O Problema da Subseqüência Comum Máxima sem Repetições

Christian Tjandraatmadja
Supervisora: Cristina Gomes Fernandes
Orientador: Carlos Eduardo Ferreira

Instituto de Matemática e Estatística
Universidade de São Paulo
Apoio: FAPESP (proc. 07/54282-6)

O que é subseqüência comum máxima?

Vagamente falando, mede similaridade entre seqüências.

O que é subseqüência comum máxima?

Vagamente falando, mede similaridade entre seqüências.

Aplicações

- Diferenciação de arquivos

O que é subseqüência comum máxima?

Vagamente falando, mede similaridade entre seqüências.

Aplicações

- Diferenciação de arquivos
- Correção de ortografia

O que é subseqüência comum máxima?

Vagamente falando, mede similaridade entre seqüências.

Aplicações

- Diferenciação de arquivos
- Correção de ortografia
- Comparação entre seqüências de DNA e de proteínas (Surge desse contexto a idéia da subseqüência comum máxima sem repetições.)

O que é subseqüência comum máxima?

Vagamente falando, mede similaridade entre seqüências.

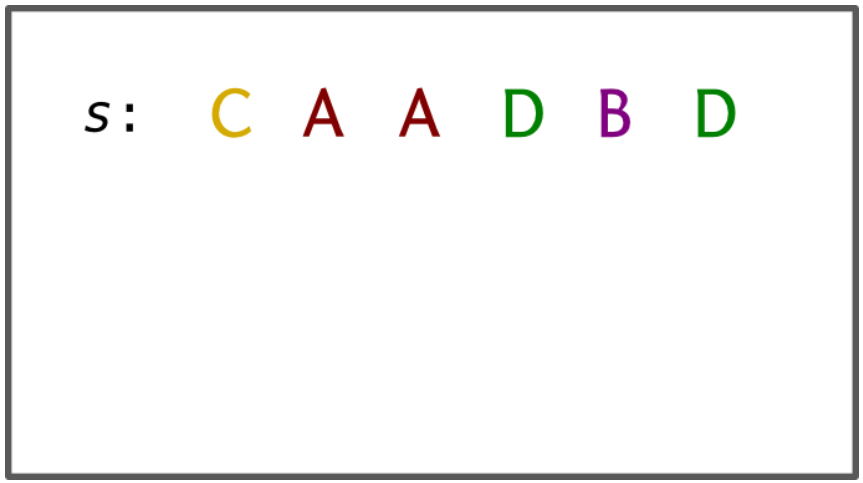
Aplicações

- Diferenciação de arquivos
- Correção de ortografia
- Comparação entre seqüências de DNA e de proteínas (Surge desse contexto a idéia da subseqüência comum máxima sem repetições.)

Objetivo

Estudar a estrutura desses problemas e implementar um algoritmo para resolver o problema da subseqüência comum máxima sem repetições.

Definições

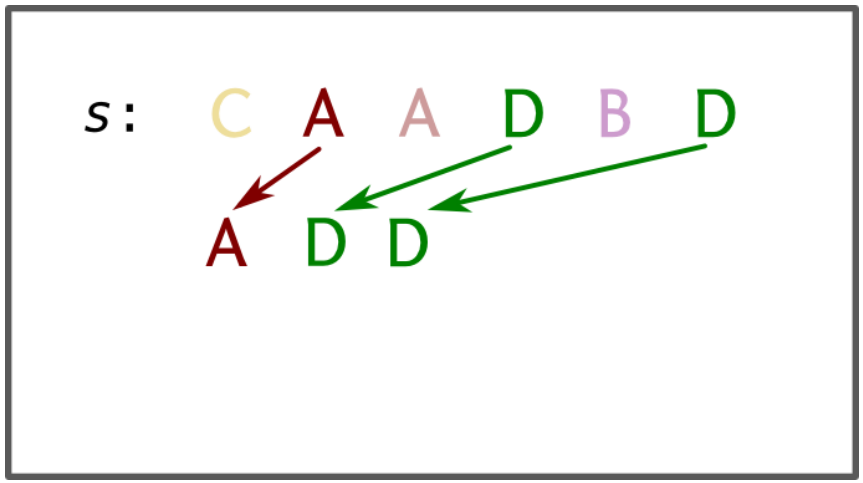


seqüência

Definições

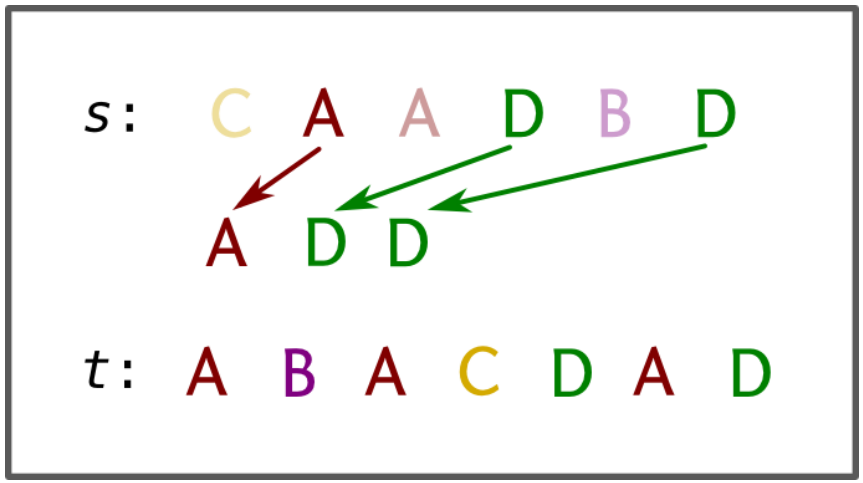
subseqüência

Definições



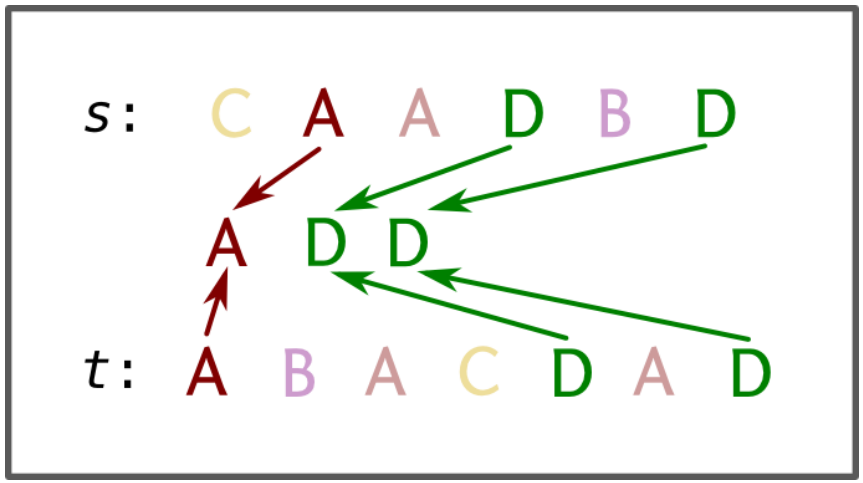
subseqüência

Definições



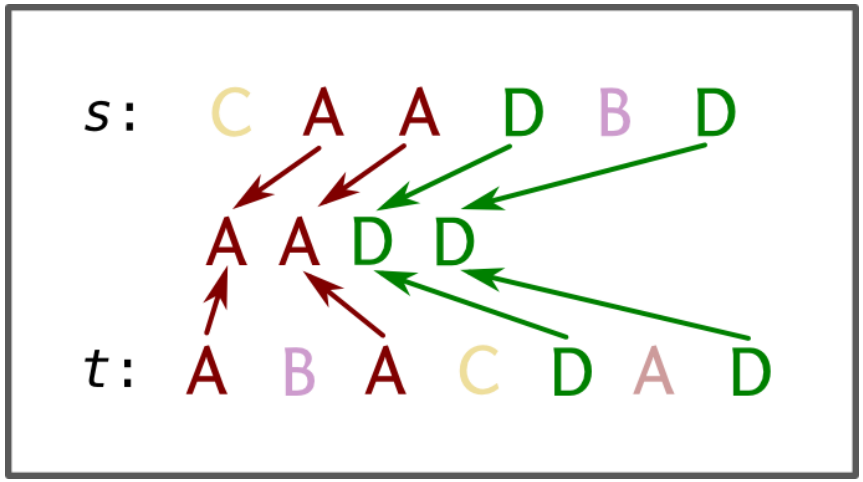
subseqüência

Definições



subseqüência comum

Definições

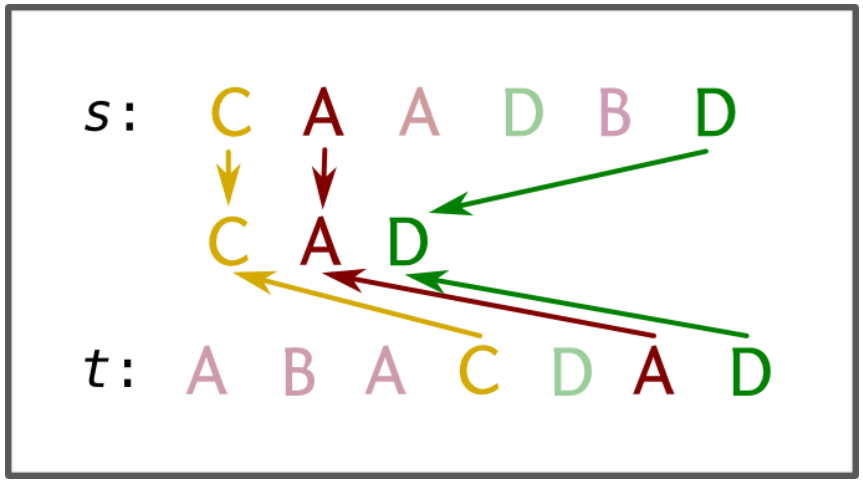


subseqüência comum máxima

Abreviações:

- LCS: subseqüência comum máxima (*longest common subsequence*)

Definições

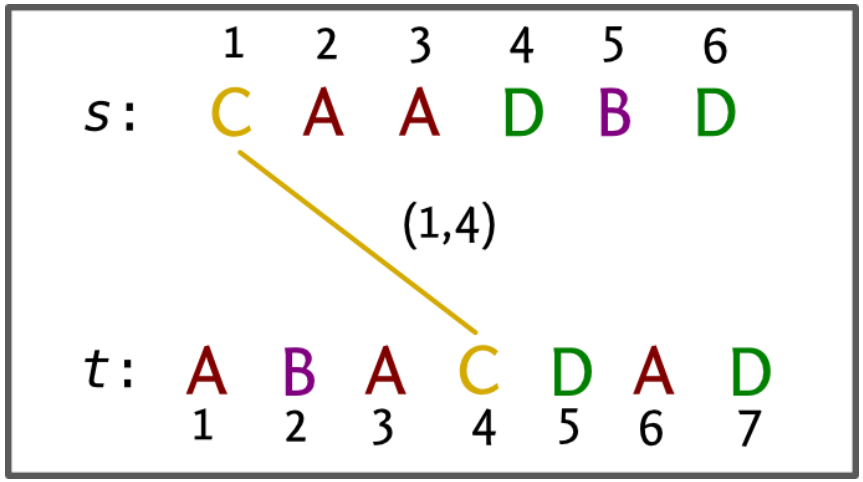


subseqüência comum máxima sem repetições

Abreviações:

- LCS: subseqüência comum máxima (*longest common subsequence*)
- RFLCS: subseqüência comum máxima sem repetições (*repetition free longest common subsequence*)

Definições



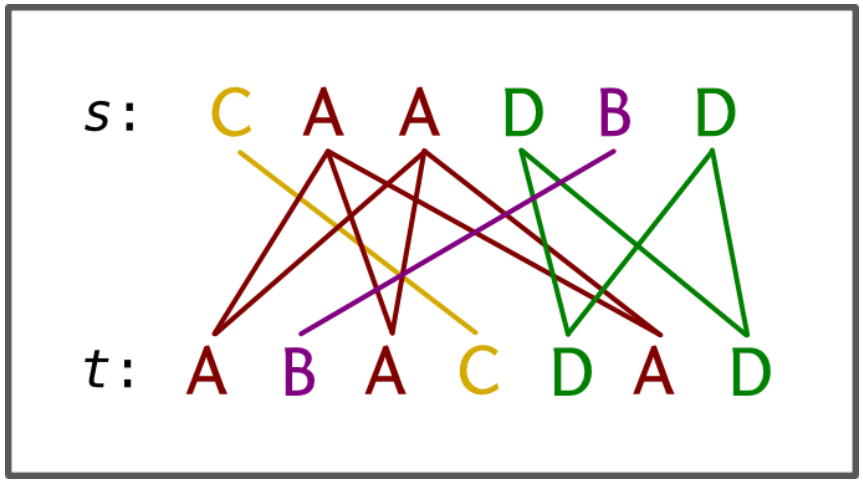
subseqüência comum máxima sem repetições

casamento: par de índices (i, j), $s_i = t_j$

Abreviações:

