Context-Freeness of Word-MIX Languages

(A paper accepted for DLT 2020)

Ryoma Sin'ya (Akita University)

DLT 2021 Aug 16.

秋田大学
Akita University

Outline

1. **Introduction**
2. Background
3. Main Result
4. Conclusion

A word puzzle

For two words $u, v \in A^*$, we denote by $|u|_v$ the number of occurrences of v in u as a factor.

Example : $|ababa|_{aba} = 2$

Question 1: Is $Z_2 = \{w \in \{a,b\}^* \mid |w|_{ab} = |w|_{ba}\}$ regular?

Answer: Yes! $Z_2 = \varepsilon + a(a^*b^+a)^* + b(b^*a^+b)^*$, indeed.

A word puzzle

For two words $u, v \in A^*$, we denote by $|u|_v$ the number of occurrences of v in u as a factor.

Example : $|ababa|_{aba} = 2$

Question 2: Is $Z_3 = \{w \in \{a, b, c\}^* \mid |w|_{ab} = |w|_{ba}\}$ regular?

Answer: No! The regularity depends not only parameter words, but also the alphabet.

A word puzzle

For two words $u, v \in A^*$, we denote by $|u|_v$ the number of occurrences of v in u as a factor.

Example： $|ababa|_{aba} = 2$

In general, the language $L_A(w_1, \ldots, w_k) = \{w \in A^* \mid |w|_{w_1} = \cdots = |w|_{w_k}\}$, which we call **Word-MIX language** (defined by k *parameter words*), can be non-context-free.

$\mathrm{MIX} = L_A(a, b, c)$ where $A = \{a, b, c\}$ for example.

This talk

Theorem (S. 2020)
For given parameter words $w_1, \ldots, w_k \in A^*$ $(k \geq 2)$, the following problems are decidable:

1: Is $L_A(w_1, \ldots, w_k) = \{w \in A^* \mid |w|_{w_1} = \cdots = |w|_{w_k}\}$ regular?

2: Is $L_A(w_1, \ldots, w_k)$ context-free?

Outline

1. Introduction
2. **Background**
3. Main Result
4. Conclusion

The 22nd International Conference on Developments in Language Theory (DLT 2018) will be held in Tokyo, Japan on September 10-14th, 2018. The conference is organized by Algorithmic "Oritatami" Self-Assembly Lab, The University of Electro-Communications (UEC) as part of UEC 100th Anniversary Commemorative Event.

The registration is NOW OPEN!!

Important Dates

Deadline for submission

~~April 30th, 2018 (AoE)~~
May 7th, 2018 (firm AoE)

Notification to authors

~~June 1st, 2018~~
~~June 6th, 2018 (firm)~~
June 8th, 2018 (firm)

Camera-ready version

~~June 20th, 2018~~
June 22nd, 2018 (firm)

Background

Theorem [Lidbetter et al. DLT 2018]:
The regularity of $L_A(w_1, w_2)$ is decidable in time $O(|A||w_1||w_2|)$.

Theorem [Lidbetter et al. DLT 2018]:
$L_A(w_1, w_2)$ is finite if and only if $|A| = 1$ and $x \neq y$.

What about the finiteness for general case?

