Connecting de Bruijn Graphs

Giulia Bernadini, Huiping Chen, Inge Li Gørtz, ***Christoffer Krogh***, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

Overview

- Previous Work
- This Work

Previous Work

Making de Bruijn Graphs Eulerian

Authors: Giulia Bernardini, Huiping Chen, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

- Part of: Volume: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)
  Series: Leibniz International Proceedings in Informatics (LIPIcs)
  Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

- Publication Date: 2022-06-22


Making de Bruijn Graphs Eulerian

de Bruijn Graphs

de Bruijn Graphs

- Collection of length k strings

de Bruijn Graphs

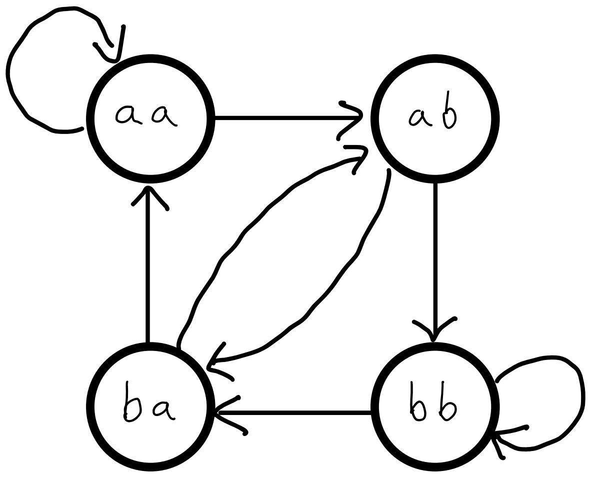
- Collection of length k strings
- Vertices are length $k-1$ subtrings

de Bruijn Graphs

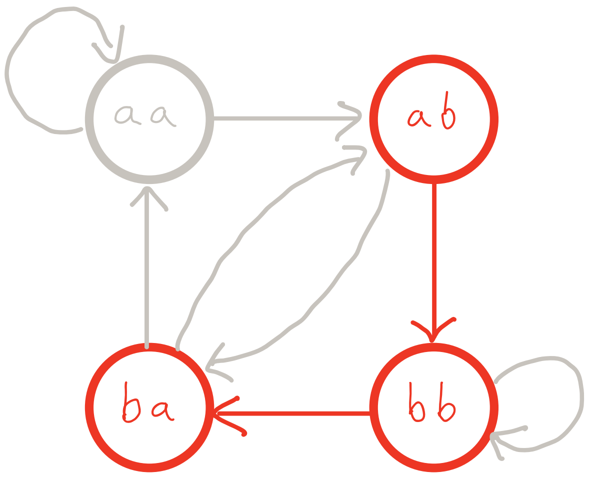
- Collection of length k strings
- Vertices are length $k - 1$ subtrings
- Edges iff corresponding string exists in collection

de Bruijn Graphs

- Collection of length k strings
- Vertices are length $k-1$ subtrings
- Edges iff corresponding string exists in collection



$\Sigma = \{a, b\}$
$k = 3$



$\{abb, bba\}$
$k = 3$

Eulerian Graphs

Eulerian Graphs

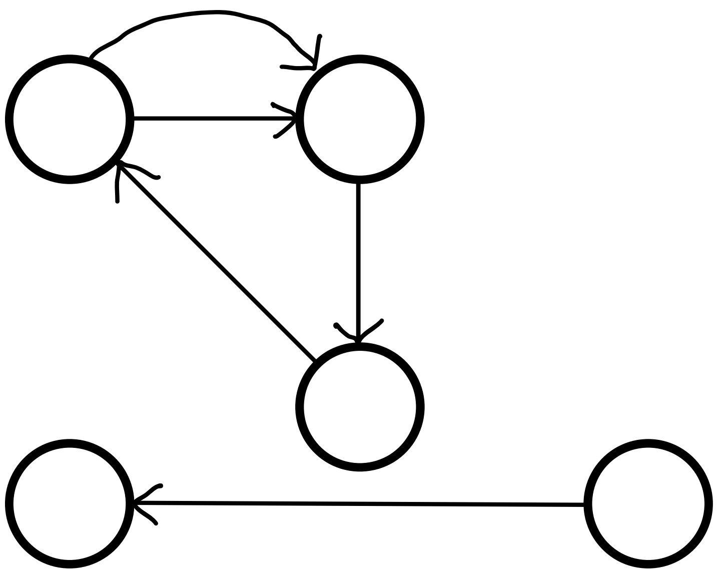
- Circuit of every edge exactly once

Eulerian Graphs

- Circuit of every edge exactly once
- Euler’s Theorem:
    1. Edges must be connected
    2. Vertices must be balanced

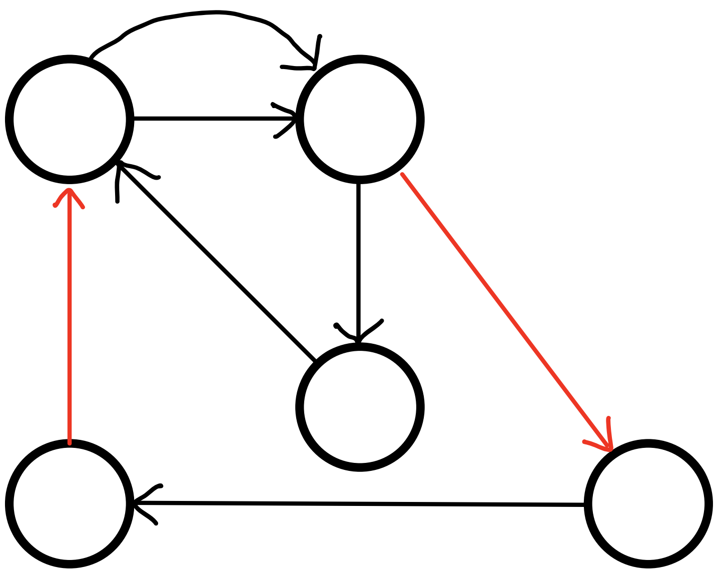
Eulerian Graphs

- Circuit of every edge exactly once
- Euler's Theorem:
  1. Edges must be connected
  2. Vertices must be balanced



Eulerian Graphs

- Circuit of every edge exactly once
- Euler's Theorem:
  1. Edges must be connected
  2. Vertices must be balanced



Main Problem

Making de Bruijn Graphs Eulerian

Authors: Giulia Bernardini, Huiping Chen, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

- Part of:
  - Volume: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)
  - Series: Leibniz International Proceedings in Informatics (LIPIcs)
  - Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

- Publication Date: 2022-06-22

Main Problem

Making de Bruijn Graphs Eulerian

Authors Giulia Bernardini, Huiping Chen, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

Part of: Volume: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)
Series: Leibniz International Proceedings in Informatics (LIPIcs)
Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

Publication Date: 2022-06-22


Making de Bruijn Graphs Eulerian

Eulerian Extension of de Bruijn Graphs (EXTEND-DBG)

Main Problem

Making de Bruijn Graphs Eulerian

Authors Giulia Bernardini, Huiping Chen, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

- Part of: Volume: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)
  Series: Leibniz International Proceedings in Informatics (LIPIcs)
  Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)
- License: Creative Commons Attribution 4.0 International license
- Publication Date: 2022-06-22


Making de Bruijn Graphs Eulerian

Eulerian Extension of de Bruijn Graphs (EXTEND-DBG)

- Given a de Bruijn graph

Main Problem

Making de Bruijn Graphs Eulerian

Authors ⓘ Giulia Bernardini, Huiping Chen, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

› Part of: Volume: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)
Series: Leibniz International Proceedings in Informatics (LIPIcs)
Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

› Publication Date: 2022-06-22

Eulerian Extension of de Bruijn Graphs (EXTEND-DBG)

- Given a de Bruijn graph
- Make Eulerian from the complete de Bruijn graph

Main Problem

Making de Bruijn Graphs Eulerian

Authors Giulia Bernardini, Huiping Chen, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

- Part of: Volume: 33rd Annual Symposium on Combinatorial Pattern Matching (CPM 2022)
  Series: Leibniz International Proceedings in Informatics (LIPIcs)
  Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)
- License: CC BY Creative Commons Attribution 4.0 International license
- Publication Date: 2022-06-22


Making de Bruijn Graphs Eulerian

Eulerian Extension of de Bruijn Graphs (EXTEND-DBG)

- Given a de Bruijn graph
- Make Eulerian from the complete de Bruijn graph
- Minimize number of new edges

Motivation

DNA sequencing

Motivation

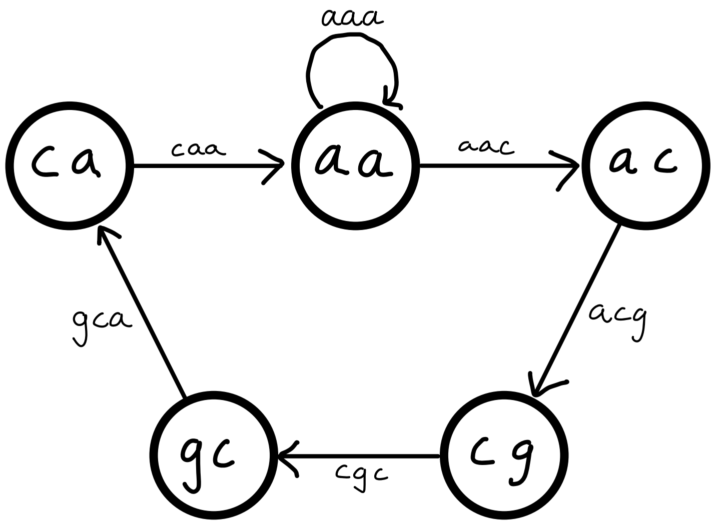
DNA sequencing

- $caaacgca \Rightarrow \{caa, aaa, aac, acg, cgc, gca\}$

Motivation

DNA sequencing

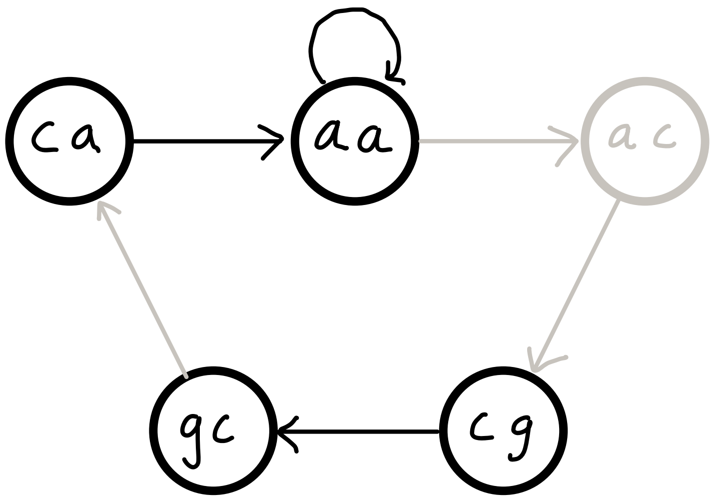
- $caaacgca \Rightarrow \{caa, aaa, aac, acg, cgc, gca\}$