- LCS: subseqüência comum máxima (*longest common subsequence*)
- RFLCS: subseqüência comum máxima sem repetições (*repetition free longest common subsequence*)

Definições



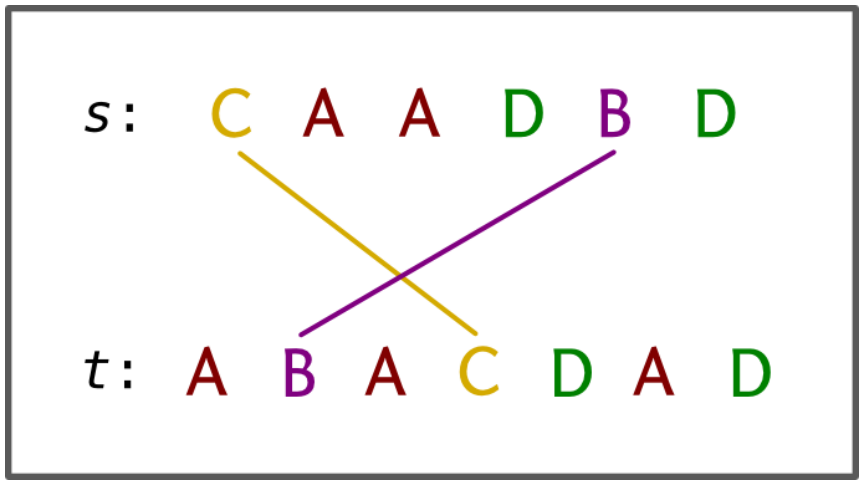
subseqüência comum máxima sem repetições

casamento: par de índices (i, j), $s_i = t_j$

Abreviações:

- LCS: subseqüência comum máxima (*longest common subsequence*)
- RFLCS: subseqüência comum máxima sem repetições (*repetition free longest common subsequence*)

Definições



subseqüência comum máxima sem repetições

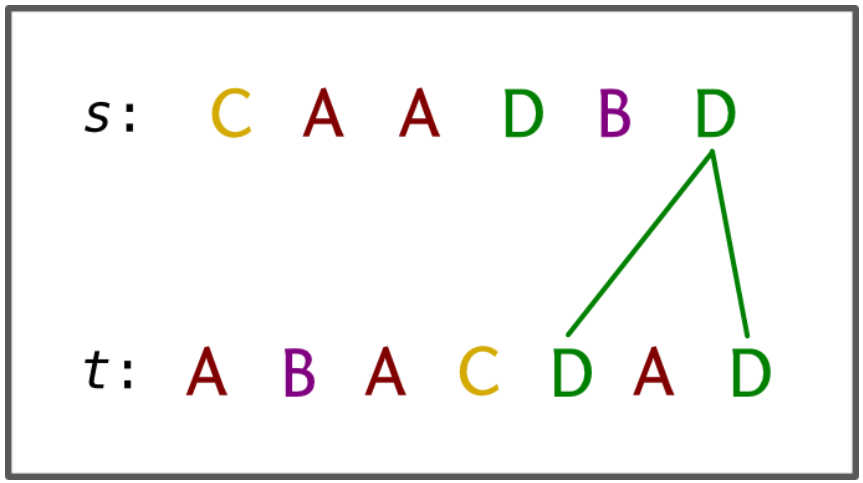
casamento: par de índices (i, j), $s_i = t_j$

cruzamento: $(i \leq k$ e $j \geq \ell)$ ou $(k \leq i$ e $\ell \geq j)$

Abreviações:

- LCS: subseqüência comum máxima (*longest common subsequence*)
- RFLCS: subseqüência comum máxima sem repetições (*repetition free longest common subsequence*)

Definições



subseqüência comum máxima sem repetições

casamento: par de índices (i, j), $s_i = t_j$

cruzamento: $(i \leq k$ e $j \geq \ell)$ ou $(k \leq i$ e $\ell \geq j)$

Abreviações:

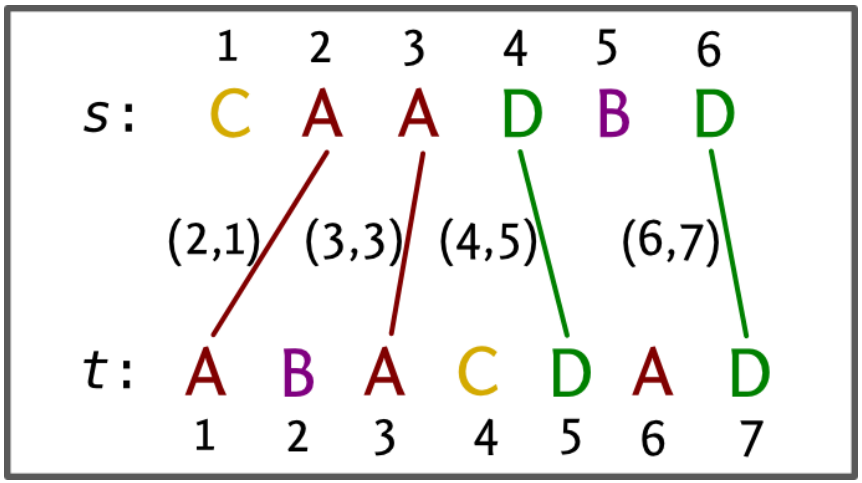
- LCS: subseqüência comum máxima (*longest common subsequence*)
- RFLCS: subseqüência comum máxima sem repetições (*repetition free longest common subsequence*)

Casamentos?

- Um conjunto de casamentos que não se cruzam dois a dois representa uma subseqüência comum .

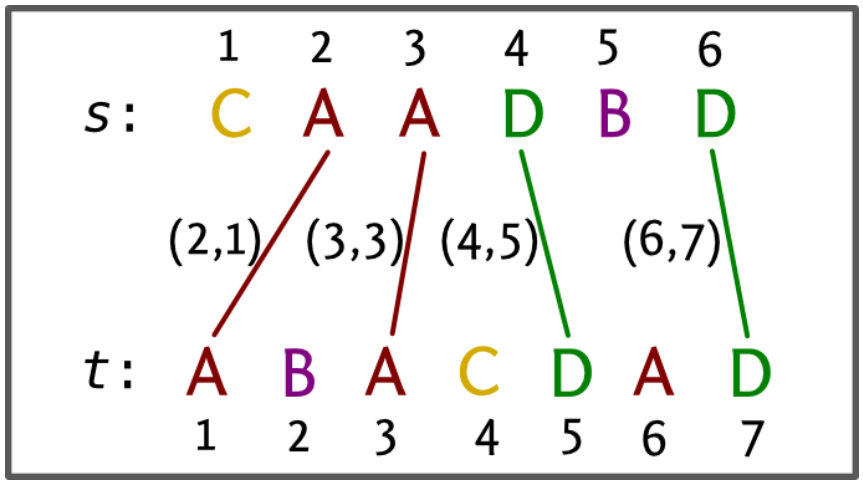
Casamentos?

- Um conjunto de casamentos que não se cruzam dois a dois representa uma subseqüência comum .
- Não há cruzamento $\longrightarrow$ ordenação dos casamentos de forma estritamente crescente $\longrightarrow$ seqüência de símbolos.



Casamentos?

- Um conjunto de casamentos de cardinalidade máxima que não se cruzam dois a dois representa uma subseqüência comum máxima.
- Não há cruzamento $\longrightarrow$ ordenação dos casamentos de forma estritamente crescente $\longrightarrow$ seqüência de símbolos.



Formulação como problema de programação inteira

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

Formulação como problema de programação inteira

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\max \sum_{(i,j) \in E} z_{ij}$$

Formulação como problema de programação inteira

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} z_{ij} \\ \text{s. a} \quad & z_{ij} \in \{0,1\} \quad \text{para todo } (i,j) \text{ em } C. \end{aligned}$$

Formulação como problema de programação inteira

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} z_{ij} & & \\ \text{s. a} \quad & z_{ij} + z_{k\ell} \leq 1 & & \text{para todo } (i,j) \text{ e } (k,\ell) \\ & & & \text{em } C \text{ que se cruzam,} \\ & z_{ij} \in \{0,1\} & & \text{para todo } (i,j) \text{ em } C. \end{aligned}$$

Formulação como problema de programação inteira

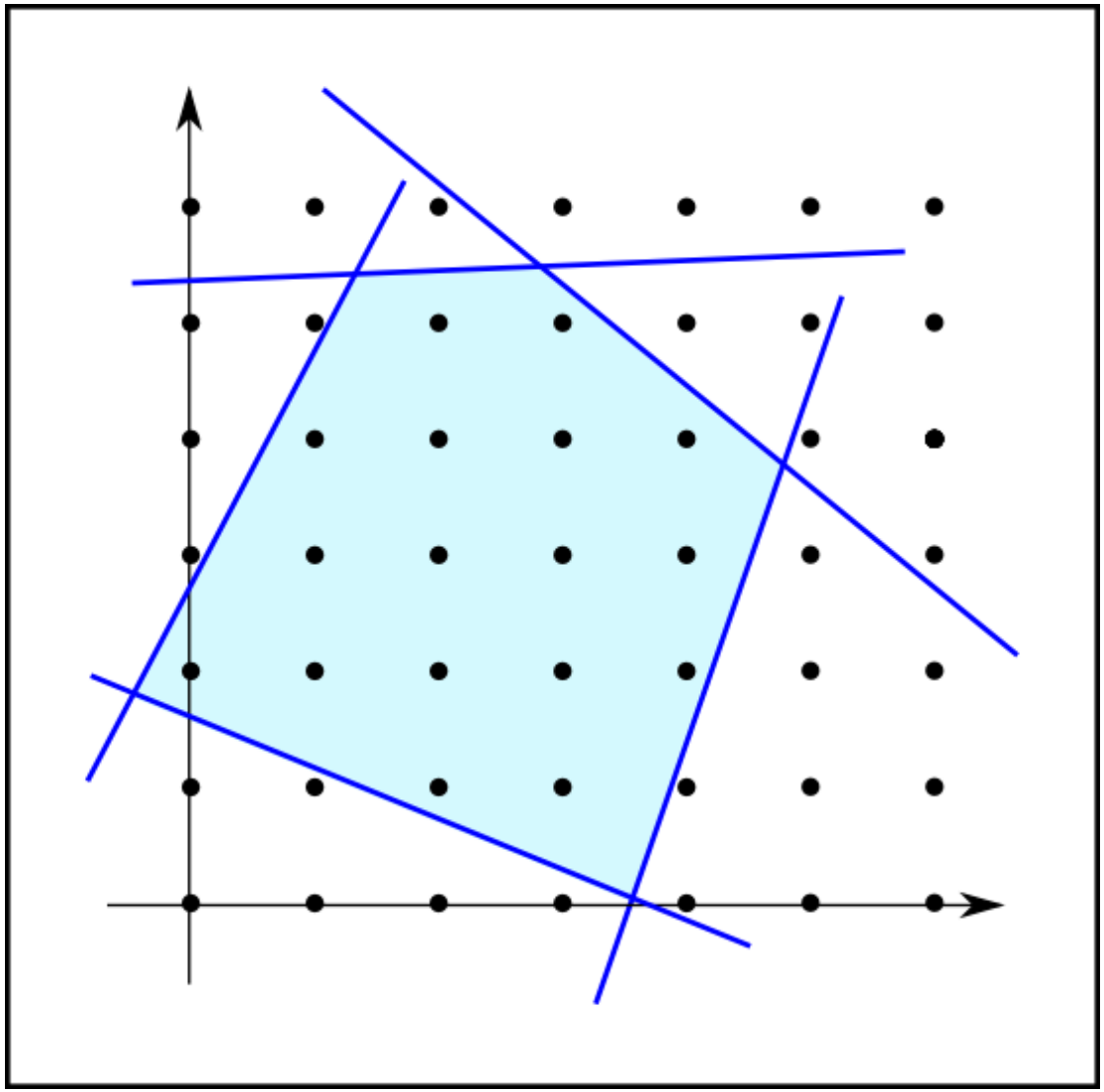
Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} z_{ij} & & \\ \text{s. a} \quad & z_{ij} + z_{k\ell} \leq 1 & & \text{para todo } (i,j) \text{ e } (k,\ell) \\ & & & \text{em } C \text{ que se cruzam,} \\ & z_{ij} \in \{0,1\} & & \text{para todo } (i,j) \text{ em } C. \end{aligned}$$

Será que há formulação melhor?