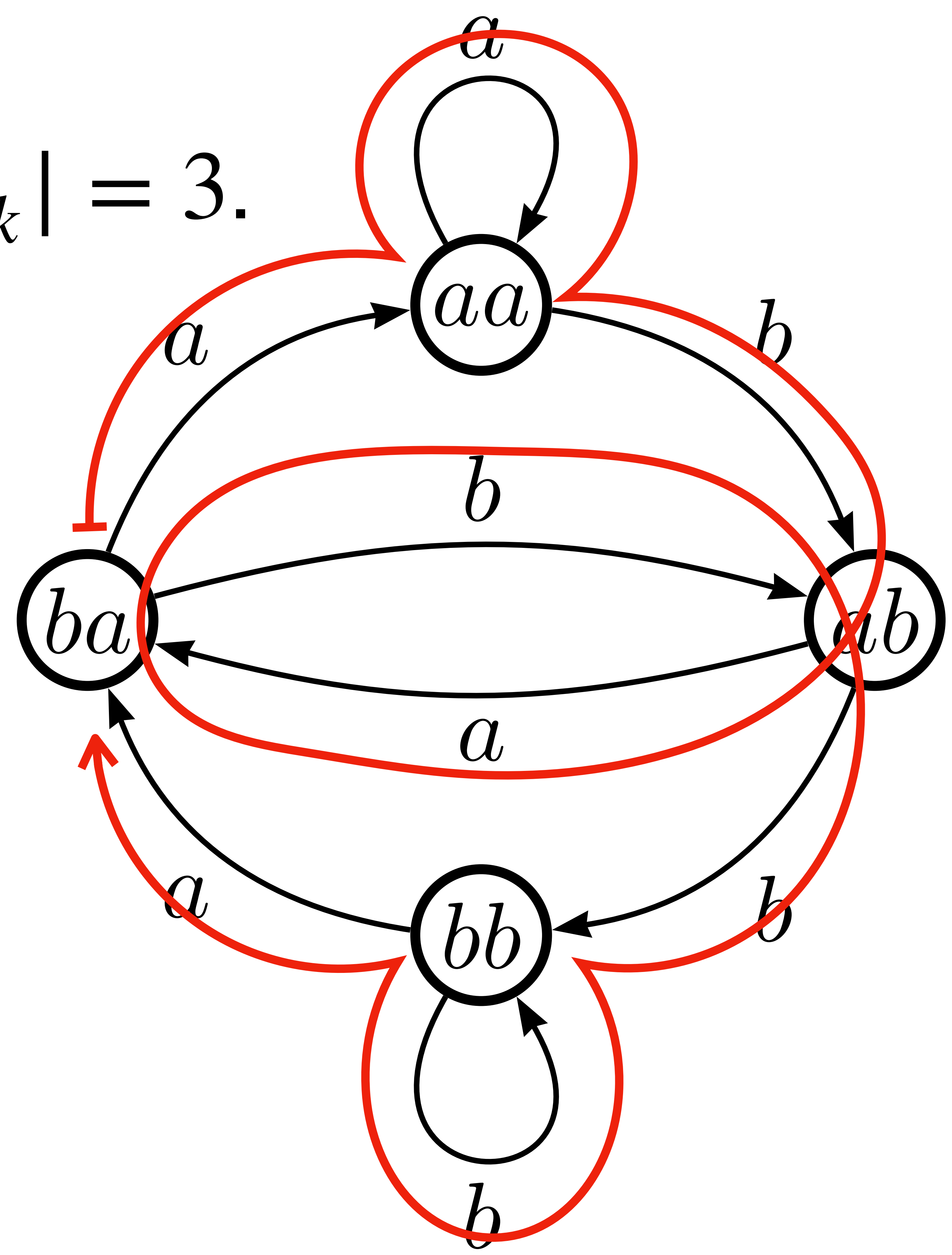
Theorem [de Bruijn 1946]:
If $|w_1| = \cdots = |w_k|$ then $L_A(w_1, \ldots, w_k)$ is infinite.

de Bruijn graph

Proof:
Consider the case $A = \{a, b\}$ and $|w_1| = \cdots = |w_k| = 3$.

Consider the two dimensional *de Bruijn graph*
$\mathcal{G}_A^2 = (A^2, E)$ where
$E = \{(a_1a_2, a_3, a_2a_3) \mid a_1, a_2, a_3 \in A\}$.

$\mathcal{G}_A^2$ has an Eulerian cycle as follows:



The word corresponding this cycle is
$w = baaababbba$. And $w^* \subseteq L_A(w_1, \ldots, w_k)$ holds since $|w^n|_v = n$ for any $n \geq 0$ and $v \in A^3$.