Motivation

DNA sequencing

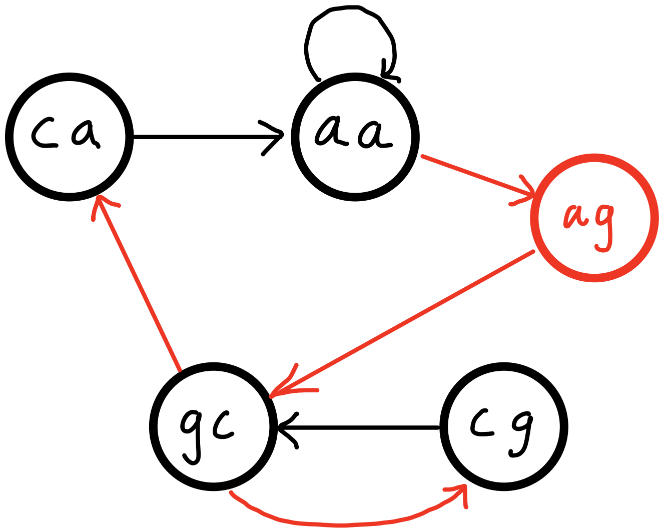
- $caaacgca \Rightarrow \{caa, aaa, \underline{aac}, \underline{acg}, cgc, \underline{gca}\}$



Motivation

DNA sequencing

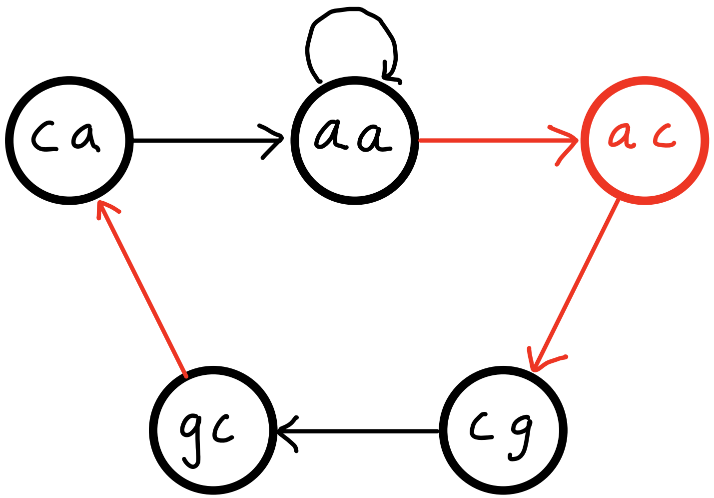
- $caaacgca \Rightarrow \{caa, aaa, \underline{aac}, \underline{acg}, cgc, \underline{gca}\}$



Motivation

DNA sequencing

- $caaacgca \Rightarrow \{caa, aaa, \underline{aac}, \underline{acg}, cgc, \underline{gca}\}$



Hardness

- EXTEND-DBG is NP-hard
  - even when only adding edges

Hardness

- EXTEND-DBG is NP-hard
  - even when only adding edges
- Split problem in two:

Hardness

- Extend-DBG is NP-hard
  - even when only adding edges
- Split problem in two:

  1. Connect de Bruijn Graph

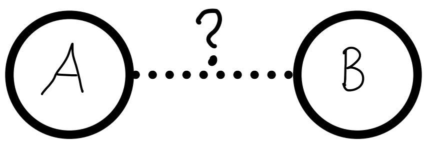
Hardness

- EXTEND-DBG is NP-hard
  - even when only adding edges
- Split problem in two:
  1. Connect de Bruijn Graph
  2. Balance de Bruijn Graph

Connect de Bruijn Graph

$$A = A_L \, X \, A_R$$

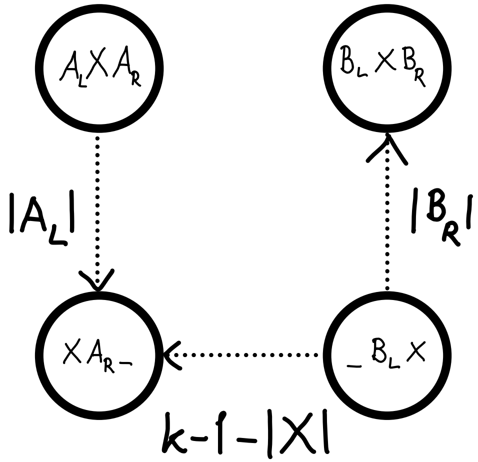
$$B = B_L \, X \, B_R$$



Connect de Bruijn Graph

$A = A_L \, X \, A_R$

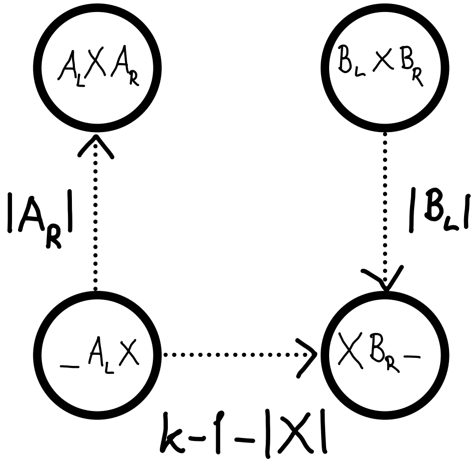
$B = B_L \, X \, B_R$



Connect de Bruijn Graph

$A = A_L \, X \, A_R$

$B = B_L \, X \, B_R$



Connect de Bruijn Graph

$A = A_L \, X \, A_R$

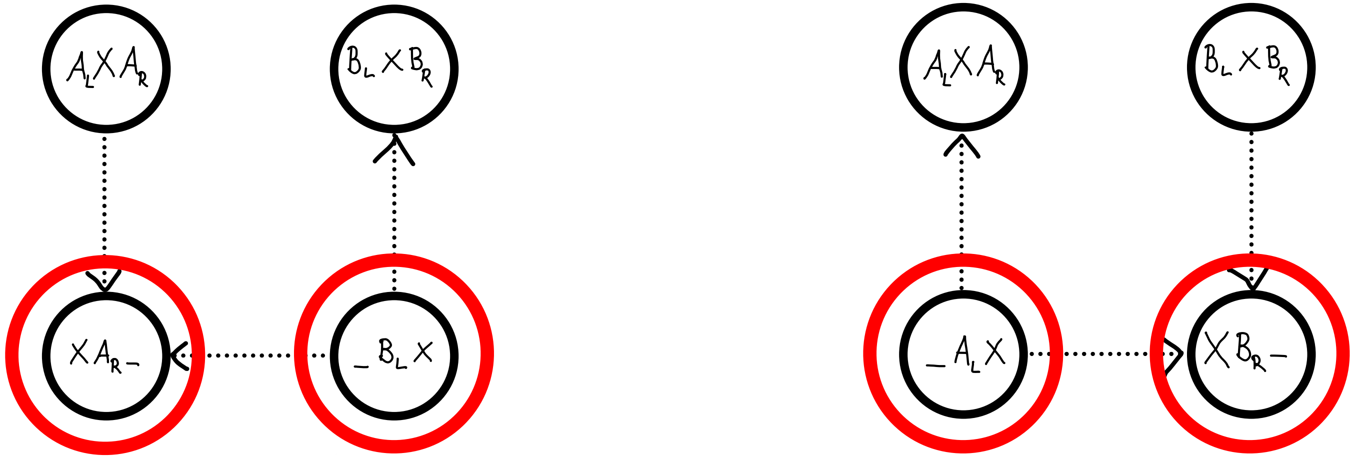
$B = B_L \, X \, B_R$

$$d(A,B) = k - 1 - |X| + \min\{A_L + B_R, A_R + B_L\}$$

Connect de Bruijn Graph

$A = A_L \, X \, A_R$

$B = B_L \, X \, B_R$



$$d(A, B) = k - 1 - |X| + \min\{A_L + B_R, A_R + B_L\}$$

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

- Given a de Bruijn Graph

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

- Given a de Bruijn Graph
- Weakly connect adding only directed paths

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

- Given a de Bruijn Graph
- Weakly connect adding only directed paths
- Minimize number of new edges

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

- Given a de Bruijn Graph
- Weakly connect adding only directed paths
- Minimize number of new edges

Solve in $\mathcal{O}(|V| k \log d + |E|)$ time,

d is the number of connected components

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

- Given a de Bruijn Graph
- Weakly connect adding only directed paths
- Minimize number of new edges

Solve in $\mathcal{O}(|V|k\log d + |E|)$ time,

d is the number of connected components

Balance de Bruijn Graph

BALANCE-DBG

- Given a de Bruijn Graph
- **Balance all vertices**
- Minimize number of new edges

Connect de Bruijn Graph with Paths

CONNECT-DBG-P

- Given a de Bruijn Graph
- Weakly connect adding only directed paths
- Minimize number of new edges

Solve in $\mathcal{O}(|V|k \log d + |E|)$ time,

d is the number of connected components

Balance de Bruijn Graph

BALANCE-DBG

- Given a de Bruijn Graph
- Balance all vertices
- Minimize number of new edges

Solve in $\mathcal{O}(k|V| + |E| + |A|)$ time,

$|A|$ is the number of added edges

Overview

- Previous Work
- This Work

DTU

This Work

Connecting de Bruijn Graphs

Authors Giulia Bernardini, Huiping Chen, Inge Li Gørtz, Christoffer Krogh, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

- Part of: Volume: 35th Annual Symposium on Combinatorial Pattern Matching (CPM 2024)
  Series: Leibniz International Proceedings in Informatics (LIPIcs)
  Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

- Publication Date: 2024-06-18


Connecting de Bruijn Graphs

This Work

Connecting de Bruijn Graphs

Authors Giulia Bernardini, Huiping Chen, Inge Li Gørtz, Christoffer Krogh, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

Part of: Volume: 35th Annual Symposium on Combinatorial Pattern Matching (CPM 2024)
Series: Leibniz International Proceedings in Informatics (LIPIcs)
Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