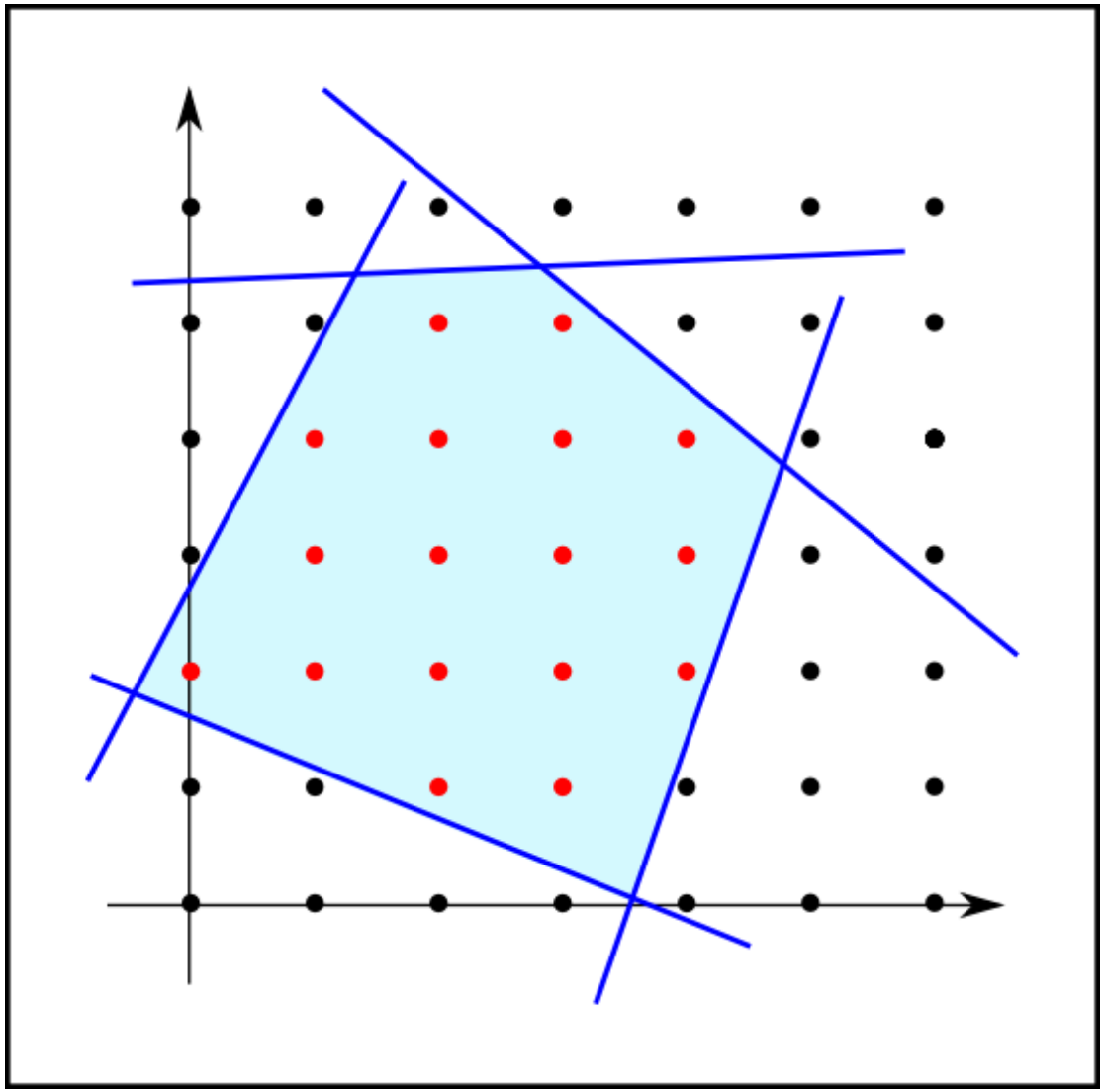
O que é uma formulação melhor?

Um exemplo:



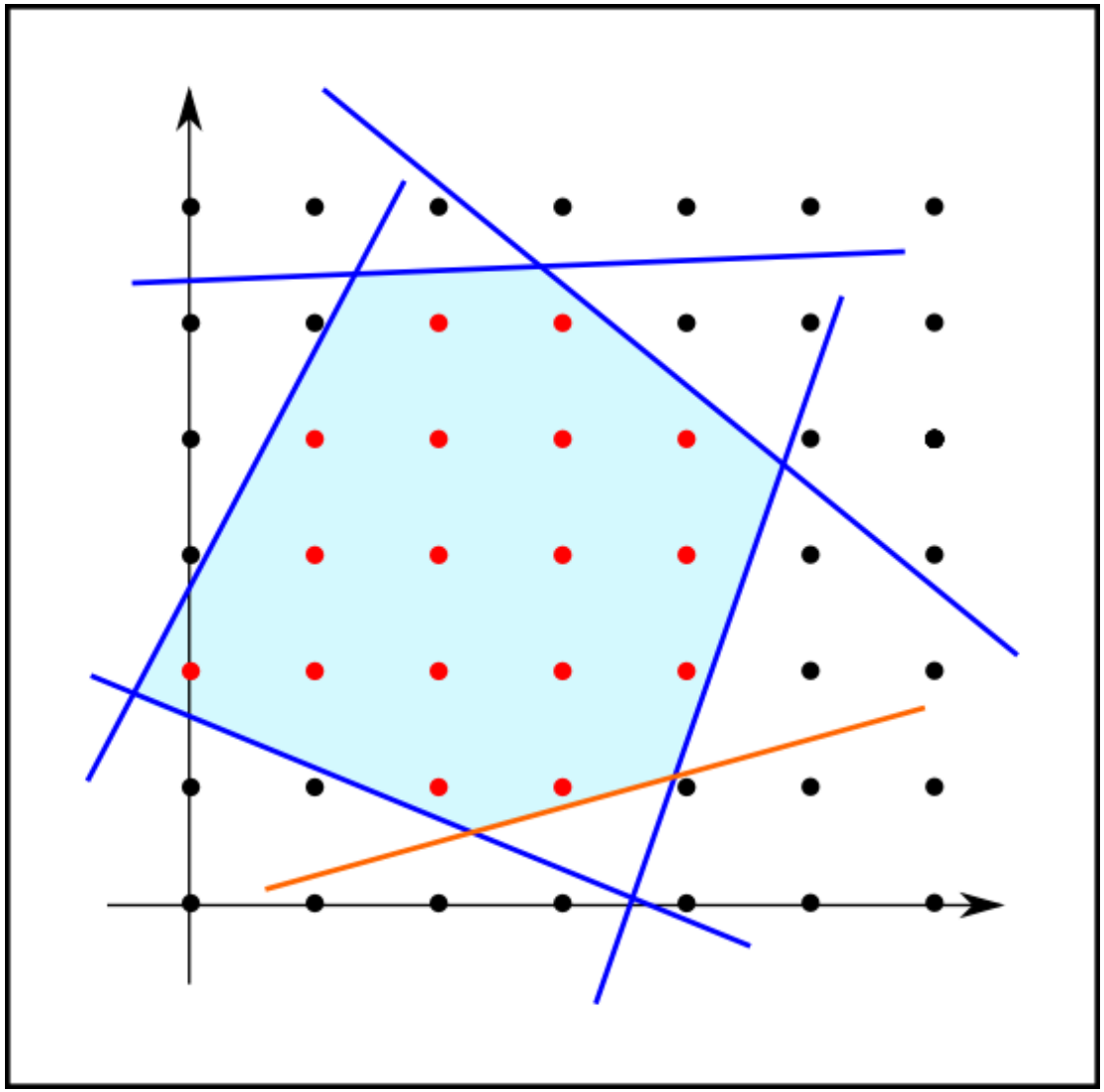
O que é uma formulação melhor?

Um exemplo:



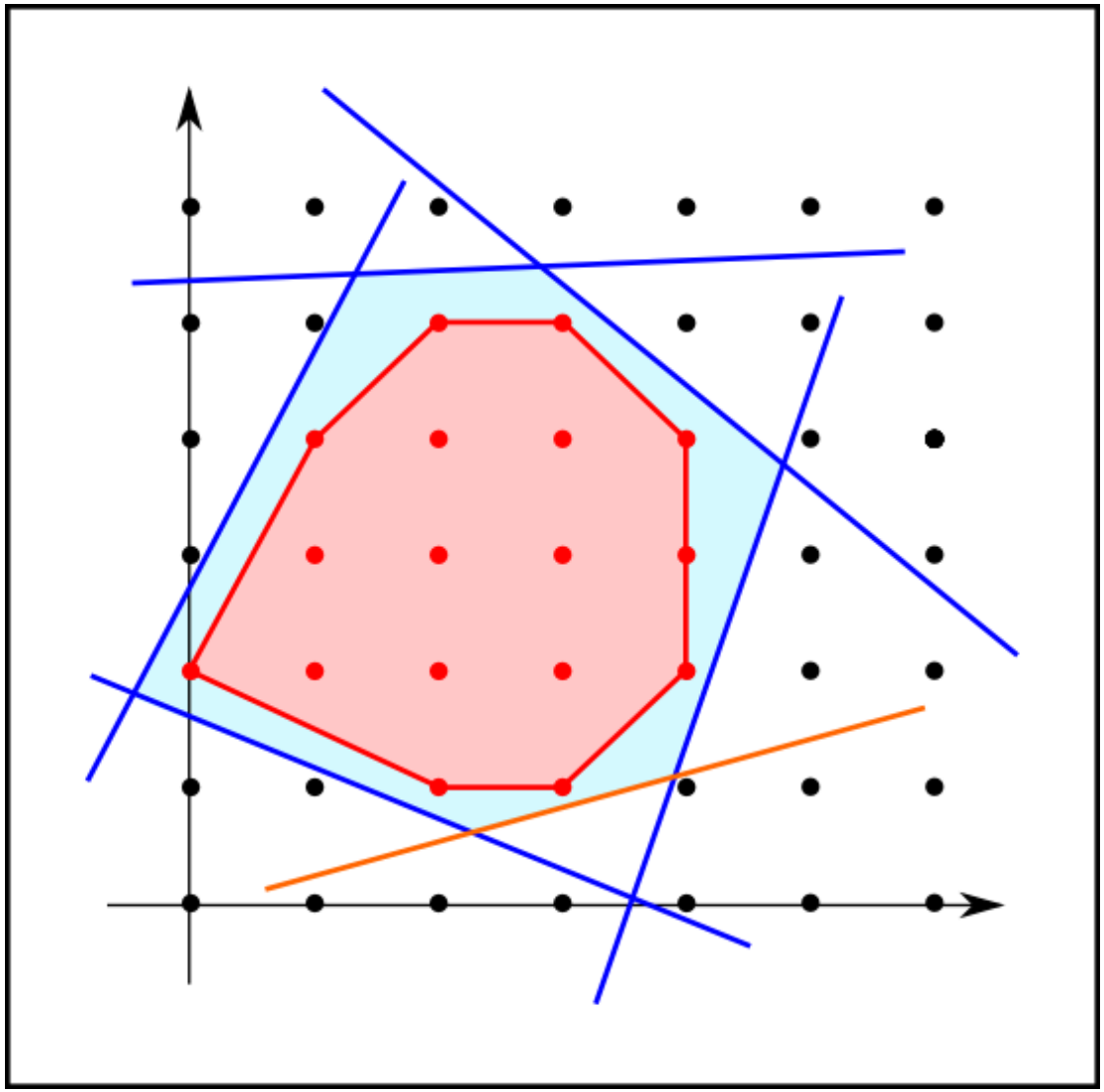
O que é uma formulação melhor?

