Reconstructing Chemical Reaction Networks by Solving Boolean Polynomial Systems

Chenqi Mou   Wei Niu

LMIB-School of Mathematics and Systems Science   École Centrale Pékin

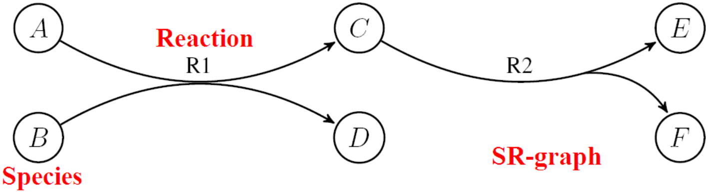
Beihang University, Beijing 100191, China

chenqi.mou, wei.niu@buaa.edu.cn

December 12, 2013 · Nanning, China

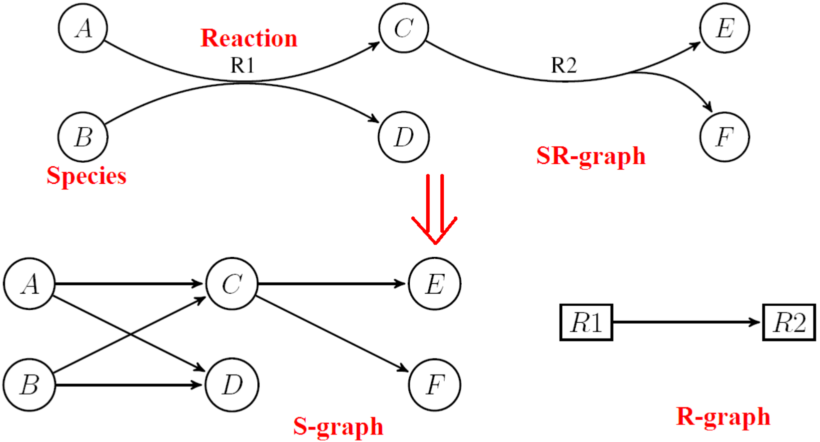
The problem

Chemical reaction networks



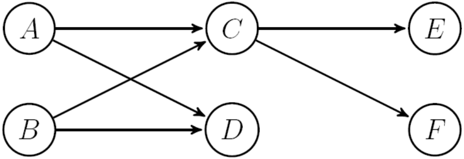
The problem

Chemical reaction networks

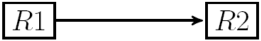


The problem

Chemical reaction networks



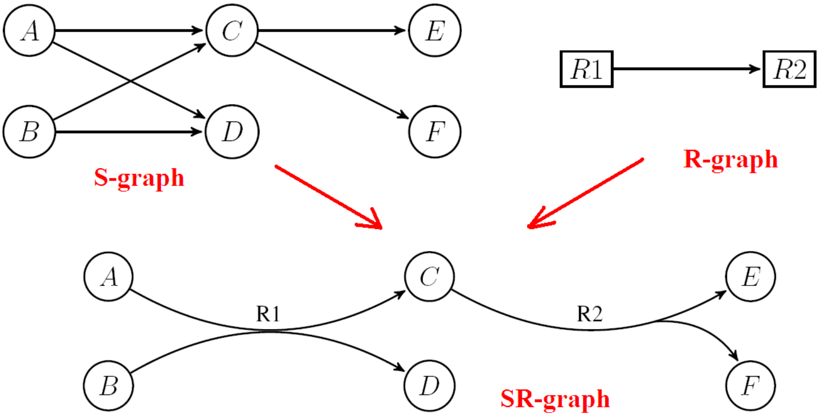
S-graph



R-graph

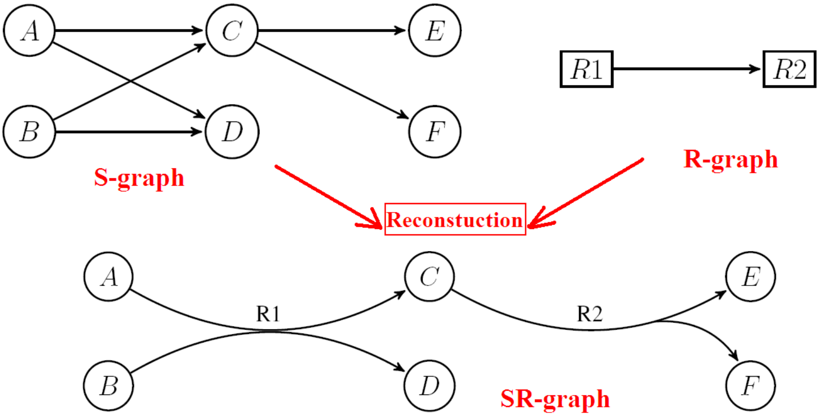
The problem

Chemical reaction networks



Reconstructing Chemical Reaction Networks

Chemical reaction networks



Why this problem?