Theorem [de Bruijn 1946]:
If $|w_1| = \cdots = |w_k|$ then $L_A(w_1, \ldots, w_k)$ is infinite.

Background

Theorem [Lidbetter et al. DLT 2018]:
The regularity of $L_A(w_1, w_2)$ is decidable in time $O(|A||w_1||w_k|)$.

Theorem [Lidbetter et al. DLT 2018]:
$L_A(w_1, w_2)$ is finite if and only if $|A| = 1$ and $x \neq y$.

What about the regularity for general case?

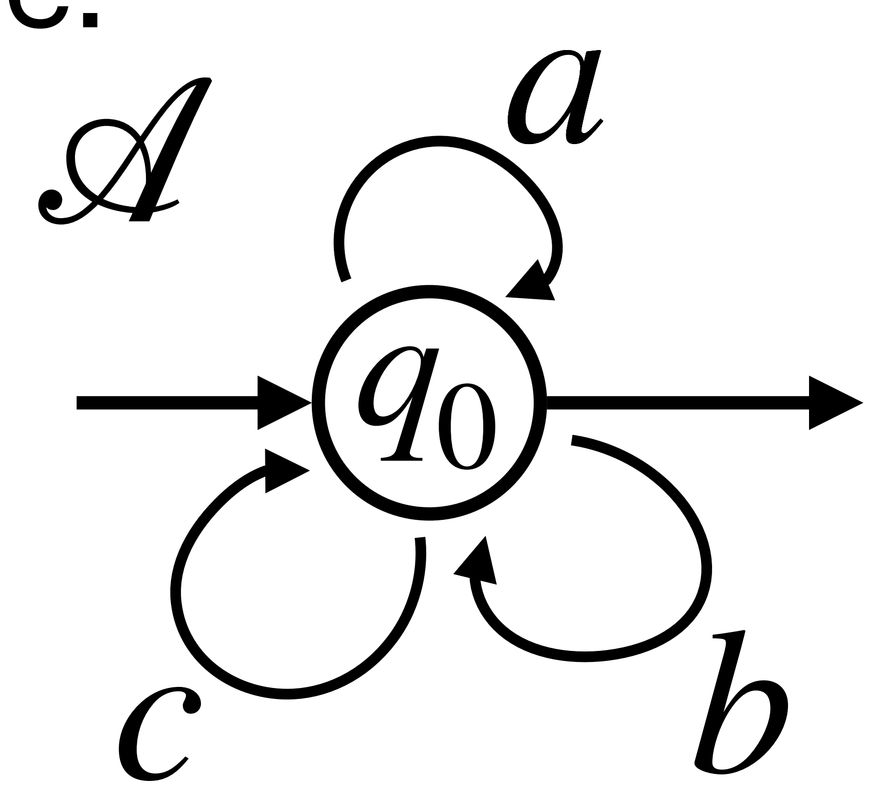
Corollary of [Cadilhac et al. DLT 2012]:
The regularity for WMIX languages is decidable.

Constrained automaton [Klaedtke-Rueß 2003]

- A constrained automaton is a pair $(\mathscr{A}, S)$ where $\mathscr{A}$ is usual finite automaton and $S \subseteq \mathbb{N}^d$ is a semi-linear set whose dimension d equals the number of transition rules in $\mathscr{A}$ (assume that the transition rules are linearly ordered).

- $(\mathscr{A}, S)$ accepts $w \Leftrightarrow$ there is an accepting run ρ in $\mathscr{A}$ labelled by w and its Parikh image $\Phi(\rho)_i = \#$ of occurrences of i-th transition rule in ρ is in S.

