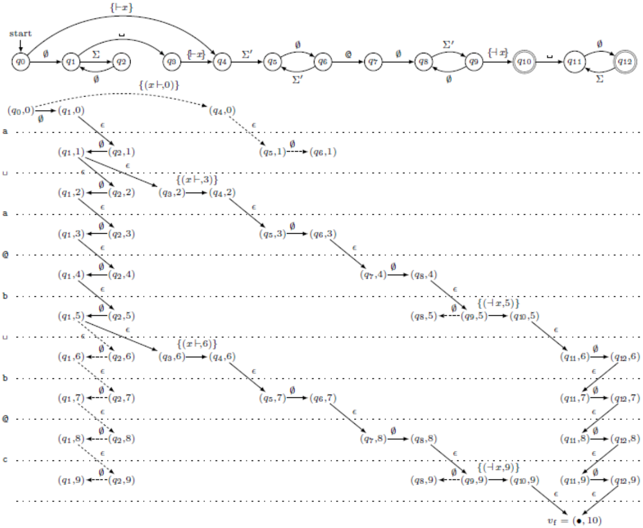
# Non-Emptiness, Model-Checking and Computation

Follows (non-trivial) from known results about the regular membership problem for SLP-compressed words:

| | |
|---|---|
| Non-emptiness: | $\mathsf{O}(\mathsf{size}(\mathcal{S}))$ |
| Model Checking: | $\mathsf{O}(\mathsf{size}(\mathcal{S}))$ |
| Computation: | $\mathsf{O}(\mathsf{size}(\mathcal{S}) \cdot \mathsf{size}(\llbracket M \rrbracket(\mathsf{D})))$ |

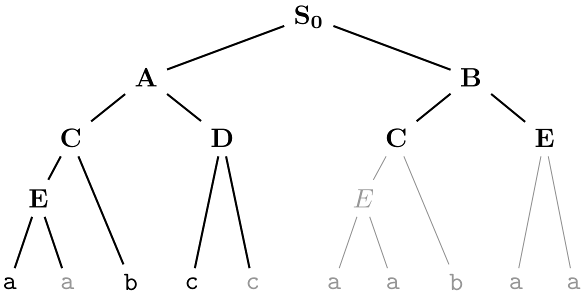
In the following: Sketch for Enumeration!

# Non-Compressed Enumeration
# (Florenzano et al. PODS 2018, Amarilli et al. ICDT 2019)

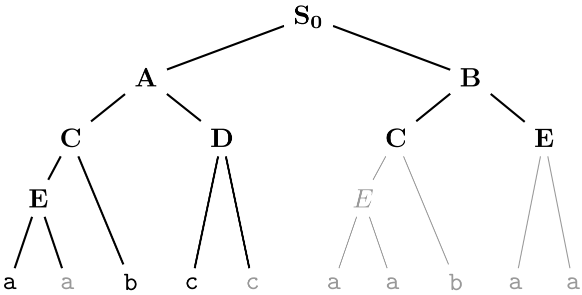


[Amarilli et al. ICDT 2019, SIGMOD Rec., 2020]

# Marking SLPs

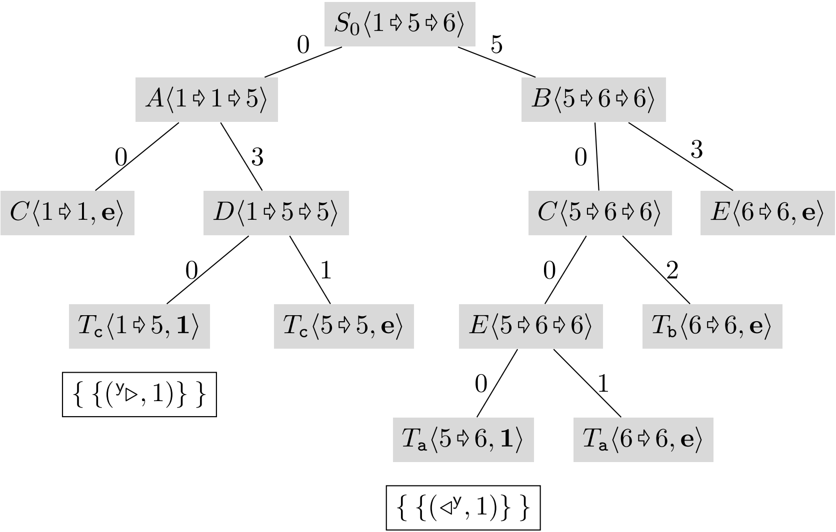


# Marking SLPs




$\Longrightarrow$ enumerate *partially decompressed* SLPs.

# Enumerating Partially Decompressed SLPs



# Balancing SLPs

SLP Balancing Theorem, Ganardi, Jez and Lohrey, FOCS 2019:

**Theorem**

Any given SLP $\mathcal{S}$ can be *balanced*$^*$ in linear time.

$^*$ $\mathrm{depth}(\mathcal{S}) = O(\log(|\,D\,|))$.

# Future Work

Dynamic setting with updates!

Thank you very much for your attention.