Publication Date: 2024-06-18

Results

1. Connecting de Bruijn Graphs is NP-hard

This Work

Connecting de Bruijn Graphs

Authors Giulia Bernardini, Huiping Chen, Inge Li Gørtz, Christoffer Krogh, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

Part of: Volume: 35th Annual Symposium on Combinatorial Pattern Matching (CPM 2024)
Series: Leibniz International Proceedings in Informatics (LIPIcs)
Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

Publication Date: 2024-06-18


Connecting de Bruijn Graphs

Results

1. Connecting de Bruijn Graphs is NP-hard

2. 2-approximation for CONNECT-DBG

This Work

Connecting de Bruijn Graphs

Authors Giulia Bernardini, Huiping Chen, Inge Li Gørtz, Christoffer Krogh, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

Part of: Volume: 35th Annual Symposium on Combinatorial Pattern Matching (CPM 2024)
Series: Leibniz International Proceedings in Informatics (LIPIcs)
Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

Publication Date: 2024-06-18

Results

1. Connecting de Bruijn Graphs is NP-hard
2. 2-approximation for CONNECT-DBG*
3. Improved and simplified solution to CONNECT-DBG-P

This Work

Connecting de Bruijn Graphs

Authors Giulia Bernardini, Huiping Chen, Inge Li Gørtz, Christoffer Krogh, Grigorios Loukides, Solon P. Pissis, Leen Stougie, Michelle Sweering

Part of: Volume: 35th Annual Symposium on Combinatorial Pattern Matching (CPM 2024)
Series: Leibniz International Proceedings in Informatics (LIPIcs)
Conference: Annual Symposium on Combinatorial Pattern Matching (CPM)

Publication Date: 2024-06-18


Connecting de Bruijn Graphs

Results

1. Connecting de Bruijn Graphs is NP-hard
2. 2-approximation for Connect-DBG*
3. Improved and simplified solution to Connect-DBG-P
4. Integer linear program formulation

Hardness of CONNECT-DBG

Hardness of Connect-DBG

Reduction from Vertex Cover

Hardness of Connect-DBG

Reduction from

Vertex Cover

- Given an undirected graph

Hardness of Connect-DBG

Reduction from

Vertex Cover

- Given an undirected graph
- Choose vertices such that at least one endpoint of every edge is chosen

Hardness of Connect-DBG

Reduction from

Vertex Cover

- Given an undirected graph
- Choose vertices such that at least one endpoint of every edge is chosen
- Minimize number of chosen vertices

Hardness of Connect-DBG

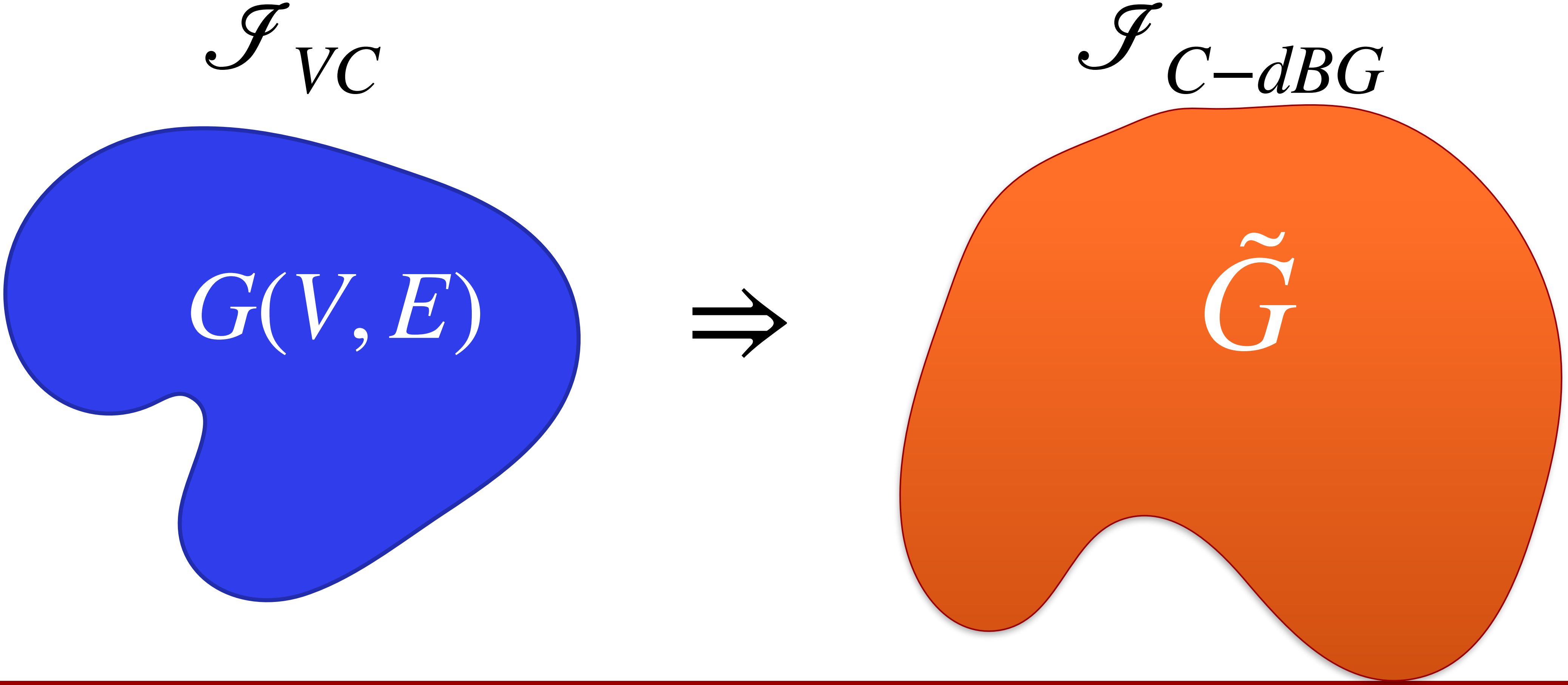
Reduction from Vertex Cover

$$\mathcal{I}_{VC} = G(V,E) \quad \Rightarrow \quad \mathcal{I}_{C-dBG}$$

Hardness of Connect-DBG

Reduction from Vertex Cover

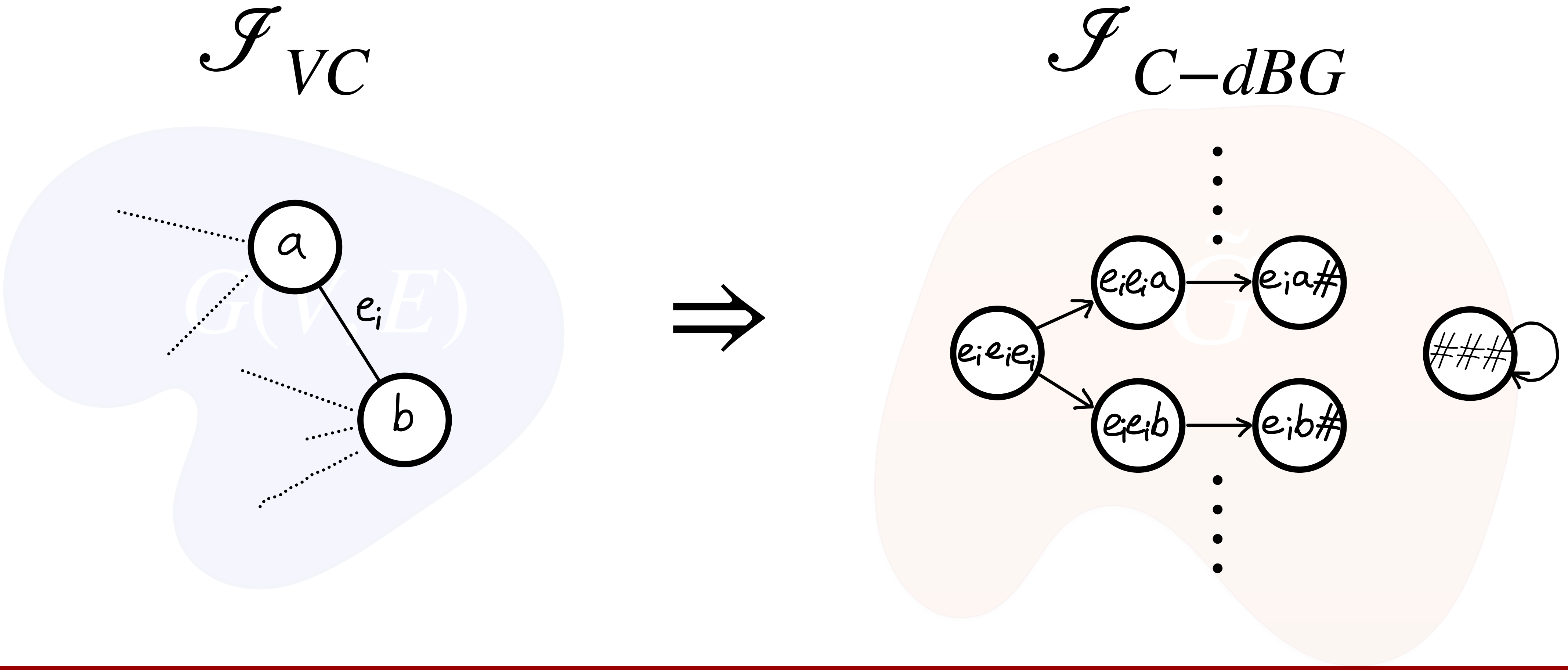
$\mathcal{I}_{VC} = G(V,E) \quad \Rightarrow \quad \mathcal{I}_{C-dBG}$