- S- and R-graphs: easier for detecting
- Can the same S- and R-graphs lead to different SR-graphs?
- What do these SR-graphs mean?

Why this problem?

- S- and R-graphs: easier for detecting
- Can the same S- and R-graphs lead to different SR-graphs?
- What do these SR-graphs mean?

CRR (Compound-Reaction-Reconstruction) problem

[Fagerberg et. al. 2013]
Existence / NP-hard / SAT, SMT, ILP

Why this problem?

- S- and R-graphs: easier for detecting
- Can the same S- and R-graphs lead to different SR-graphs?
- What do these SR-graphs mean?

CRR (Compound-Reaction-Reconstruction) problem

[Fagerberg et. al. 2013]
Existence / NP-hard / SAT, SMT, ILP

$\Longrightarrow$ CRR$^+$ problem: all the potential SR-graphs

Why Polynomial System Solving (PoSSo)?

CRR problem

Existence	Hilbert's Nullstellensatz
NP-hardness	PoSSo is also NP-hard [Garey & Johnson 1979]
SAT, SMT, ILP	Polynomial system solvers

Why Polynomial System Solving (PoSSo)?

CRR problem

Existence	Hilbert's Nullstellensatz
NP-hardness	PoSSo is also NP-hard [Garey & Johnson 1979]
SAT, SMT, ILP	Polynomial system solvers
All the solutions	
feasible	natural
	Complexity:
	⇝ Worst: doubly exponential (in #var) [Mayr & Meyer 1982]
	⇝ Dedicated complexity (structured): bidegree (1,1) [Faugère, Safey El Din, Spaenlehauer 2010]

Matrix representation

R: a reaction $\Longrightarrow$ Input species: $I(R)$; Output species: $O(R)$;

SR-graph $\rightleftarrows$ two Boolean matrices

Matrix representation

R: a reaction $\Longrightarrow$ Input species: $I(R)$; Output species: $O(R)$;

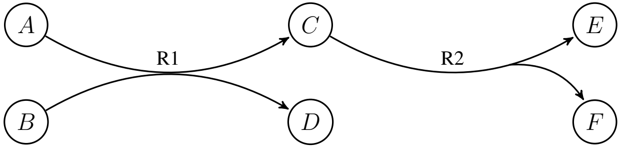
SR-graph $\rightleftarrows$ two Boolean matrices

$\mathbf{E}_{m\times n}$ such that

$$\mathbf{E}_{i,k} := \begin{cases} 1, & S_i \in I(R_k) \\ 0, & \text{Otherwise} \end{cases}$$

$\mathbf{P}_{n\times m}$ such that

$$\mathbf{P}_{k,j} := \begin{cases} 1, & S_j \in O(R_k) \\ 0, & \text{Otherwise} \end{cases}$$



Matrix representation

- S-graphs: Boolean matrix $\mathbf{S}_{m \times m}$ such that

$$\mathbf{S}_{i,j} := \begin{cases} 1, & \exists R_k \text{ s.t. } S_i \in I(R_k) \text{ and } S_j \in O(R_k) \\ 0, & \text{Otherwise} \end{cases}$$

- R-graphs: Boolean matrix $\mathbf{R}_{n \times n}$ such that

$$\mathbf{R}_{k,l} := \begin{cases} 1, & \exists S_i \text{ s.t. } S_i \in O(R_k) \text{ and } S_i \in I(R_k) \\ 0, & \text{Otherwise} \end{cases}$$

Matrix representation

- S-graphs: Boolean matrix $\mathbf{S}_{m\times m}$ such that

$$\mathbf{S}_{i,j} := \begin{cases} 1, & \exists R_k \text{ s.t. } S_i \in I(R_k) \text{ and } S_j \in O(R_k) \\ 0, & \text{Otherwise} \end{cases}$$

- R-graphs: Boolean matrix $\mathbf{R}_{n\times n}$ such that

$$\mathbf{R}_{k,l} := \begin{cases} 1, & \exists S_i \text{ s.t. } S_i \in O(R_k) \text{ and } S_i \in I(R_k) \\ 0, & \text{Otherwise} \end{cases}$$

Input: $\mathbf{S}$, $\mathbf{R} \Longrightarrow$ **Output**: $\mathbf{E}$, $\mathbf{P}$

- CRR: existence of $\mathbf{E}$ and $\mathbf{P}$
- CRR$^+$: all the possible $\mathbf{E}$ and $\mathbf{P}$