Um exemplo:



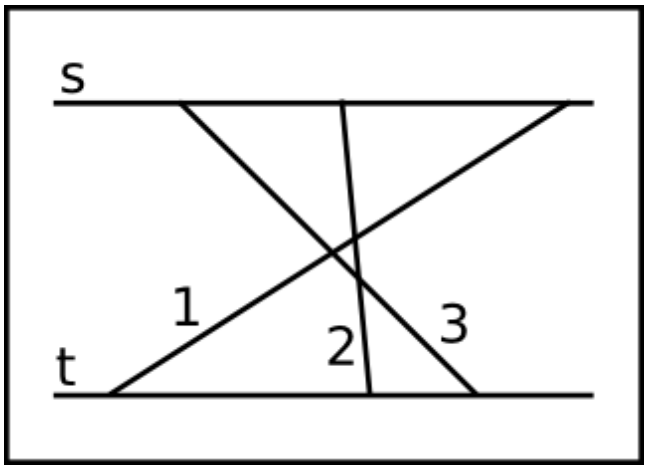
O que é uma formulação melhor?

Um exemplo:

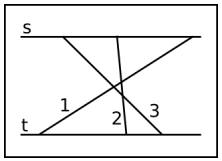


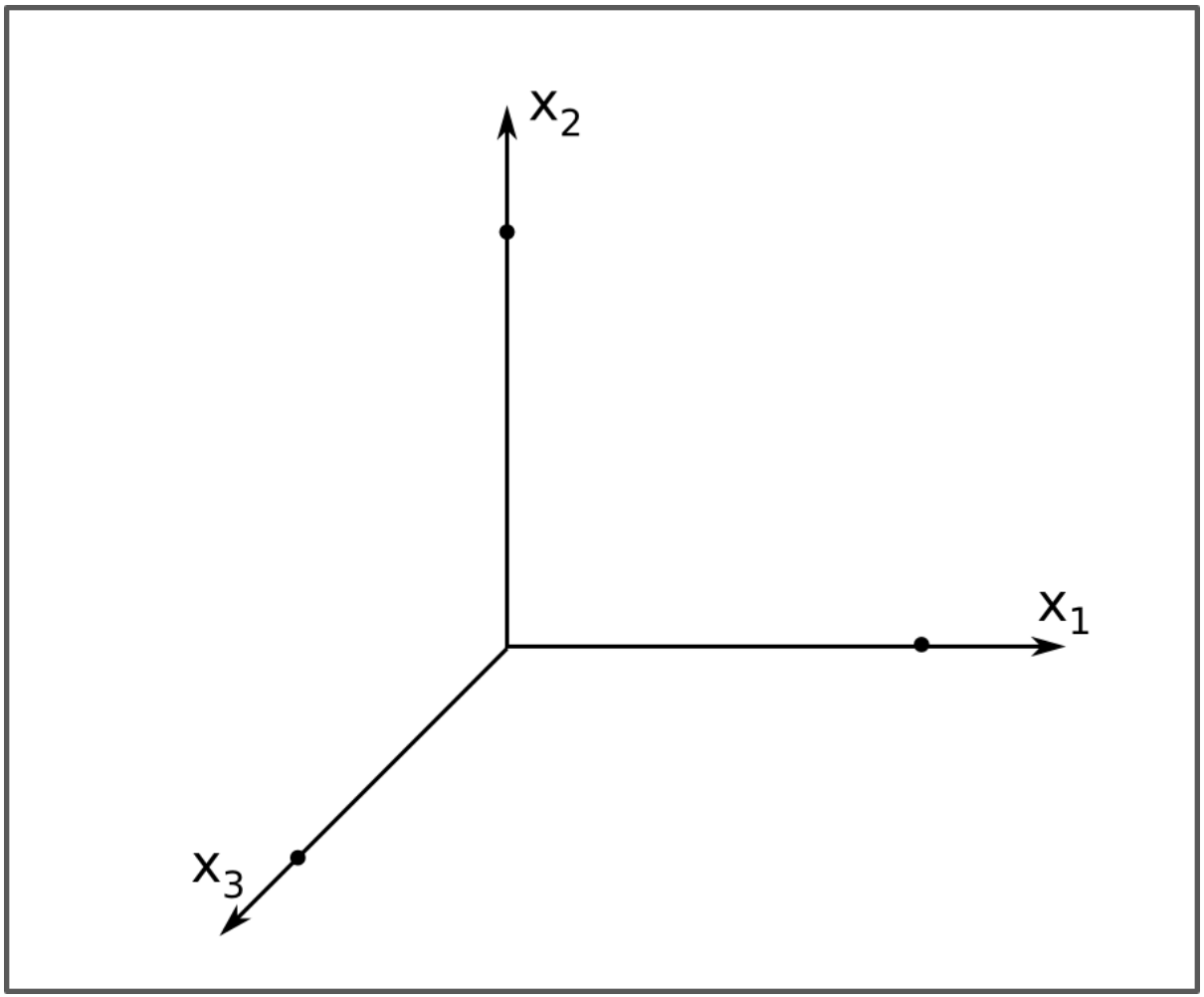
E no nosso contexto?

Suponha que temos três casamentos 1, 2 e 3, com variáveis x_1, x_2 e x_3 respectivamente. Suponha também que eles se cruzam dois a dois, como a seguir.

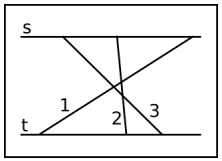


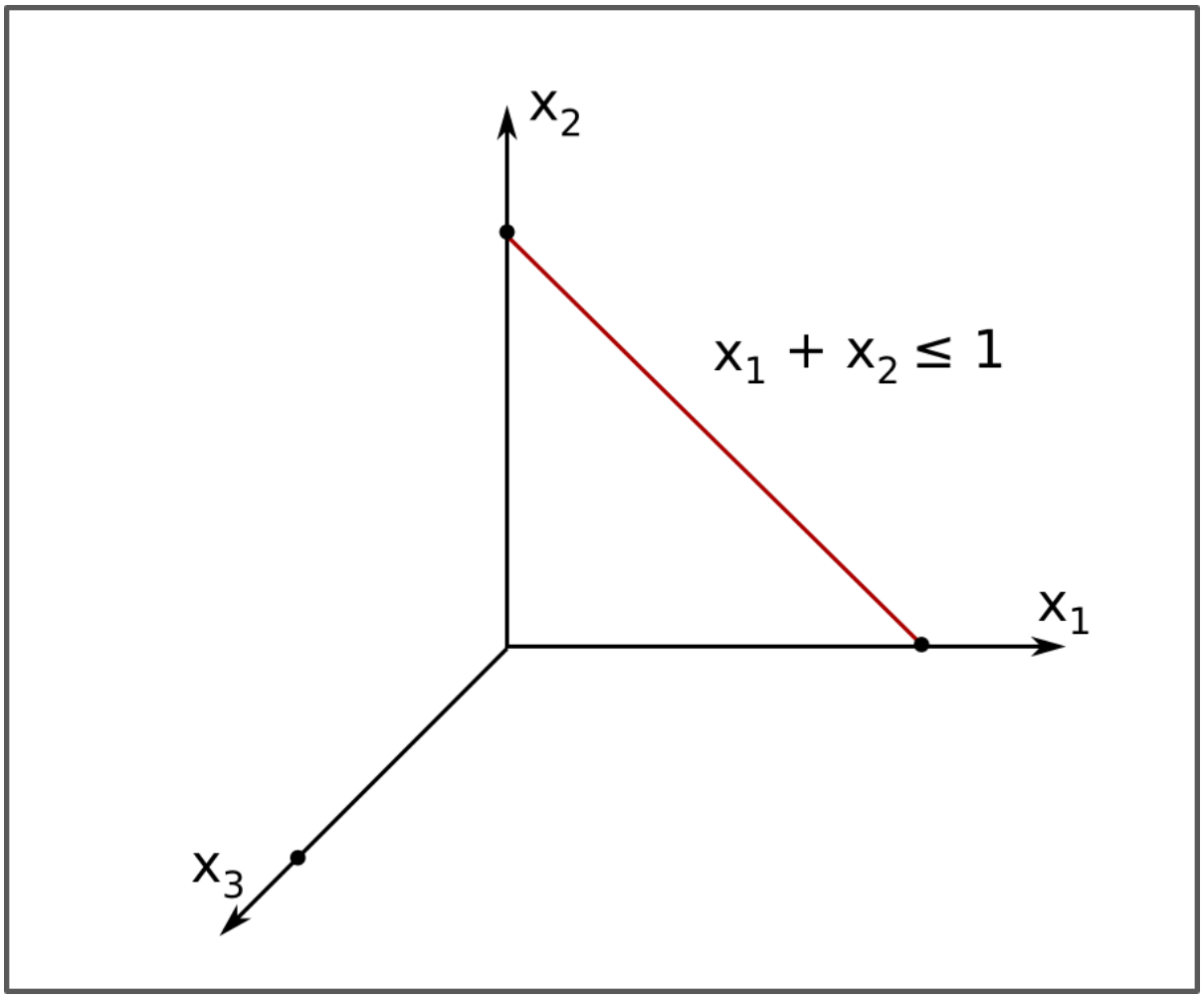
E no nosso contexto?



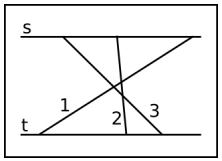


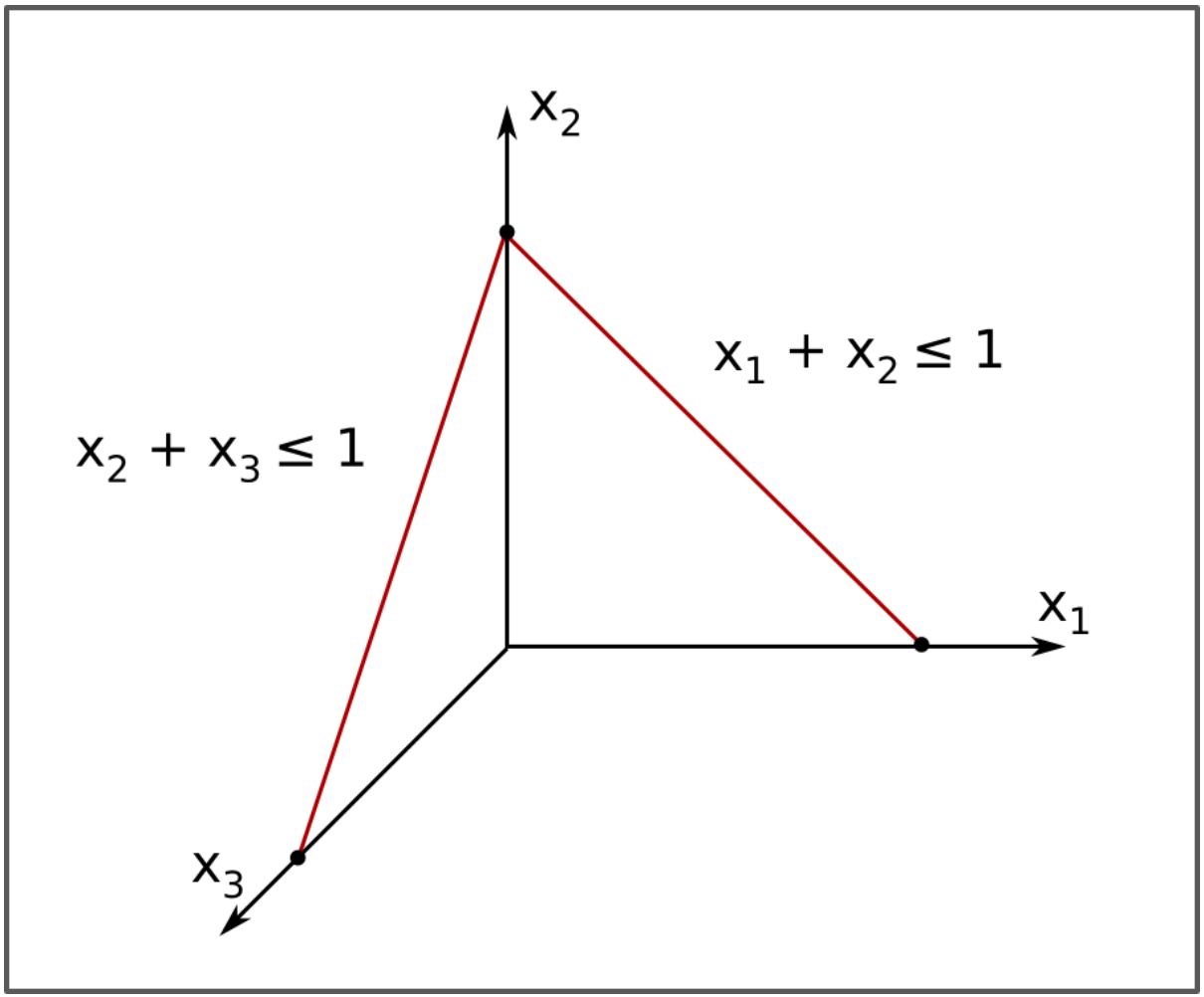
E no nosso contexto?