Hardness of Connect-DBG

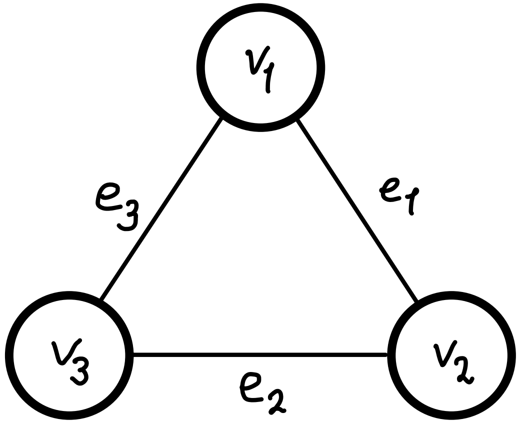
Reduction from Vertex Cover

$\mathcal{I}_{VC} = G(V,E) \quad \Rightarrow \quad \mathcal{I}_{C\text{-}dBG}$

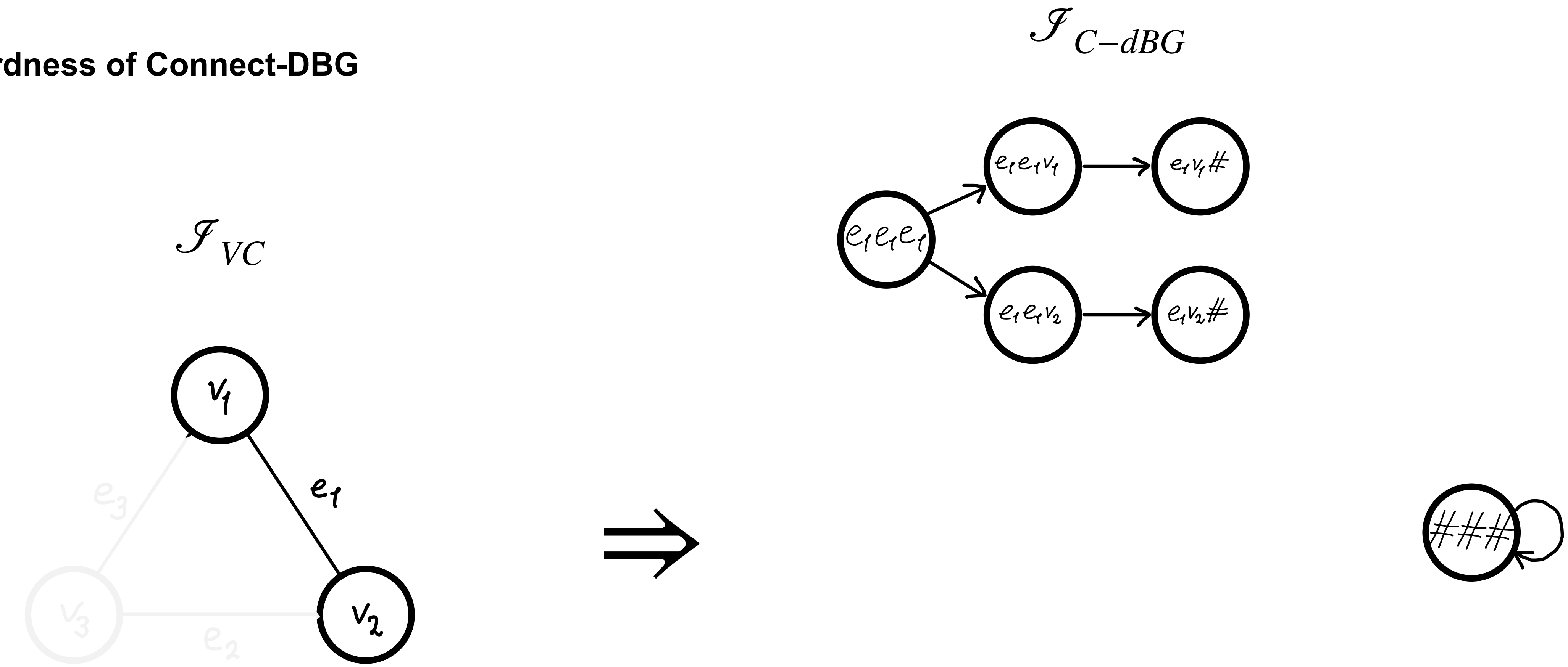


Hardness of Connect-DBG

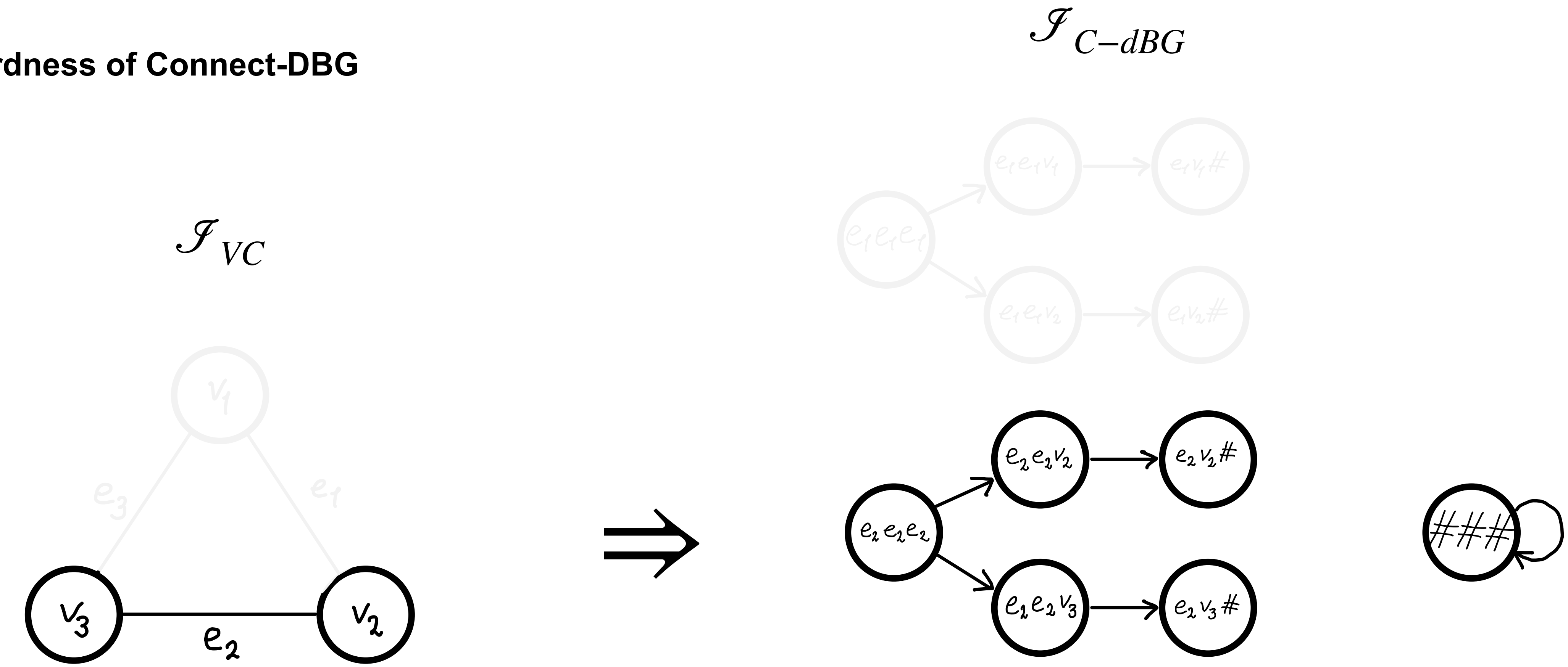
$\mathcal{I}_{VC}$



Hardness of Connect-DBG