Example:

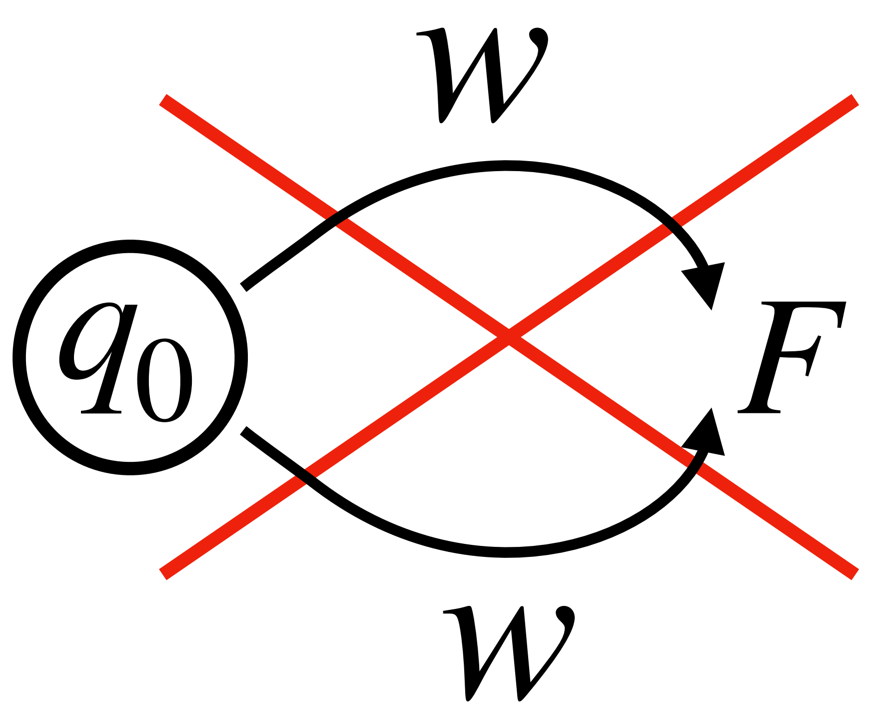


$L(\mathscr{A}) = \{a, b, c\}^*$

But $L((\mathscr{A}, S)) = \mathrm{MIX}$ for $S = \{(n, n, n) \mid n \in \mathbb{N}\} \subseteq \mathbb{N}^3$.

Unambiguous CA (UnCA)

- A constraint automaton $(\mathscr{A}, S)$ is **unambiguous** if $\mathscr{A}$ is unambiguous, i.e., for each word accepted by $\mathscr{A}$ there is only one accepting run.



No different accepting runs!

- $L_A(w_1, \ldots, w_k)$ can be recognised by an UnCA whose underlying automaton is the de Bruijn graph of dimension $\max\{\, |w_i| \mid 1 \leq i \leq k\}$.

Theorem [Cadilhac et al. DLT 2012]: The regularity for UnCA is decidable.

Remark: their proof uses the regularity condition of bounded languages [GS1966].

Background

Theorem [Lidbetter et al. DLT 2018]:
The regularity of $L_A(w_1, w_2)$ is decidable in time $O(|A||w_1||w_k|)$.

Theorem [Lidbetter et al. DLT 2018]:
$L_A(w_1, w_2)$ is finite if and only if $|A| = 1$ and $x \neq y$.

What about the regularity for general case?

Corollary of [Cadilhac et al. DLT 2012]:
The regularity for WMIX languages is decidable.

What about the **context-freeness** for general case?

Outline

1. Introduction
2. Background
3. **Main Result**
4. Conclusion

Main result

Theorem (S. 2020)
For given parameter words $w_1, \ldots, w_k \in A^*$ $(k \geq 2)$, the following problems are decidable:

1: Is $L_A(w_1, \ldots, w_k) = \{w \in A^* \mid |w|_{w_1} = \cdots = |w|_{w_k}\}$ regular?

2: Is $L_A(w_1, \ldots, w_k)$ context-free?

(very rough) Sketch of the proof

For given parameter words $w_1, \ldots, w_k$, we can calculate finitely many (basis of) vectors spaces $\mathbf{V}_1, \ldots, \mathbf{V}_n \subseteq \mathbb{Q}^k$ *from the de Bruijn graph of dimension* $\max\{\, |w_i| \mid 1 \leq i \leq k\}$.