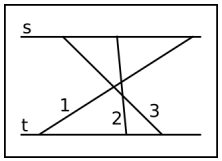


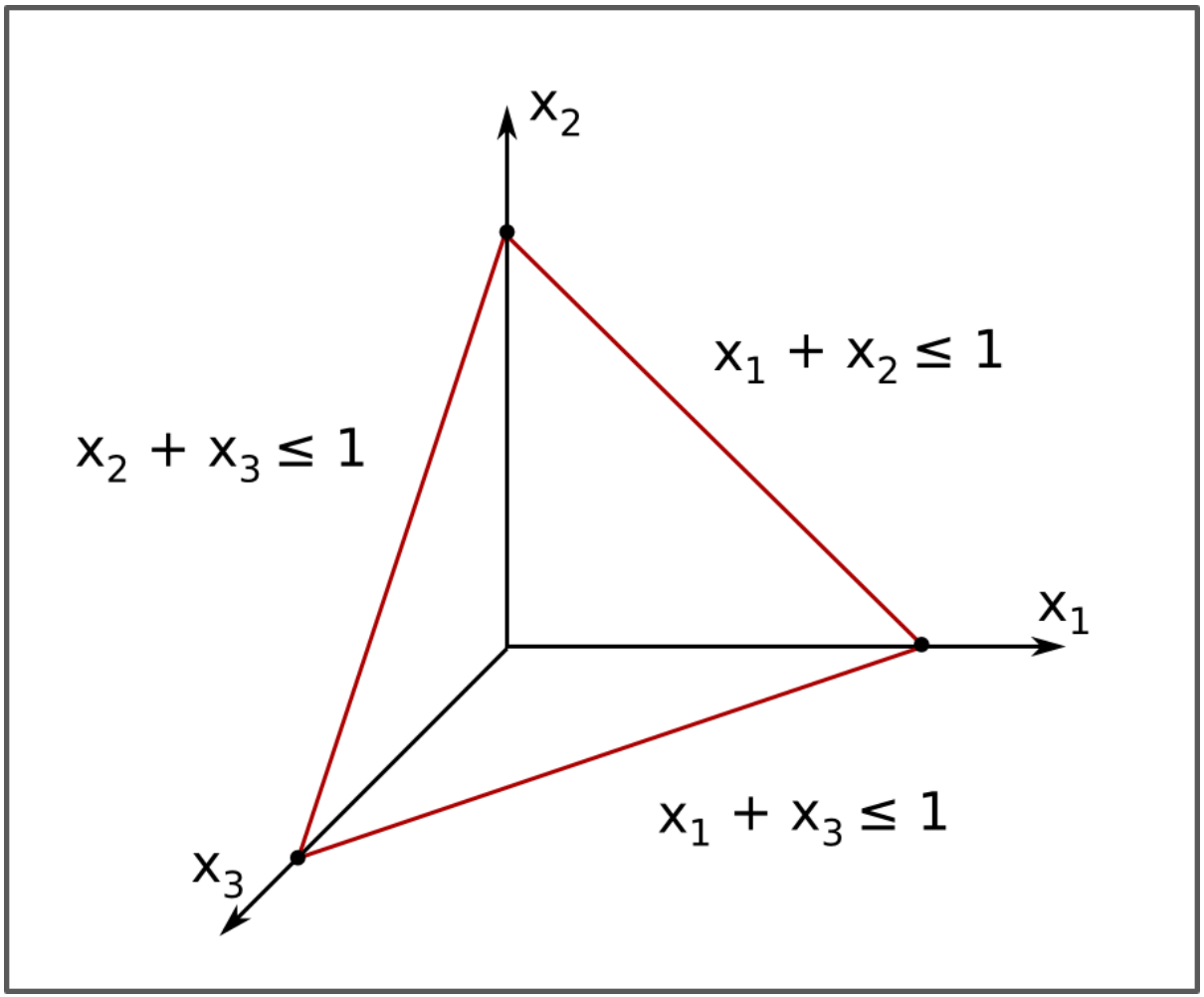
E no nosso contexto?



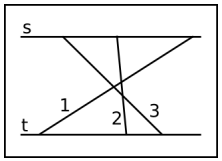


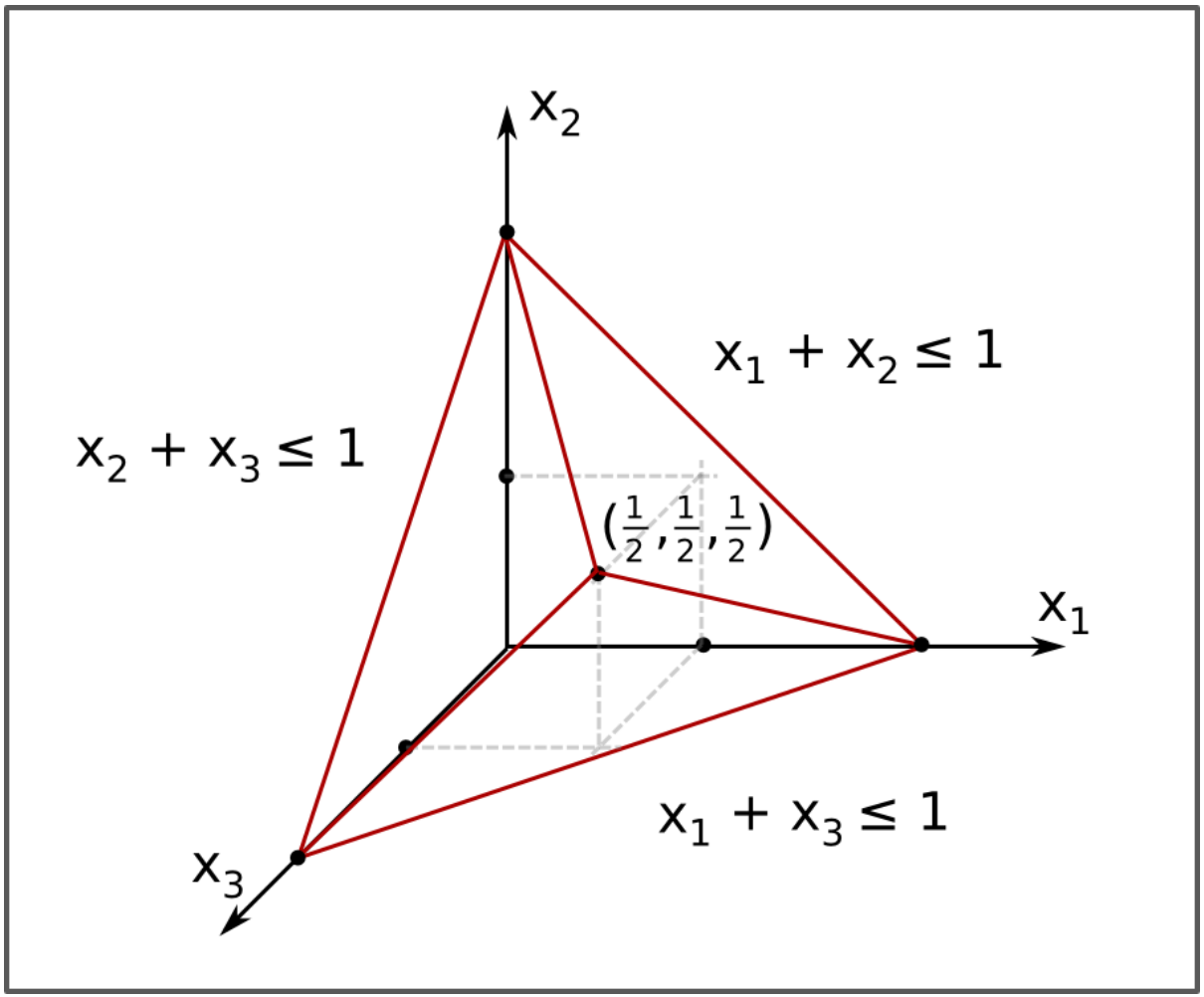
E no nosso contexto?



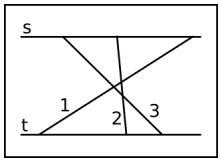


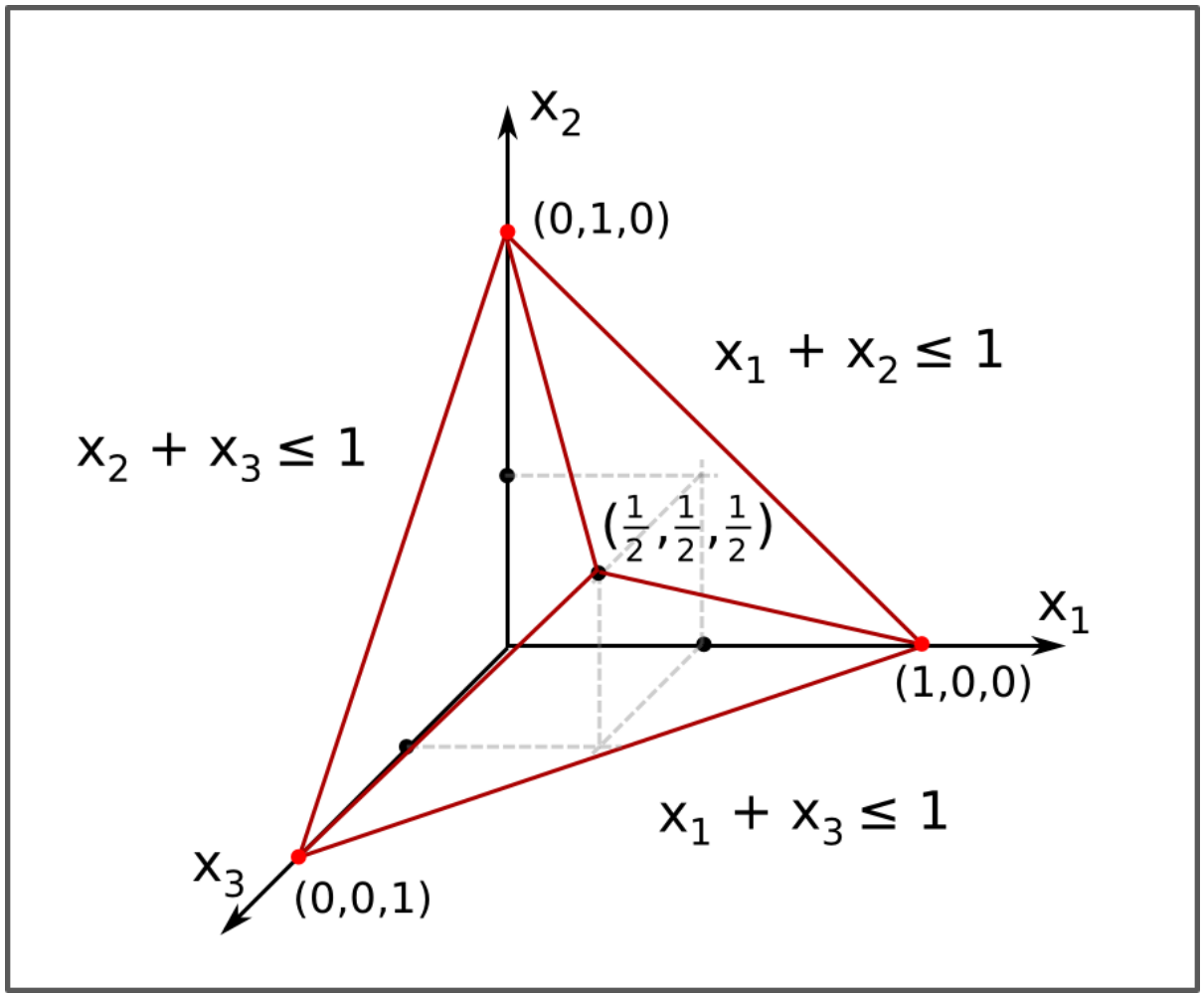
E no nosso contexto?