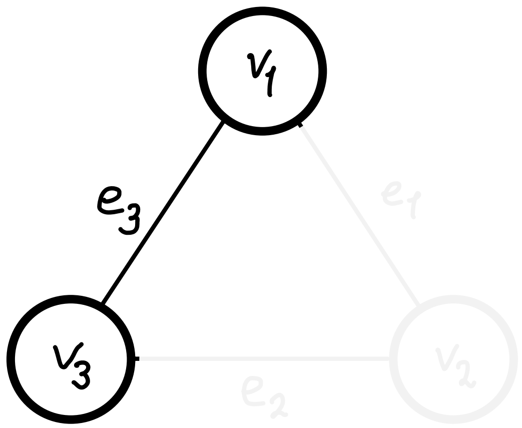


Hardness of Connect-DBG



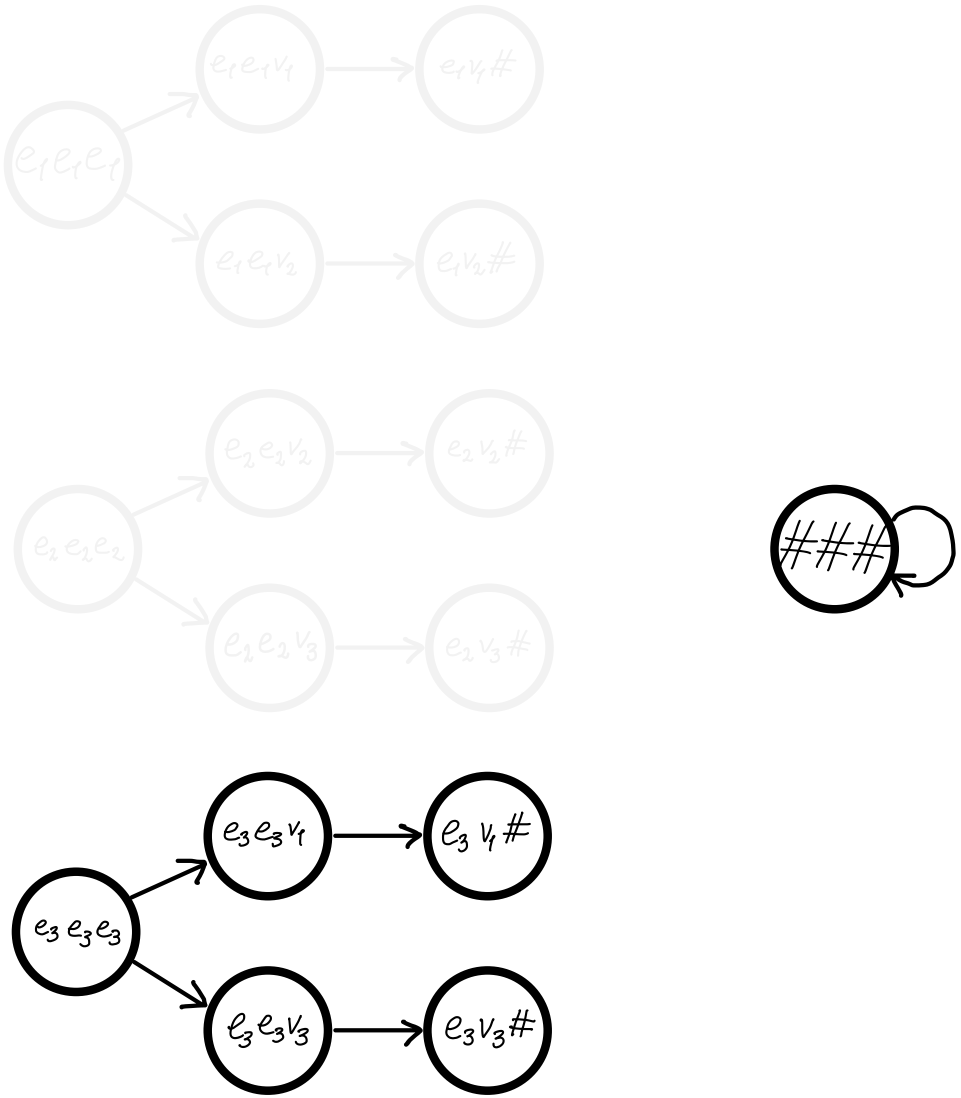
Hardness of Connect-DBG

$\mathcal{I}_{VC}$

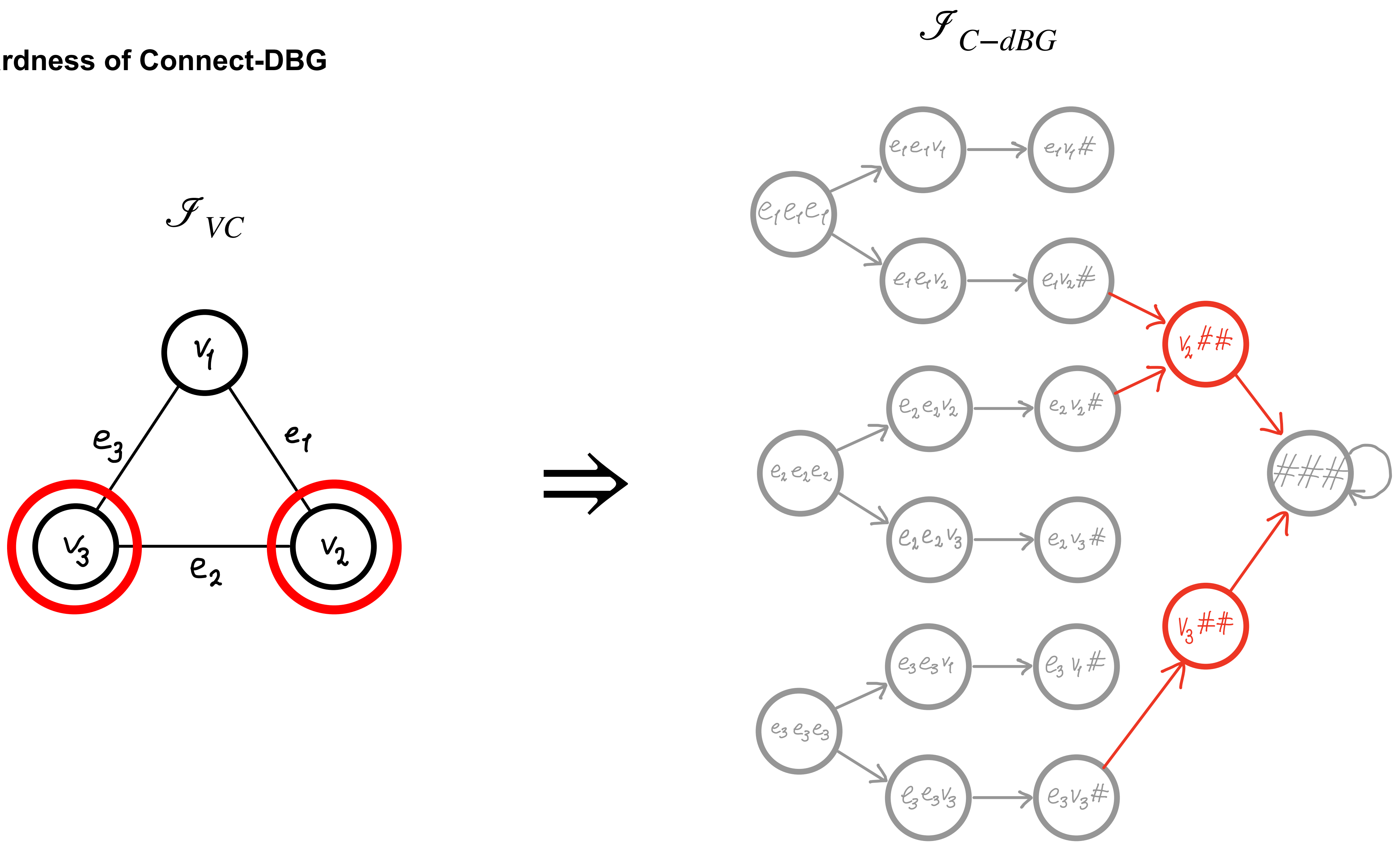


$\Rightarrow$

$\mathcal{I}_{C-dBG}$

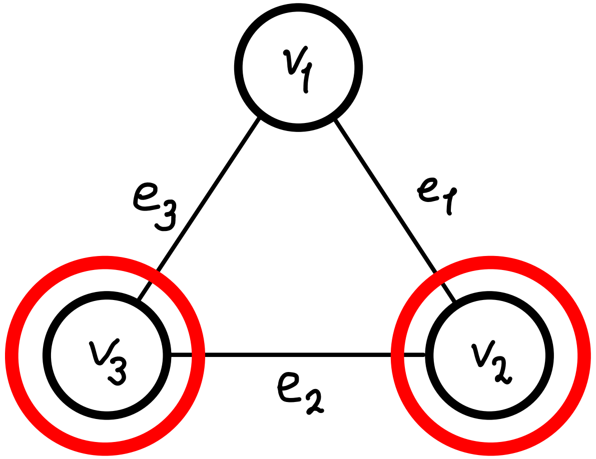


Hardness of Connect-DBG



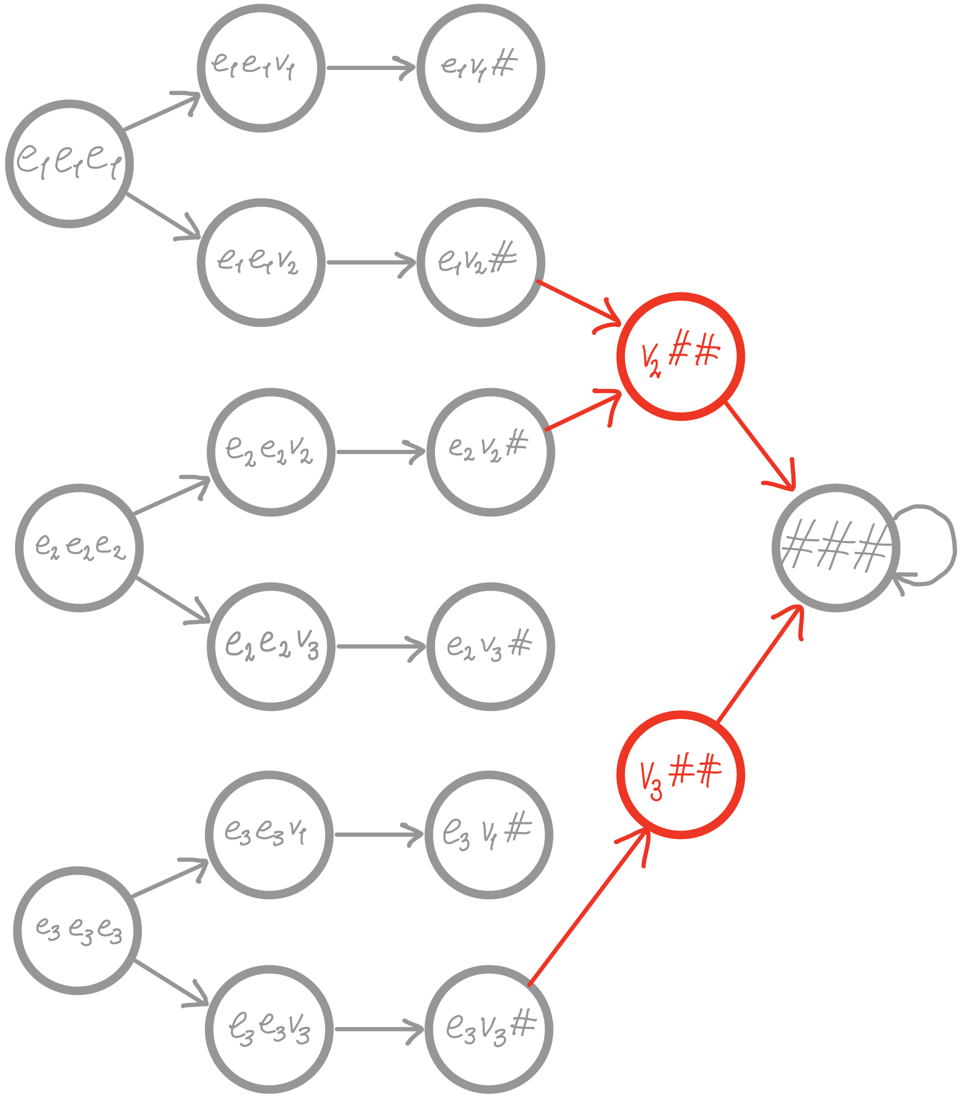
Hardness of Connect-DBG