Relationship

$\mathbf{S}$, $\mathbf{R}$, $\mathbf{E}$, and $\mathbf{P}$

$$\mathbf{S}_{i,j} = \bigwedge_{k=1,\dots,n} (\mathbf{E}_{i,k} \vee \mathbf{P}_{k,j}), \qquad \mathbf{R}_{k,l} = \bigwedge_{i=1,\dots,m} (\mathbf{P}_{k,i} \vee \mathbf{E}_{i,l}).$$

- Direct translation to PoSSo problem

Background

Boolean polynomial ring $\mathbb{F}_2[\mathbf{E}_{1,1}, \dots, \mathbf{E}_{m,n}, \mathbf{P}_{1,1}, \dots, \mathbf{P}_{n,m}]$

$\Downarrow$

$x \wedge y = x \cdot y$ and $x \vee y = x + y + x \cdot y$

$\Downarrow$

Boolean polynomial system

Structure

$$\mathbf{S}_{i,j} = \bigwedge_{k=1,\ldots,n}(\mathbf{E}_{i,k} \vee \mathbf{P}_{k,j})$$

$$x \wedge y = x \cdot y \text{ and } x \vee y = x + y + x \cdot y$$

- $\mathbf{S}_{i,j} = 1 \implies 1$ polynomial equation (degree $2n$; variable $2n$)
  $\implies$ of type s (or r if $\mathbf{R}_{i,j} = 1$)
- $\mathbf{S}_{i,j} = 0 \implies n$ bivariate quadratic equations
  $\implies$ of type 0

Structure

$$\mathbf{S}_{i,j} = \bigwedge_{k=1,\ldots,n} (\mathbf{E}_{i,k} \vee \mathbf{P}_{k,j})$$

$$x \wedge y = x \cdot y \text{ and } x \vee y = x + y + x \cdot y$$

- $\mathbf{S}_{i,j} = 1 \implies 1$ polynomial equation (degree $2n$; variable $2n$) $\implies$ of type s (or r if $\mathbf{R}_{i,j} = 1$)
- $\mathbf{S}_{i,j} = 0 \implies n$ bivariate quadratic equations $\implies$ of type 0

Structure (p and q: #zeros in S and R)

- type 0: $np + mq$
- type s: $m^2 - p$
- type r: $n^2 - q$

#Solutions $\geq$ #Variables $\implies$ overdefined

PoSSo

Methods

- Gröbner bases [Buchberger 1965, Faugère 1999, 2002]
  triangular sets [Wang 2001, Moreno Maza 2000, Gao & Huang 2012]
  XL (overdefined) e.g., [Ars et. al. 2004]
  Polynomial system $\Longrightarrow$ in a better form $\Longrightarrow$ solutions
- Complexity (Gröbner bases): $O(\binom{n+d_{reg}}{n}^{\omega})$[Bardet, Faugère, Salvy 2004]
- Over $\mathbb{F}_2$: add the field equations $(x_k^2 + x_k = 0)$.

PoSSo

Implementation

Gröbner bases:

- Buchberger algorithm: almost in all Computer Algebra Systems
- F_4, F_5: FGb, MAGMA...
  $\Longrightarrow$ MAGMA: optimization for over $\mathbb{F}_2$ (since V2.15)

Triangular sets:

- Epsilon, RegularChains (in Maple) ...

Randomly generated $\mathbf{S}$ and $\mathbf{R}$

MAGMA V2.17-1 (F_4 implementation)
$\Longrightarrow$ V2.20 (released yesterday, F_4 updated)

m, n	P	Density (%)	#Var	$\#F$	Time	#Solutions
8	0.9	3.13/15.63	128	940	0.27	0
8	0.9	9.38/9.38	128	940	36.77	0
8	0.9	3.12/9.38	128	968	>1000	unknown
9	0.9	11.11/6.17	162	1346	8.25	0
9	0.9	12.35/6.17	162	1338	0.62	0
9	0.9	9.88/8.64	162	1338	>1000	unknown
10	0.9	10/8	200	1838	1.21	0
10	0.9	9/12	200	1811	1.17	0
11	0.9	14.05/10.74	242	2362	2.17	0
5	0.95	8/8	50	234	0.06	296
5	0.95	4/8	50	238	0.70	7759

Remarks on the experiments

- General one: no optimization is made
- for CRR:
  (1) Experimentally, not comparable to SMT / SAT in efficiency (with optimization)
  (2) Problem generation (VS CNF generation)
- There exist instances with more than 1 solution (not trivial)
- For real-world examples (Biology): size $(m, n \geq 40)$, sparsity $\geq 98\%$

Future work

- Structure $\Longrightarrow$ simplify the problem / dedicated algorithm
- Complexity analyses: better?
- CRR: NP-hardness by PoSSo?