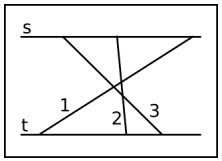


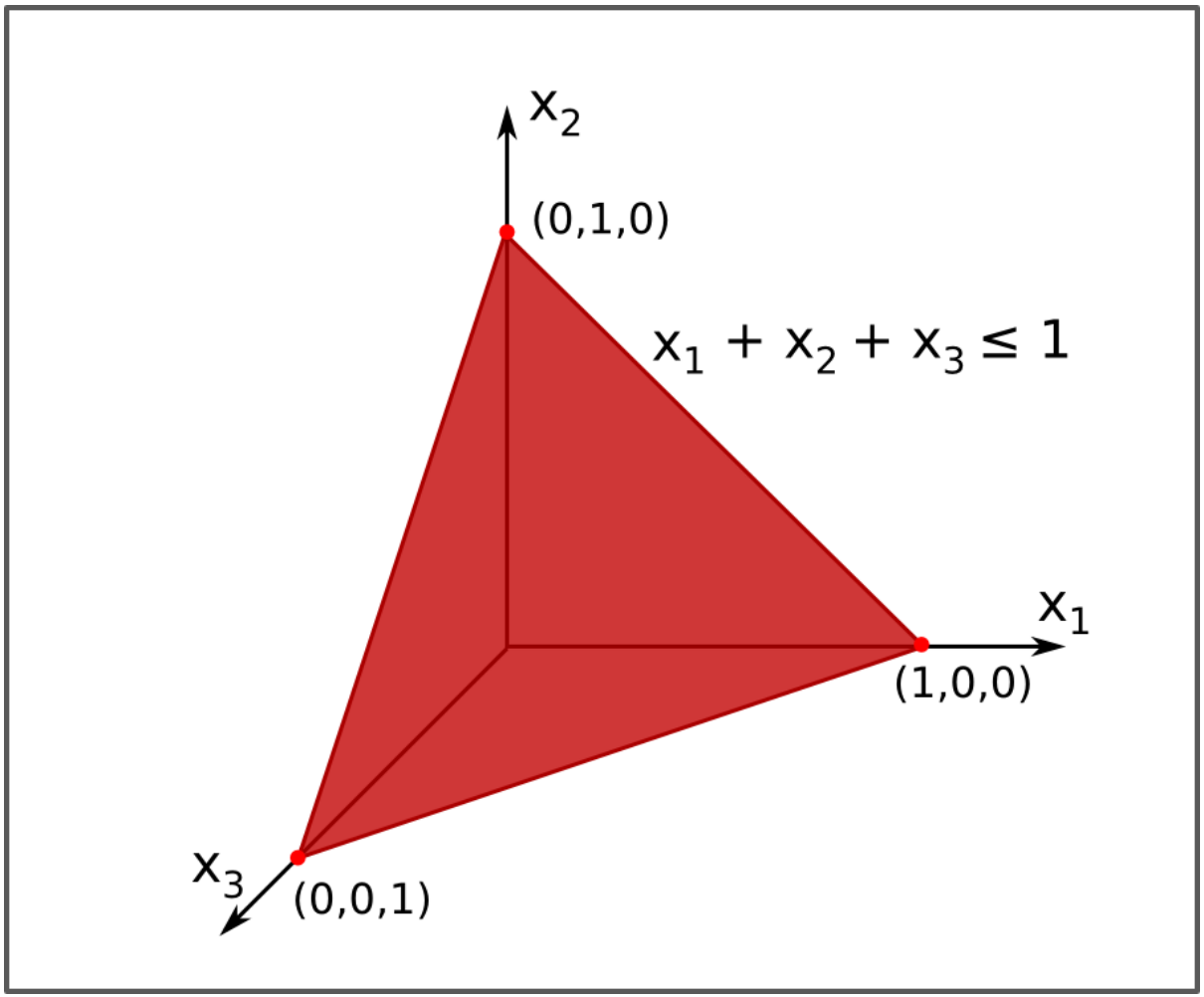
E no nosso contexto?





E no nosso contexto?

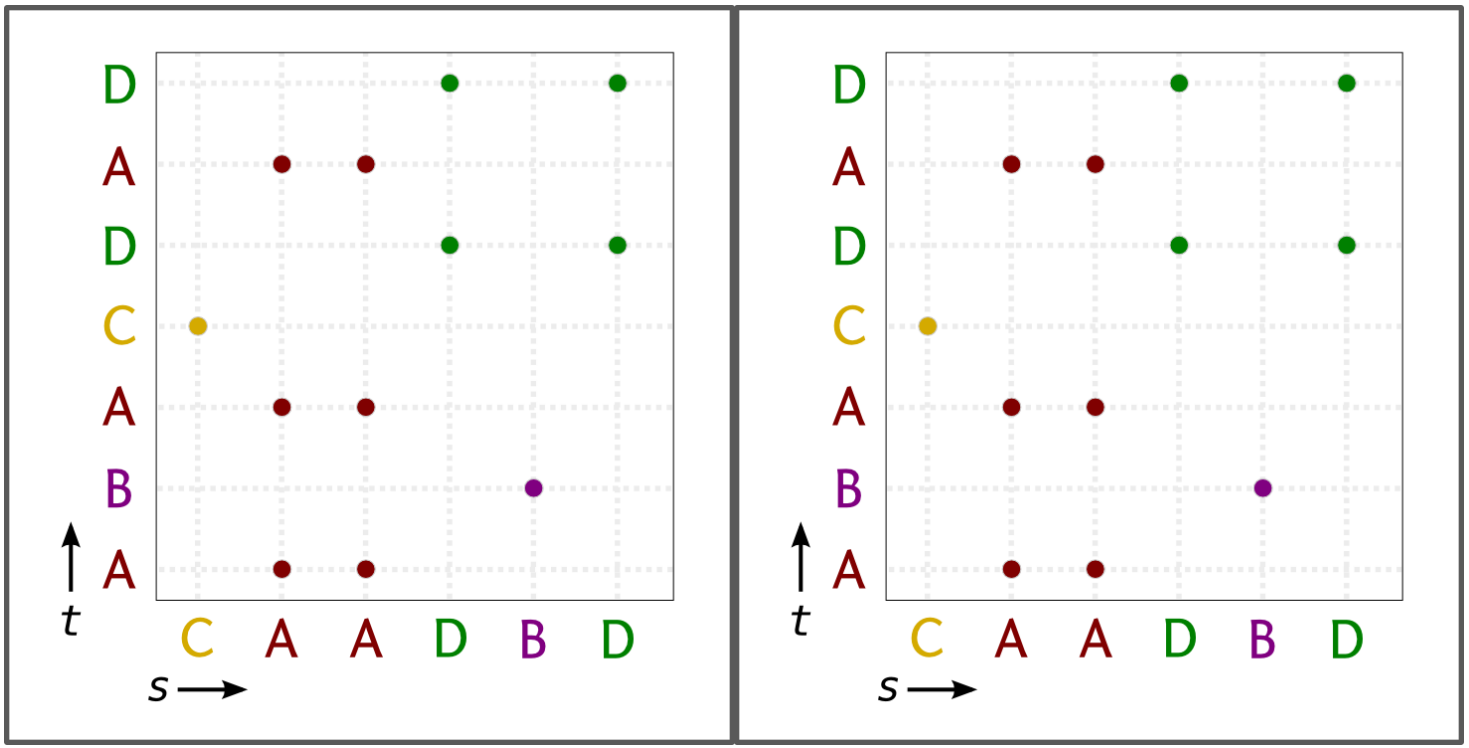




Dois grafos associados ao problema

Conjunto de vértices = conjunto de casamentos.

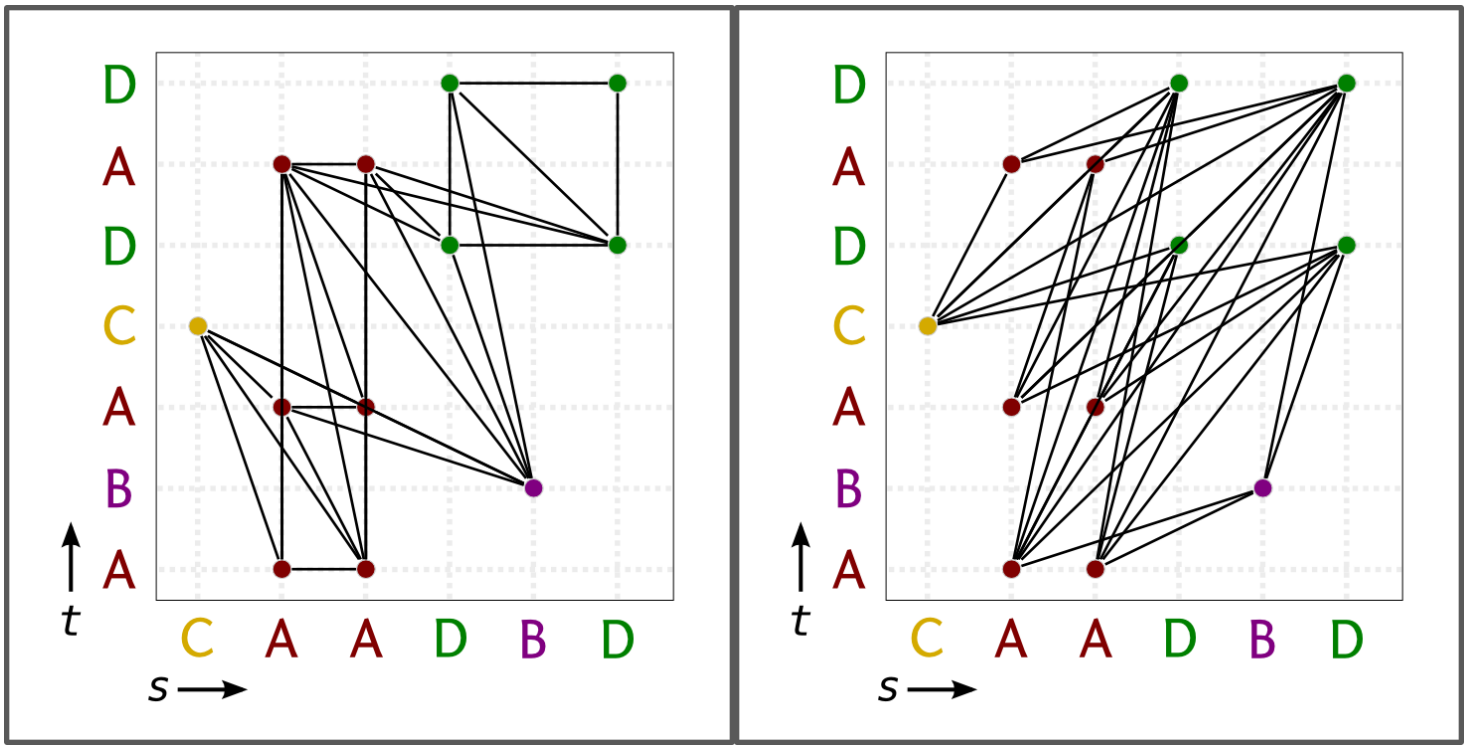
- Grafo de cruzamento (G_C): dois vértices são adjacentes se e só se eles se cruzam.
- Grafo de não-cruzamento (G_{NC}): dois vértices são adjacentes se e só se eles não se cruzam.