$$OPT(\mathcal{I}_{VC}) = 2$$



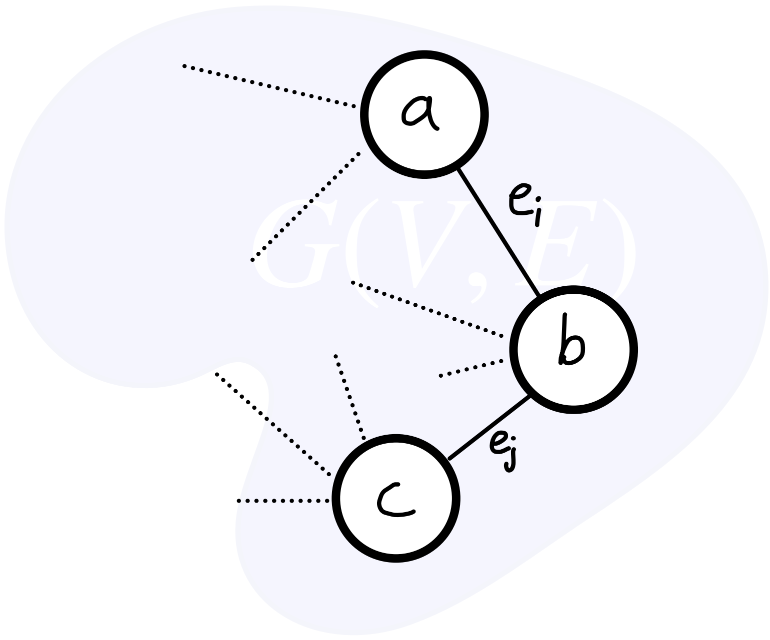
$\Rightarrow$

$$OPT(\mathcal{I}_{C-dBG}) = 2 + |E| = 5$$



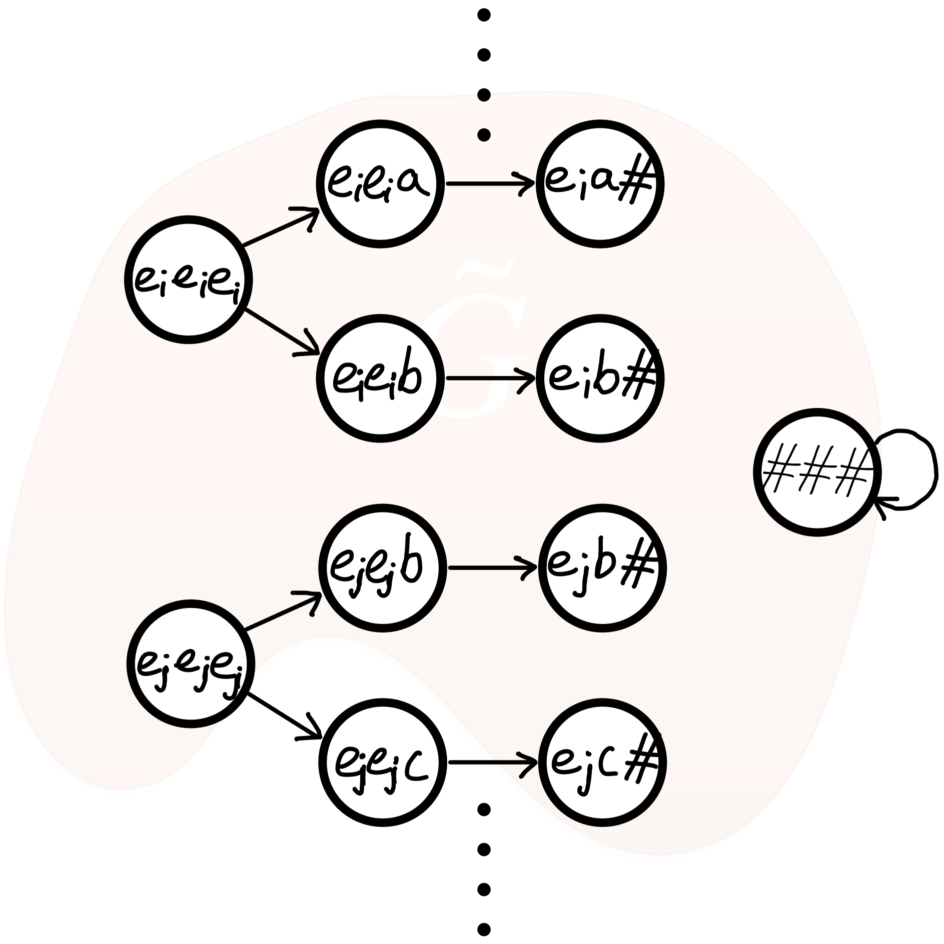
Hardness of Connect-DBG

$\mathcal{I}_{VC}$



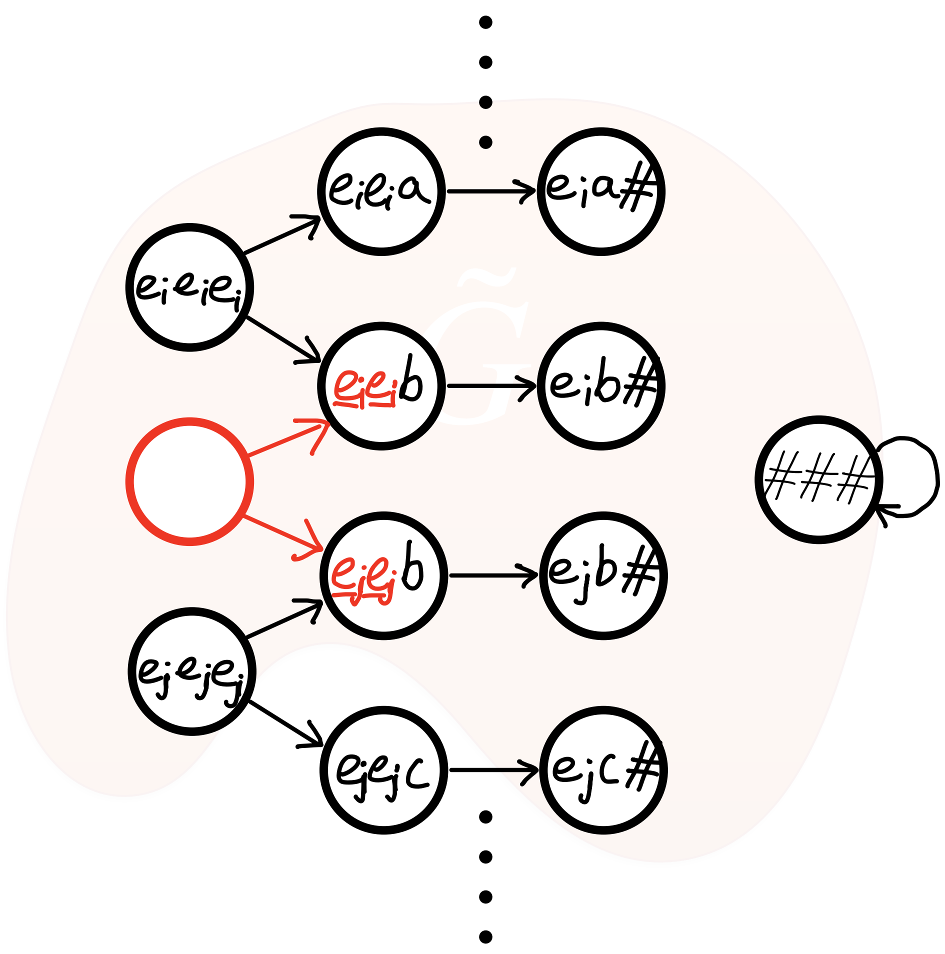
$\Leftarrow$

$\mathcal{I}_{C-dBG}$



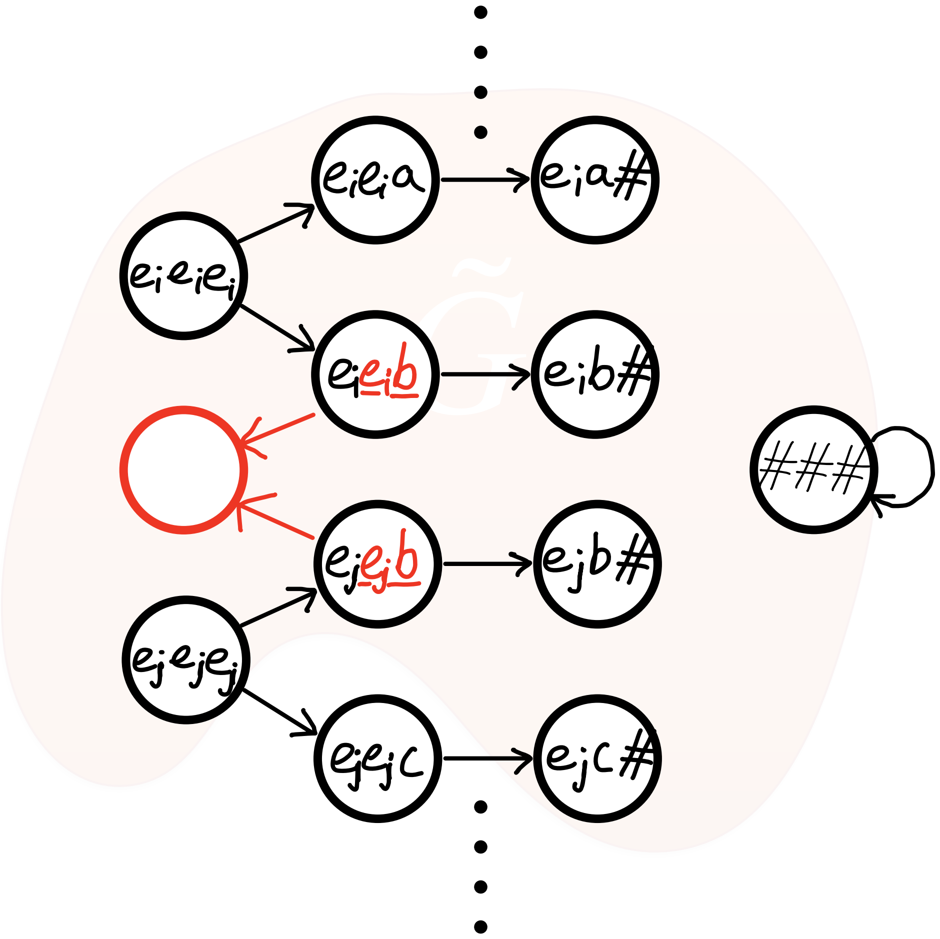