Define the **dimension** of $L = L_A(w_1, \ldots, w_k)$ as

$$\dim(L) = \max\{\dim(\mathbf{V}_i) \mid 1 \leq i \leq n\}.$$

Roughly speaking, $\dim(L) - 1$ represents **the minimum # of registers** that an automaton with such registers can recognize the WMIX language L.
For example, $\dim(L_A(a, b)) = 2$ and it can be recognised by a pushdown automaton (stack can be used as a single register).

We showed: (1) L is regular if and only if $\dim(L) \leq 1$.
(2) L is context-free if and only if $\dim(L) \leq 2$.

MIX: a simplest example of dimension three

$\mathrm{MIX} = L_A(a, b, c)$ where $A = \{a, b, c\}$.

Consider the set of **simple** cycles $S = \{\gamma_1, \gamma_2, \gamma_3\}$:

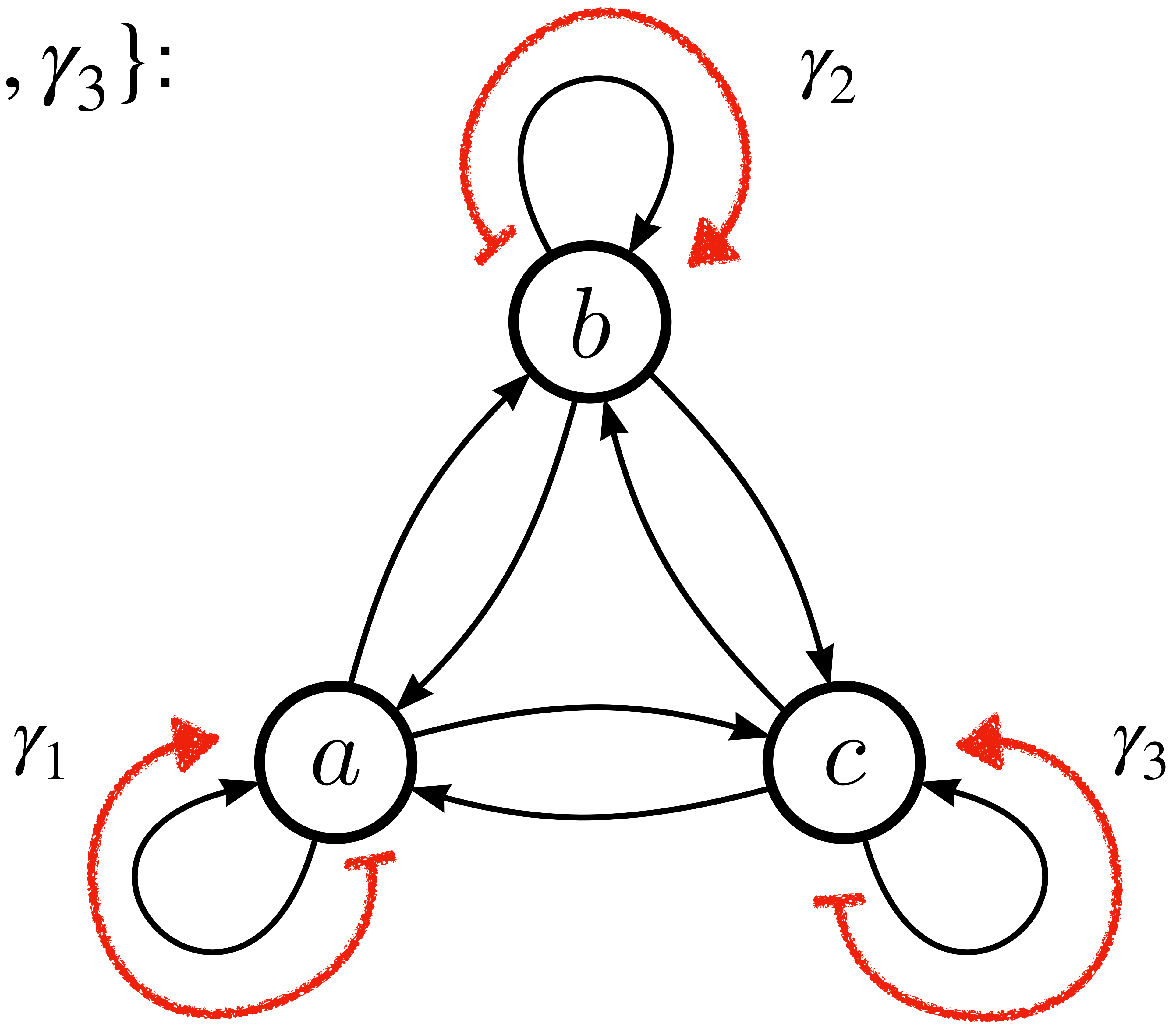
This set S has the following property:

(★) for any $n \geq 0$ there is a word $w \in \mathrm{MIX}$

such that the walk corresponding w in this de Bruijn graph passes each cycle in S **more than** n **times**, respectively.

For example $a^3b^3c^3$ passes $\gamma_1, \gamma_2, \gamma_3$ two times, respectively.

We call such S **pumpable** (in MIX).



MIX: a simplest example of dimension three

$\mathrm{MIX} = L_A(a, b, c)$ where $A = \{a, b, c\}$.

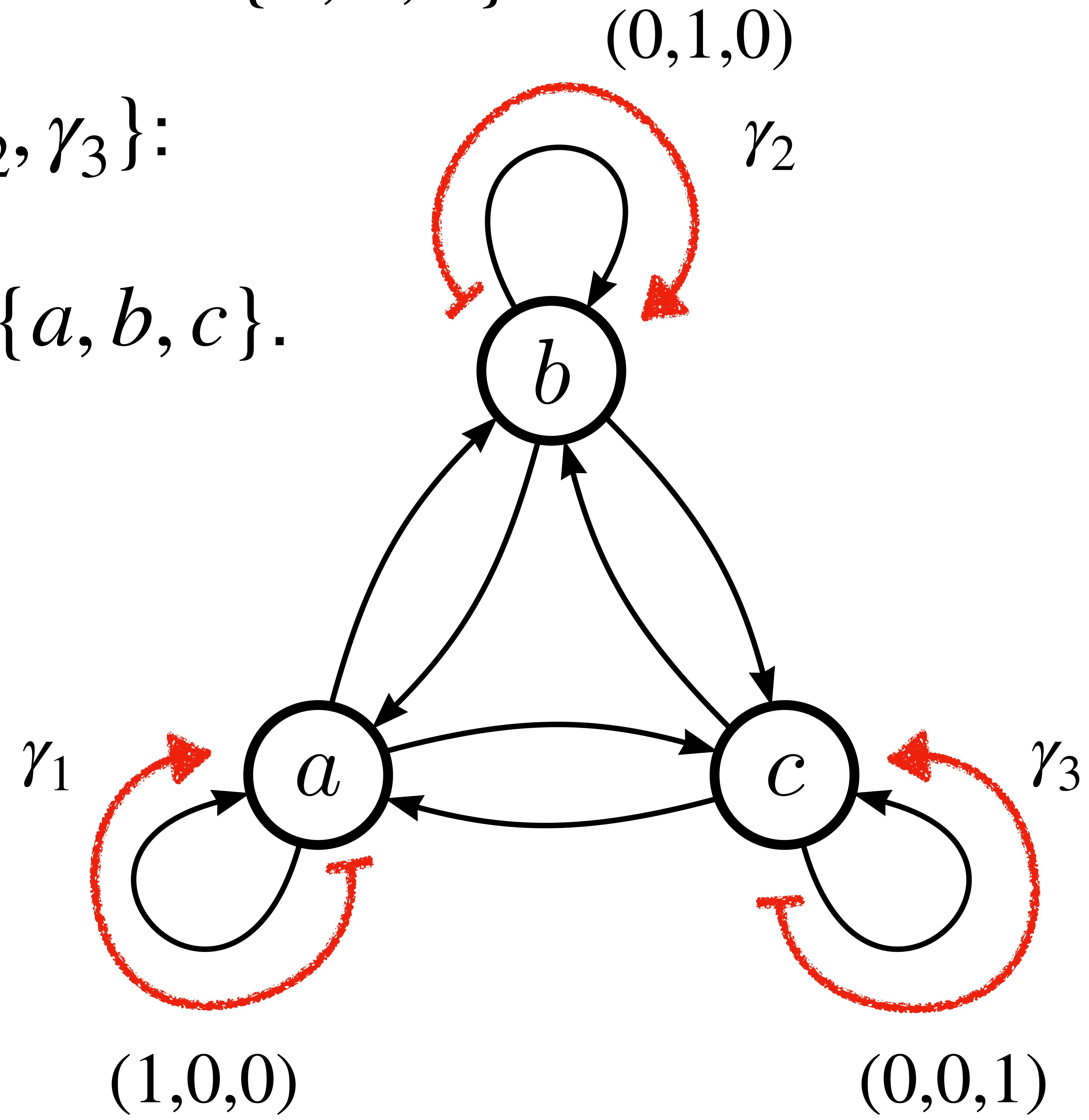
Consider the set of **simple** cycles $S = \{\gamma_1, \gamma_2, \gamma_3\}$:

The set of words corresponding to S is $W = \{a, b, c\}$.

Now the parameter words are a, b and c.
We can associate its ***occurrence vectors*** in W as $\{(|w|_a, |w|_b, |w|_c) \mid w \in W\}$
$= \{(1,0,0), (0,1,0), (0,0,1)\}$.

Define $\dim(S)$ as the maximum # of linearly independent occurrence vectors in W.

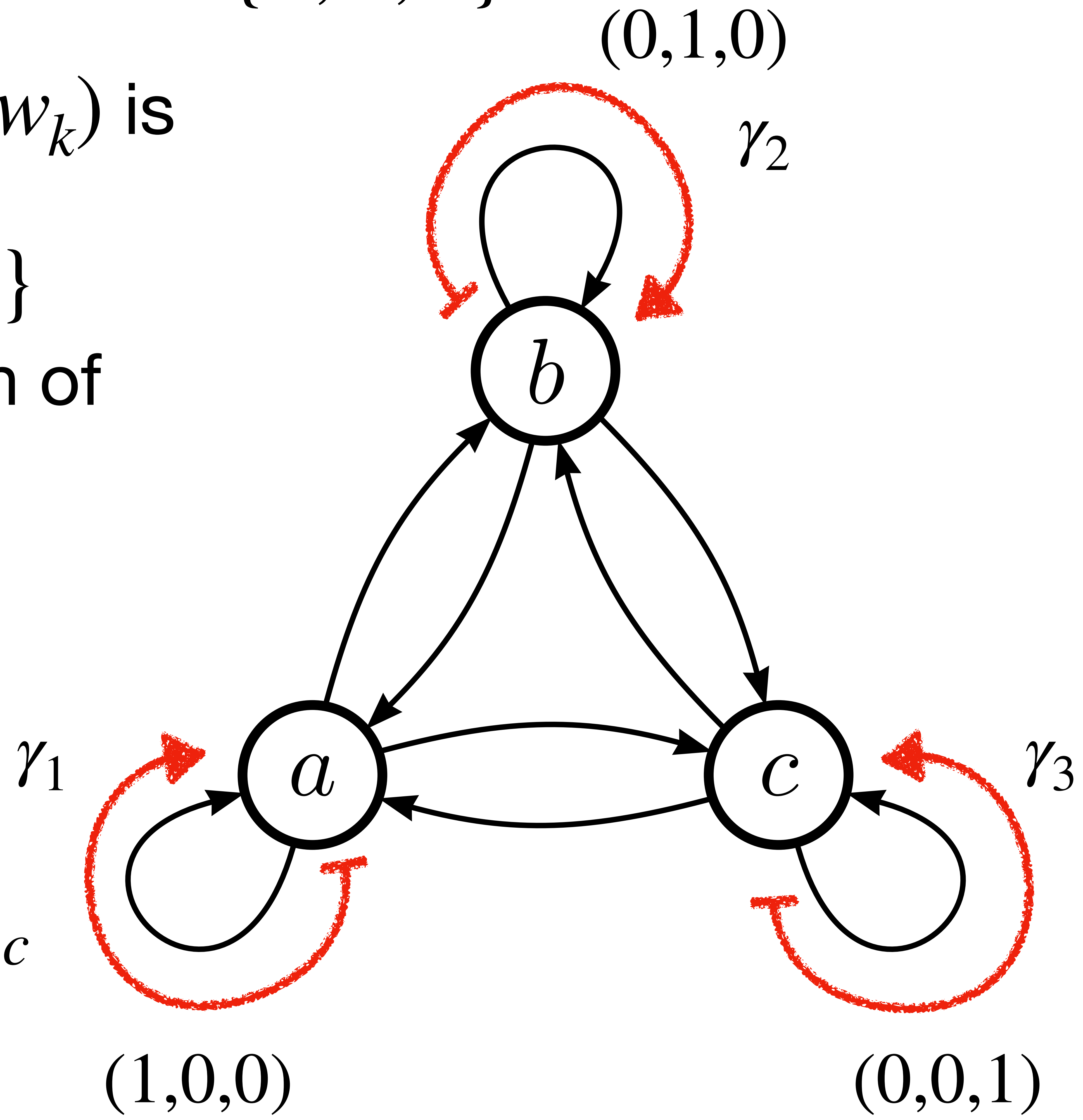
In this case $\dim(S) = 3$.



MIX: a simplest example of dimension three

$$\mathrm{MIX} = L_A(a, b, c) \text{ where } A = \{a, b, c\}.$$

In general, the dimension of $L = L_A(w_1, \ldots, w_k)$ is defined as
$\dim(L) = \max\{\dim(S) \mid S \text{ is pumpable in } L\}$
(Here we should consider the de Bruijn graph of dimension $\max\{|w_i| \mid 1 \le i \le k\}$).



In this case $\dim(\mathrm{MIX}) = 3$.
MIX is not context-free, an automaton should count $|w|_a - |w|_b$ and $|w|_b - |w|_c$ simultaneously (two registers are required).

MIX: a simplest example of dimension three

$$\mathrm{MIX} = L_A(a, b, c) \text{ where } A = \{a, b, c\}.$$

In general, the dimension of $L = L_A(w_1, \ldots, w_k)$ is defined as
$\dim(L) = \max\{\dim(S) \mid S \text{ is pumpable in } L\}$
(Here we should consider the de Bruijn graph of dimension $\max\{\,|w_i| \mid 1 \leq i \leq k\}$).

Note: It is decidable whether a given set of cycles S in L is pumpable or not. Because we can formalise "S is pumpable in L" by a Presburger arithmetic formula.

Conclusion and future work

- We showed that the regularity and context-freeness for WMIX languages are decidable.
  - Our approach is calculating the “dimension” of a WMIX language L, which represents the minimum # of registers (minus 1) for accepting L.
- I think we can generalise the notion “dimension” for more general languages of the form
$$L_A(\varphi; w_1, \ldots, w_k) = \left\{ w \in A^* \mid \varphi(|w|_{w_1}, \ldots, |w|_{w_k}) \text{ is true} \right\}$$
  where φ is a Presburger formula with n free variables $x_1, \ldots, x_k$,

  hence the context-freeness for such languages is still decidable.

Appendix