Dois grafos associados ao problema

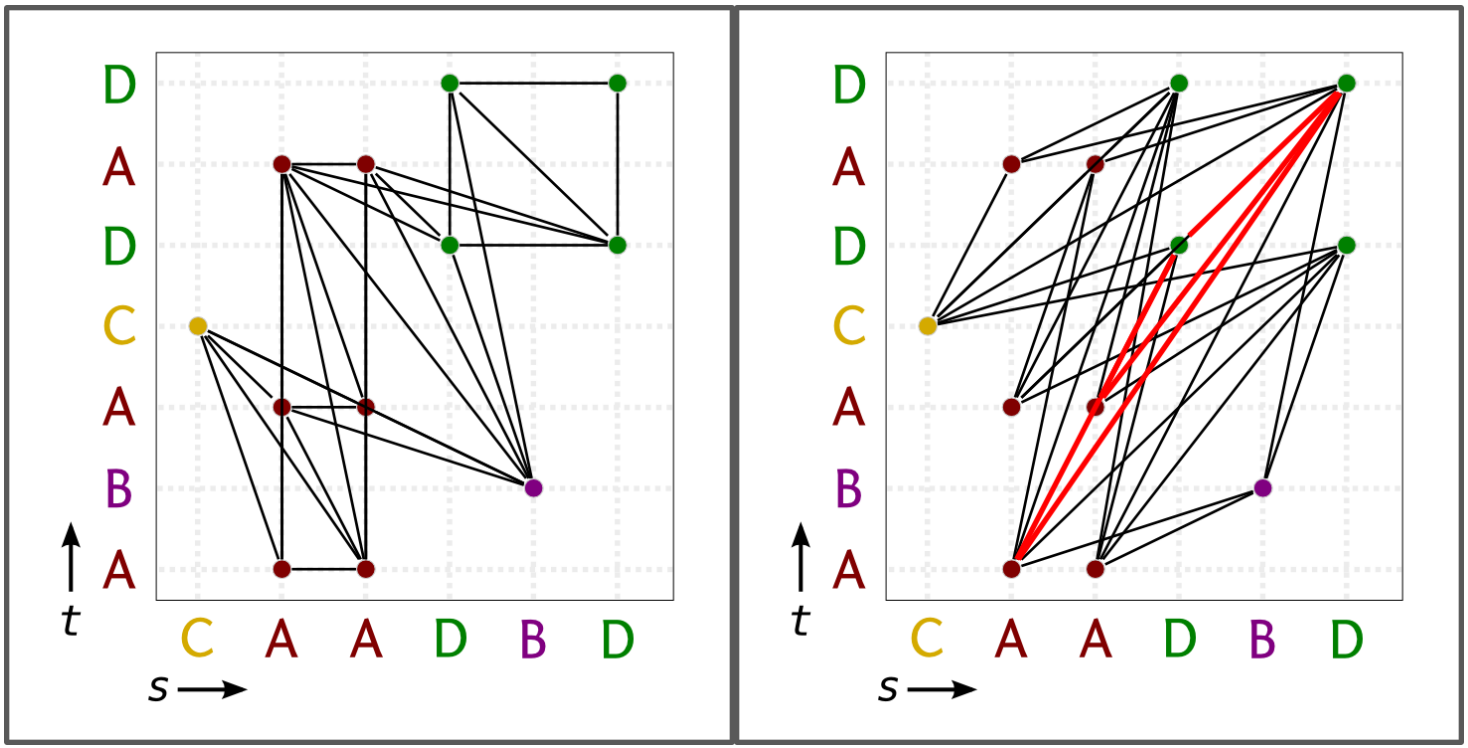
Conjunto de vértices = conjunto de casamentos.

- Grafo de cruzamento (G_C): dois vértices são adjacentes se e só se eles se cruzam.
- Grafo de não-cruzamento (G_{NC}): dois vértices são adjacentes se e só se eles não se cruzam.



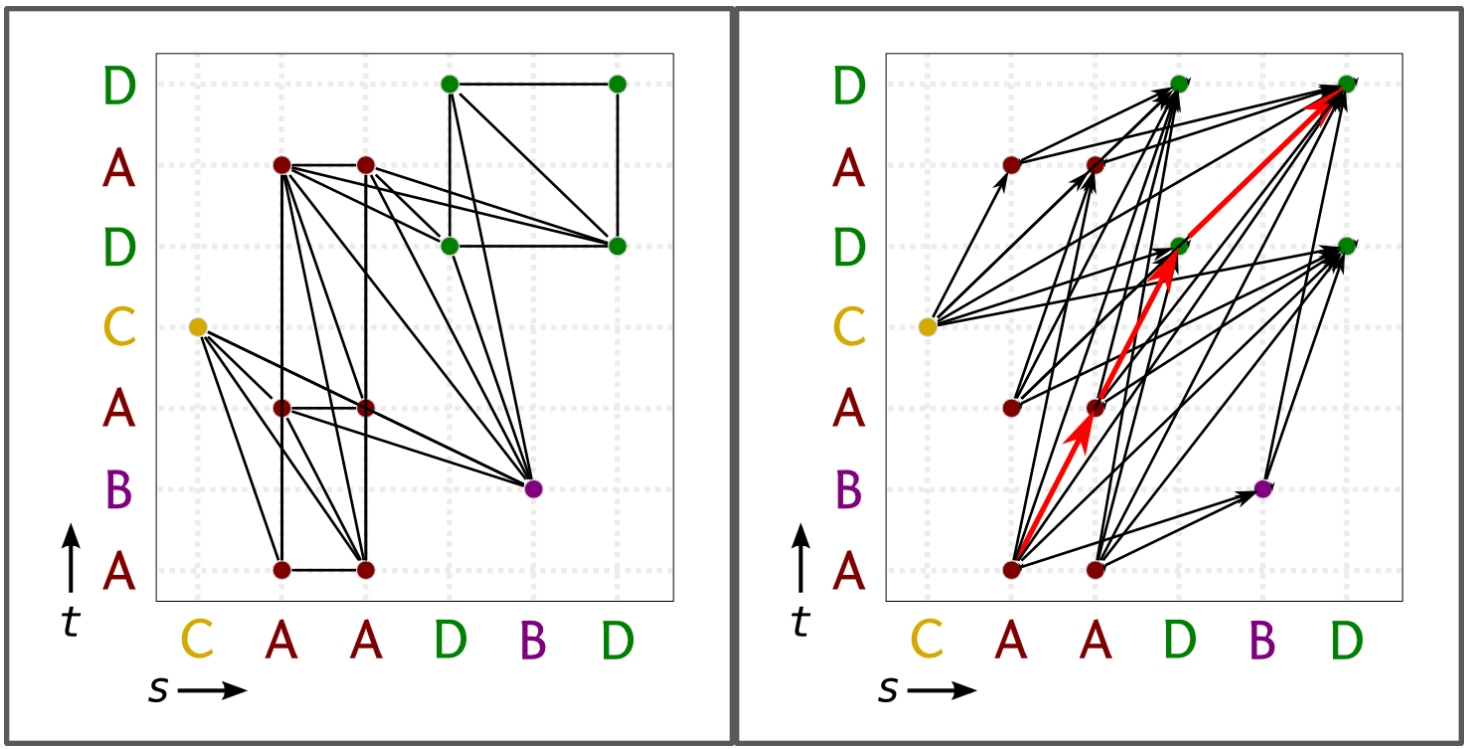
Dois grafos associados ao problema

LCS: clique máximo em G_{NC}



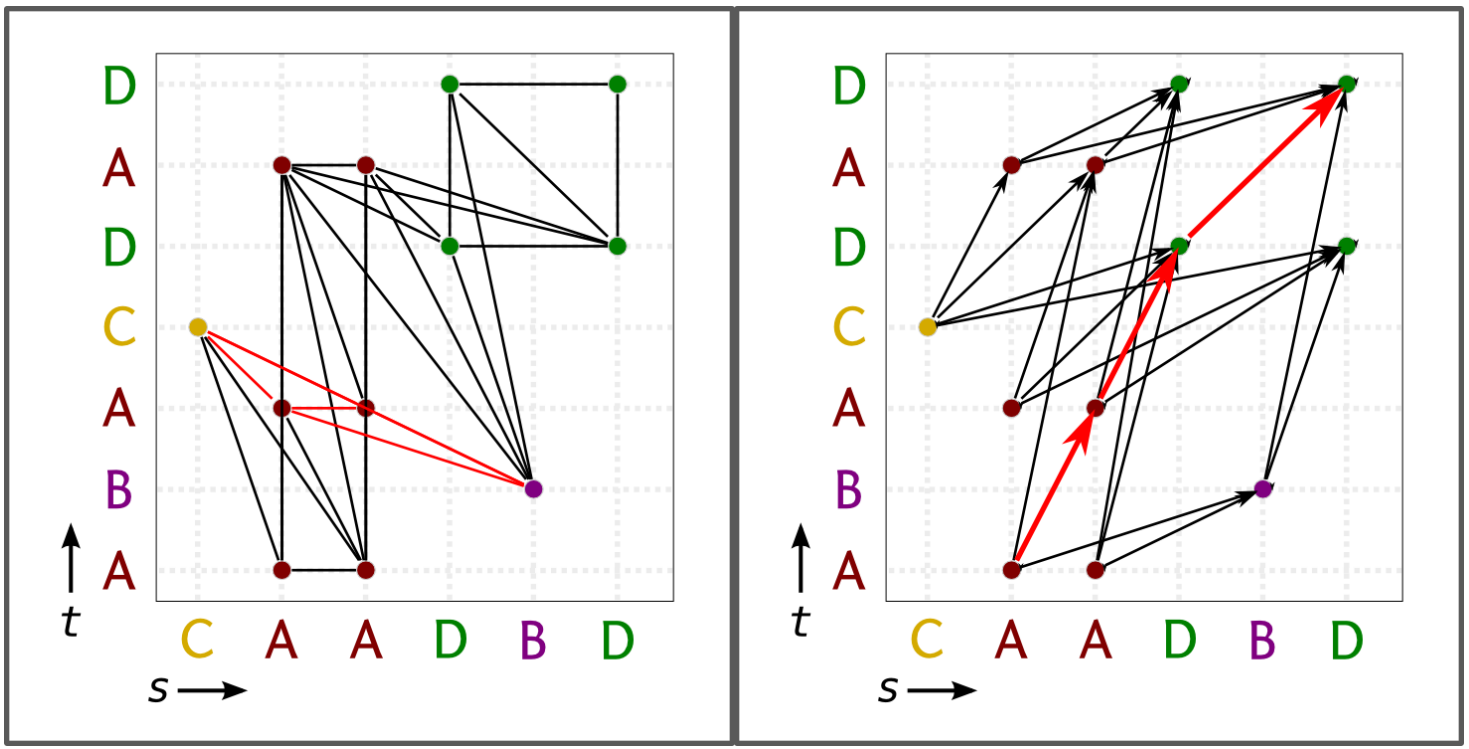
Dois grafos associados ao problema

LCS: clique máximo em G_{NC} ou caminho orientado de comprimento máximo em G_{NC} orientado.