Hardness of Connect-DBG

$\mathcal{I}_{C-dBG}$



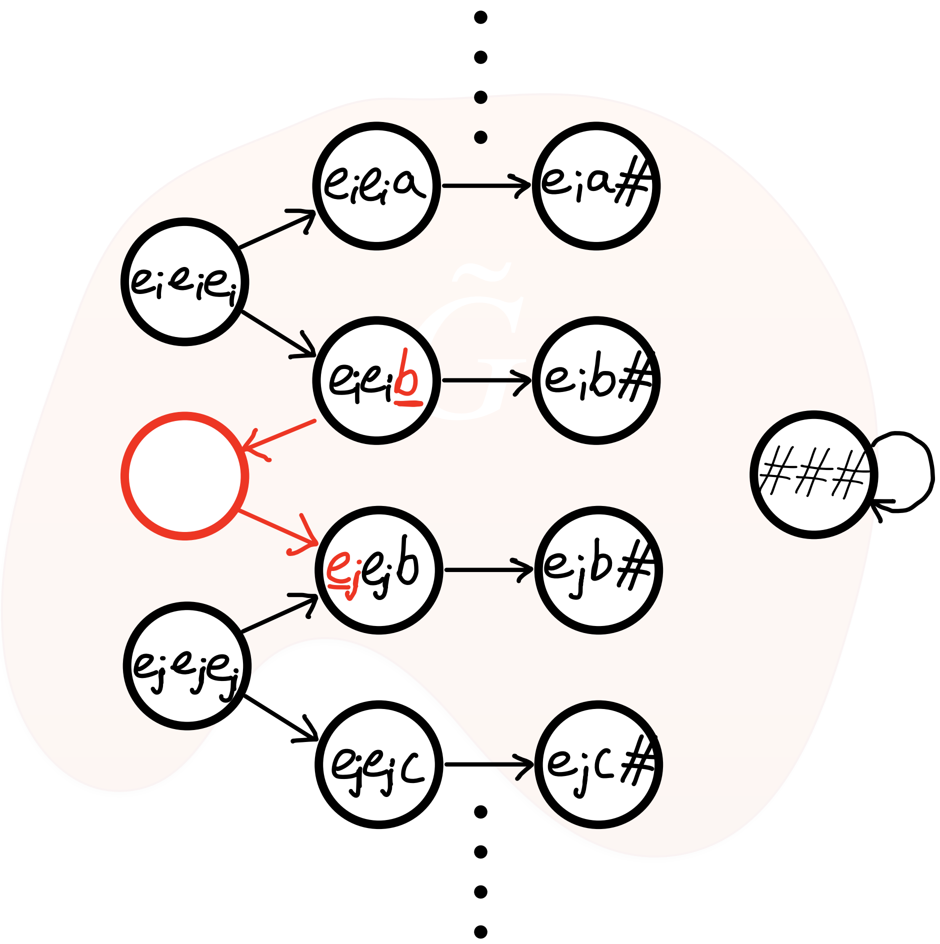
Hardness of Connect-DBG

$\mathcal{I}_{C\text{–}dBG}$



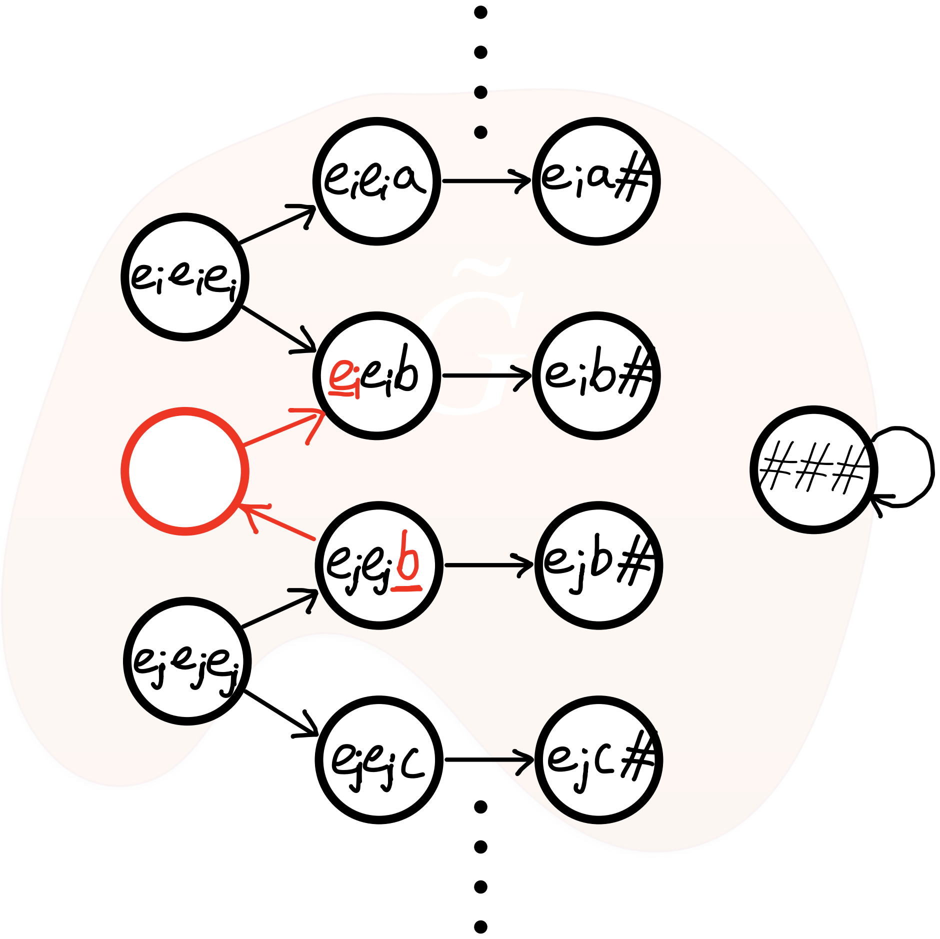
Hardness of Connect-DBG

$\mathcal{I}_{C-dBG}$



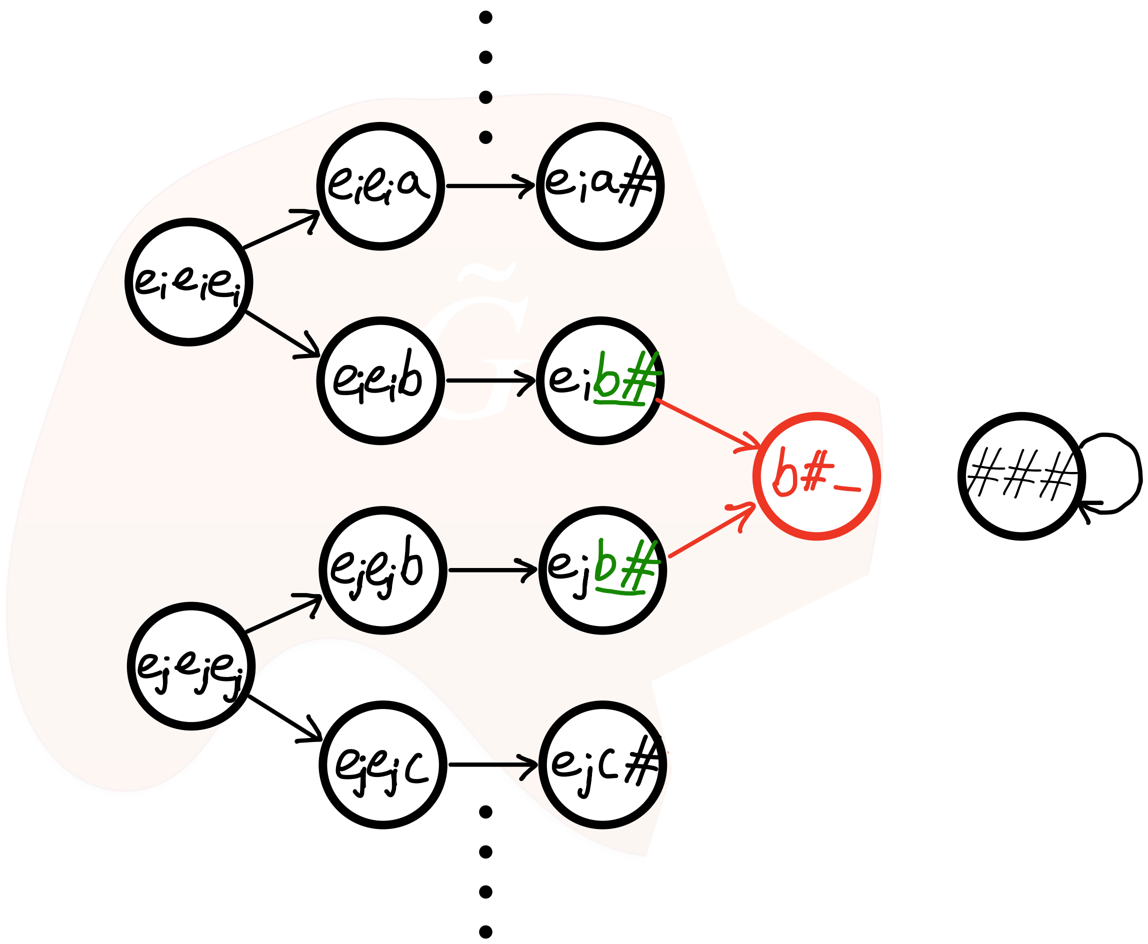
Hardness of Connect-DBG

$\mathcal{I}_{C-dBG}$