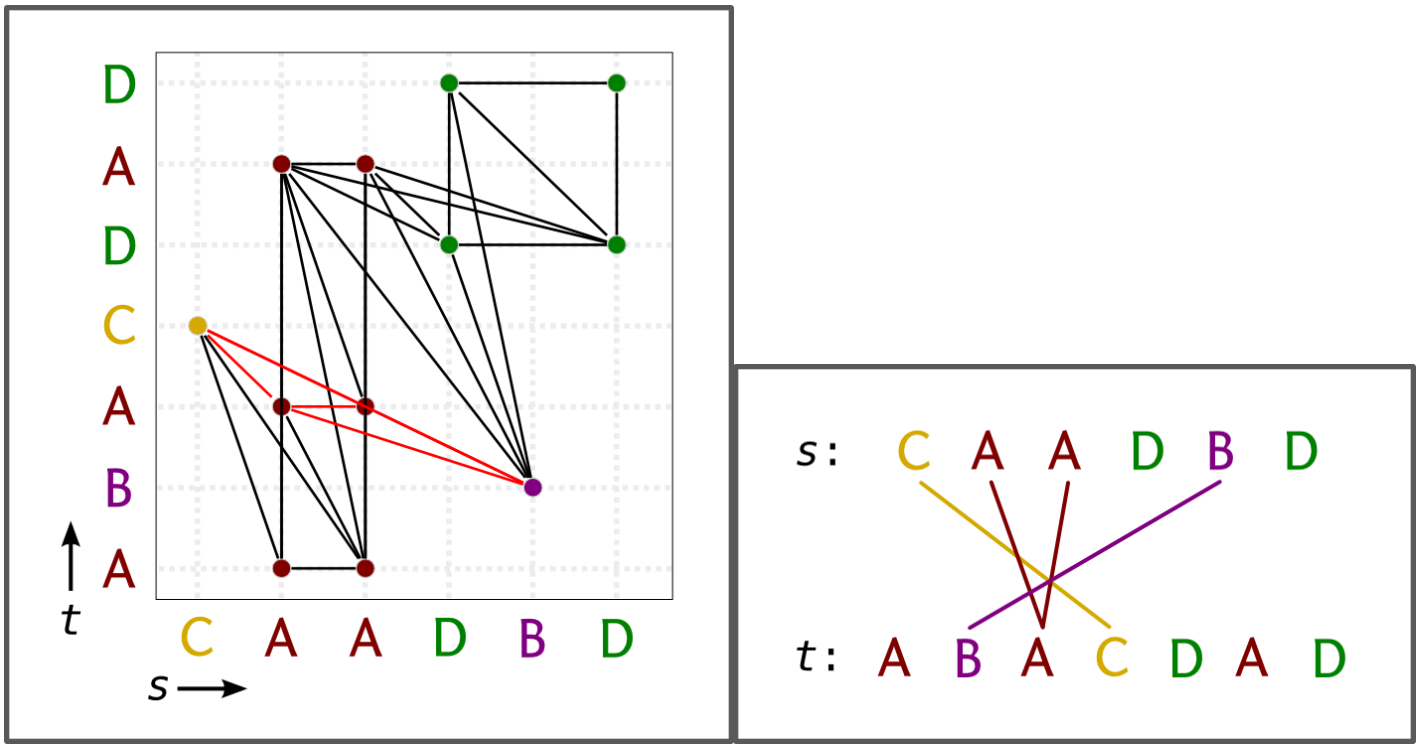
Dois grafos associados ao problema

LCS: clique máximo em G_{NC} ou caminho orientado de comprimento máximo em G_{NC} orientado.
E em G_C, o que são cliques maximais?



Dois grafos associados ao problema

LCS: clique máximo em G_{NC} ou caminho orientado de comprimento máximo em G_{NC} orientado.
E em G_C, o que são cliques maximais? Conjuntos maximais de casamentos que se cruzam dois a dois: estrelas maximais.



Melhorando a formulação

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} z_{ij} & & \\ \text{s. a} \quad & z_{ij} + z_{k\ell} \leq 1 & & \text{para todo } (i,j) \text{ e } (k,\ell) \\ & & & \text{em } C \text{ que se cruzam,} \\ & z_{ij} \in \{0,1\} & & \text{para todo } (i,j) \text{ em } C. \end{aligned}$$

Melhorando a formulação

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned}
\max \quad & \sum_{(i,j)\in E} z_{ij} \\
\text{s. a} \quad & \sum_{(i,j)\in S} z_{ij} \leq 1 \qquad \text{para toda estrela maximal } S, \\
& z_{ij} \in \{0,1\} \qquad \text{para todo } (i,j) \text{ em } C.
\end{aligned}$$

Consideremos agora o RFLCS

Nova restrição: para cada símbolo, há no máximo uma ocorrência.

Consideremos agora o RFLCS

Nova restrição: para cada símbolo, há no máximo uma ocorrência.

Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} z_{ij} \\ \text{s. a} \quad & \sum_{(i,j)\in S} z_{ij} \leq 1 && \text{para toda estrela maximal } S, \\ & z_{ij} \in \{0,1\} && \text{para todo } (i,j) \text{ em } C. \end{aligned}$$

Consideremos agora o RFLCS

Nova restrição: para cada símbolo, há no máximo uma ocorrência.

Dados: C: conjunto de todos casamentos.
$C(a)$: conjunto de todos casamentos de símbolo associado a.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

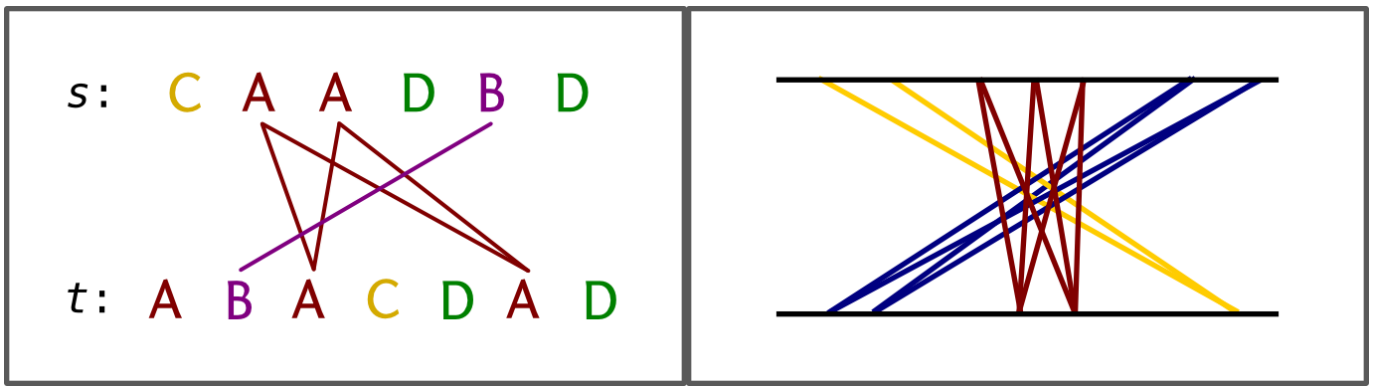
$$\begin{aligned}
\max \quad & \sum_{(i,j)\in E} z_{ij} & & \\
\text{s. a} \quad & \sum_{(i,j)\in C(a)} z_{ij} \leq 1 & & \text{para todo } a \text{ do alfabeto,} \\
& \sum_{(i,j)\in S} z_{ij} \leq 1 & & \text{para toda estrela maximal } S, \\
& z_{ij} \in \{0,1\} & & \text{para todo } (i,j) \text{ em } C.
\end{aligned}$$

Formulação melhor?

- A menos que P = NP, não podemos obter uma descrição explícita do casco convexo pois o problema do RFLCS é NP-difícil.

Formulação melhor?

- A menos que P = NP, não podemos obter uma descrição explícita do casco convexo pois o problema do RFLCS é NP-difícil.
- Defina estrela estendida como um conjunto de casamentos que, dois a dois, ou se cruzam ou têm o mesmo símbolo associado.



A formulação com estrelas estendidas

Dados: C: conjunto de todos casamentos.
$C(a)$: conjunto de todos casamentos de símbolo associado a.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned}
\max \quad & \sum_{(i,j)\in E} z_{ij} & & & \\
\text{s. a} \quad & \sum_{(i,j)\in C(a)} z_{ij} & \leq & 1 & \text{para todo } a \text{ do alfabeto,} \\
& \sum_{(i,j)\in S} z_{ij} & \leq & 1 & \text{para toda estrela maximal } S, \\
& z_{ij} & \in & \{0,1\} & \text{para todo } (i,j) \text{ em } C.
\end{aligned}$$

A formulação com estrelas estendidas

Dados: C: conjunto de todos casamentos.
$C(a)$: conjunto de todos casamentos de símbolo associado a.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned}
\max \quad & \sum_{(i,j)\in E} z_{ij} & & \\
\text{s. a} \quad & \sum_{(i,j)\in C(a)} z_{ij} \leq 1 & & \text{para todo } a \text{ do alfabeto,} \\
& \sum_{(i,j)\in S} z_{ij} \leq 1 & & \text{para toda estr. estendida max. } S, \\
& z_{ij} \in \{0,1\} & & \text{para todo } (i,j) \text{ em } C.
\end{aligned}$$

A formulação com estrelas estendidas

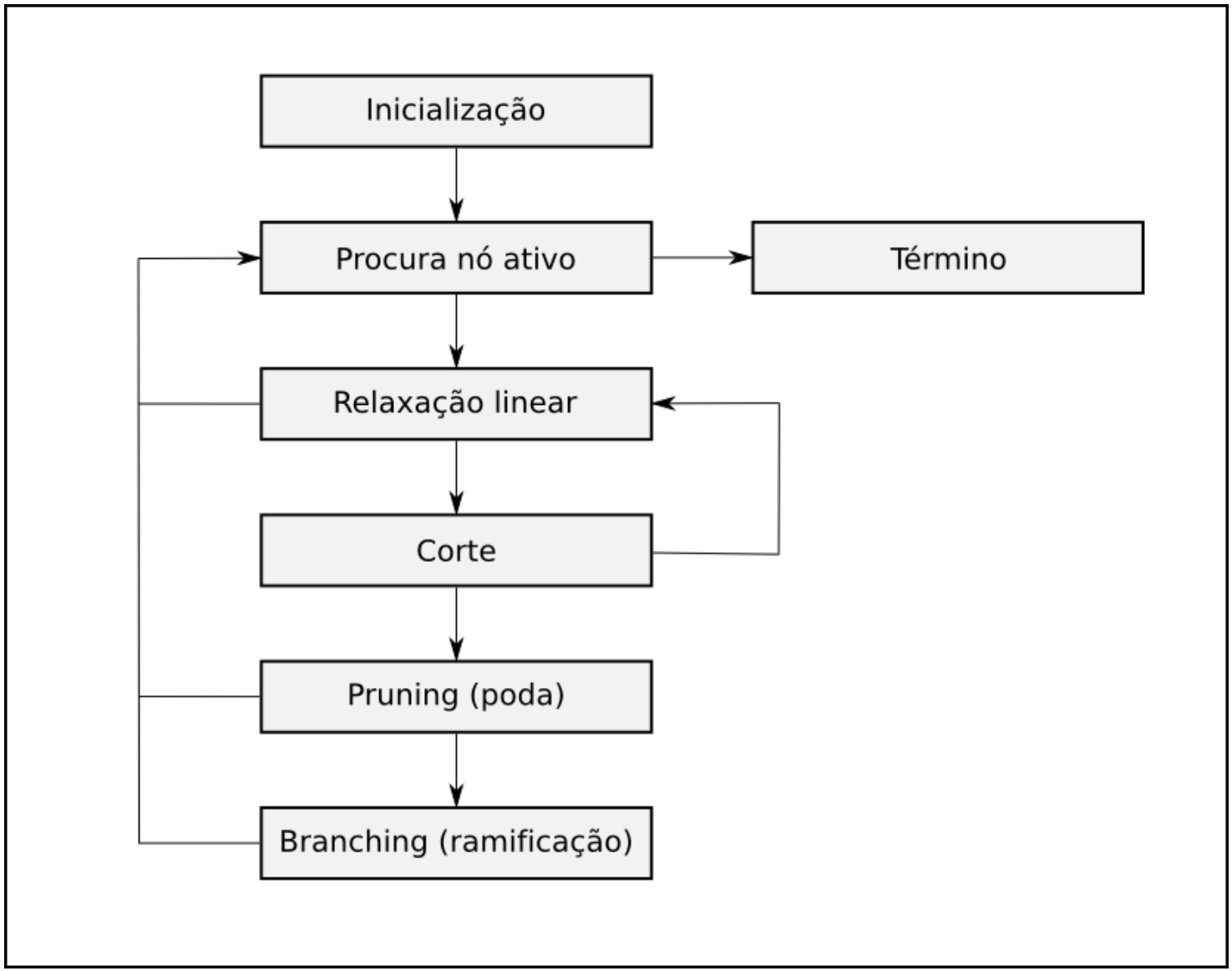
Dados: C: conjunto de todos casamentos.
Variáveis: z: vetor $z \in \{0,1\}^C$ onde $z_{ij} = 1$ se e só se escolhemos o casamento (i,j).

$$\begin{aligned} \max \quad & \sum_{(i,j)\in E} z_{ij} & & \\ \text{s. a} \quad & \sum_{(i,j)\in S} z_{ij} \leq 1 & & \text{para toda estr. estendida max. } S, \\ & z_{ij} \in \{0,1\} & & \text{para todo } (i,j) \text{ em } C. \end{aligned}$$

Implementação do algoritmo

- Pacote `glpk` (GNU Linear Programming Kit)
- Número exponencial de restrições
- Técnica de enumeração implícita com planos de corte *branch and cut*

Implementação do algoritmo



Resultados

- Formulações para os problemas do LCS e RFLCS
- Implementação de um algoritmo para a resolução do problema do RFLCS