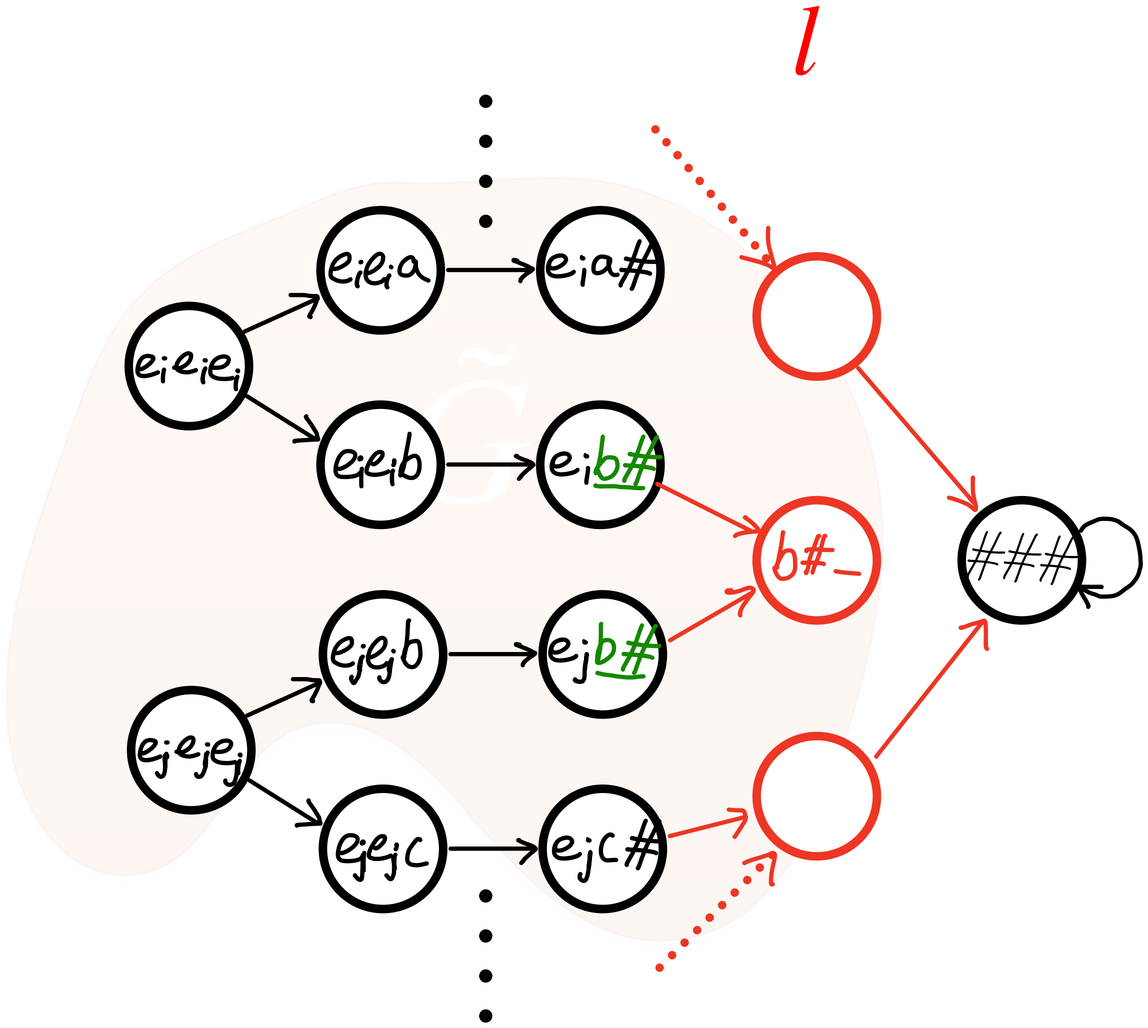
Hardness of Connect-DBG

$$\mathcal{I}_{C-dBG}$$



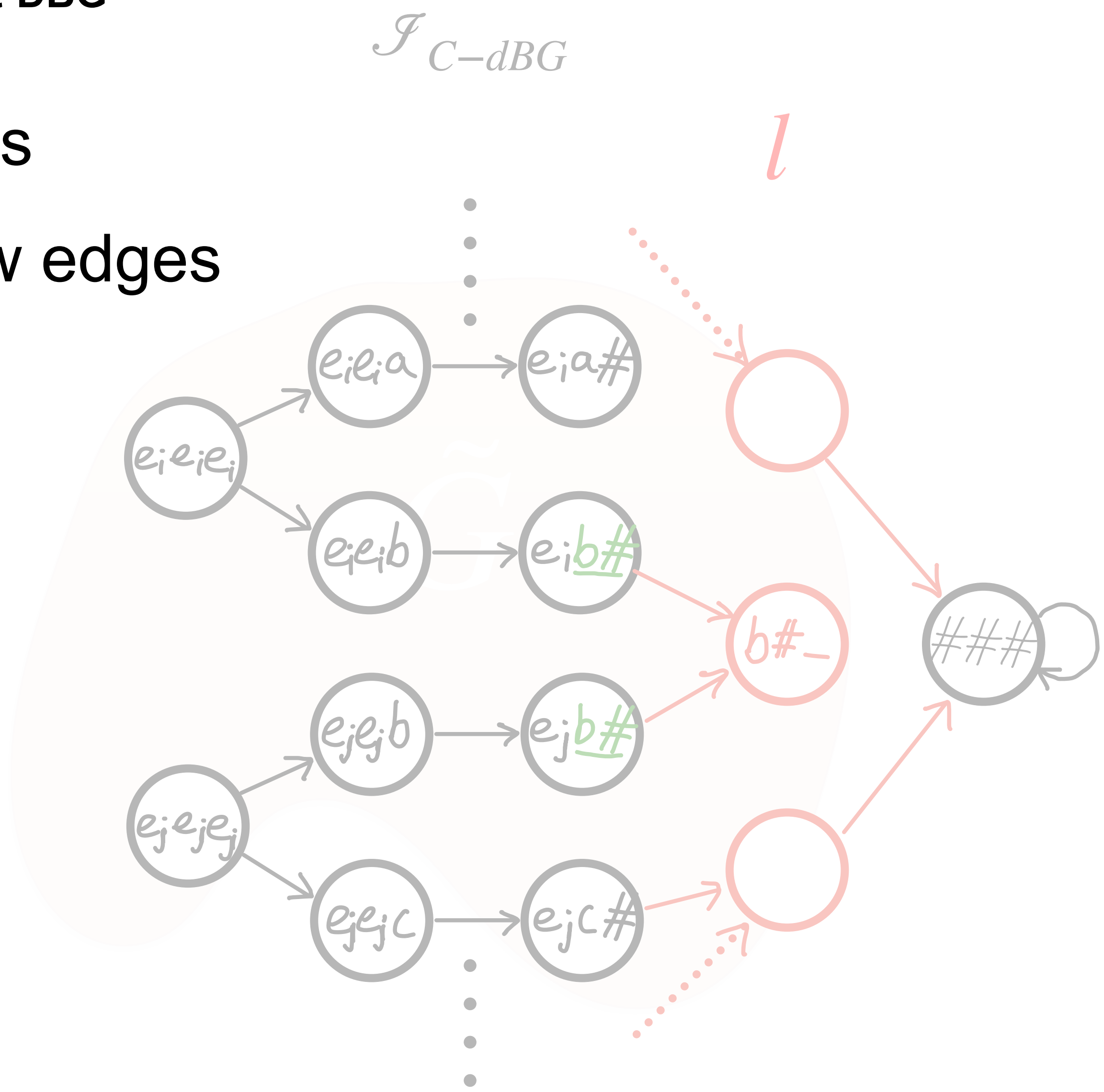
Hardness of Connect-DBG

$\mathcal{I}_{C-dBG}$

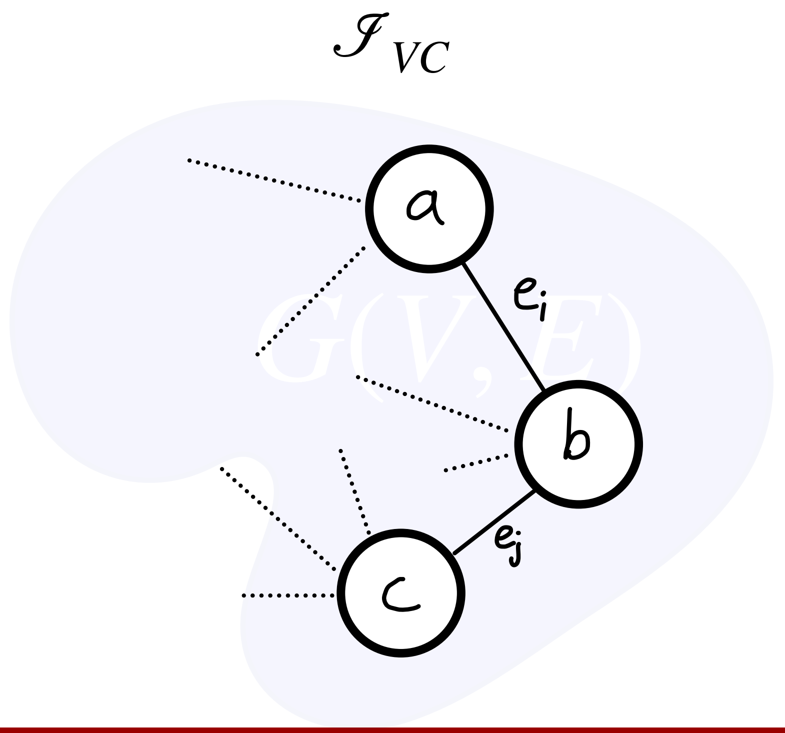


Hardness of Connect-DBG

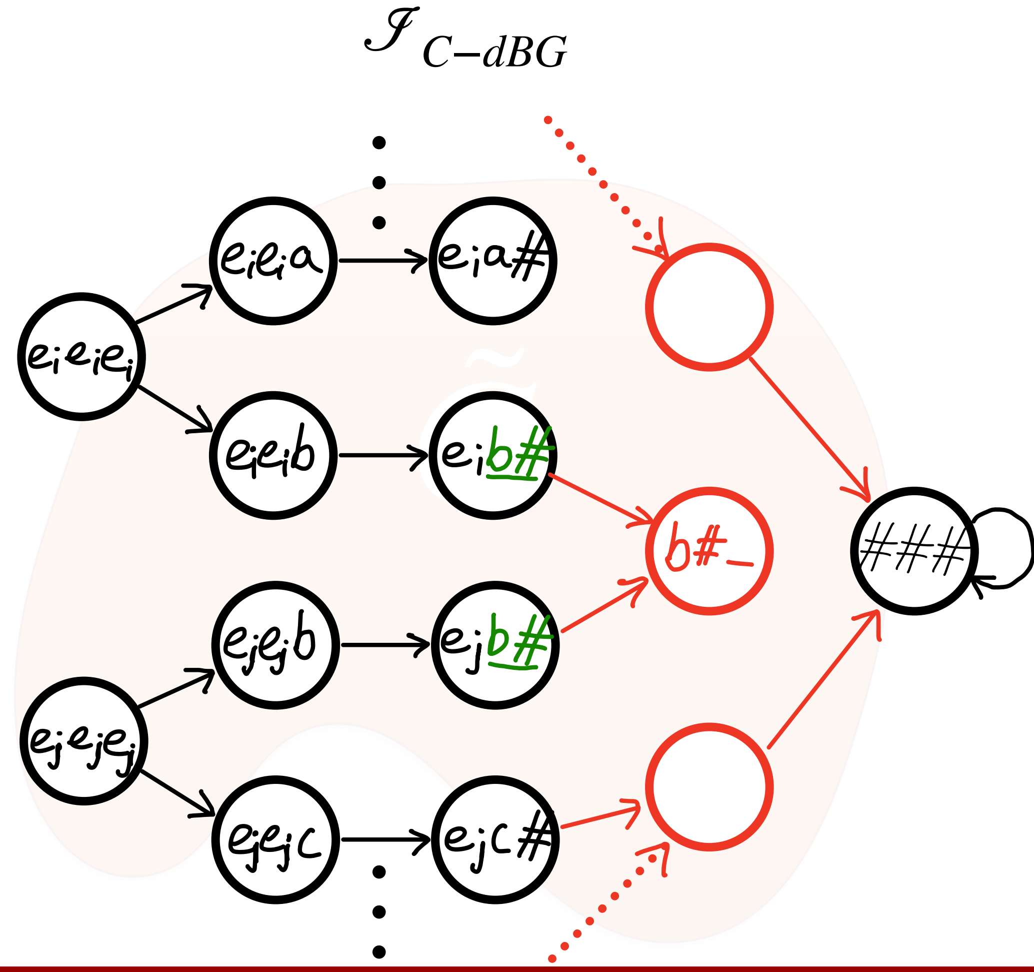
- l new vertices
- $|E| + l$ new edges



For every new vertex:



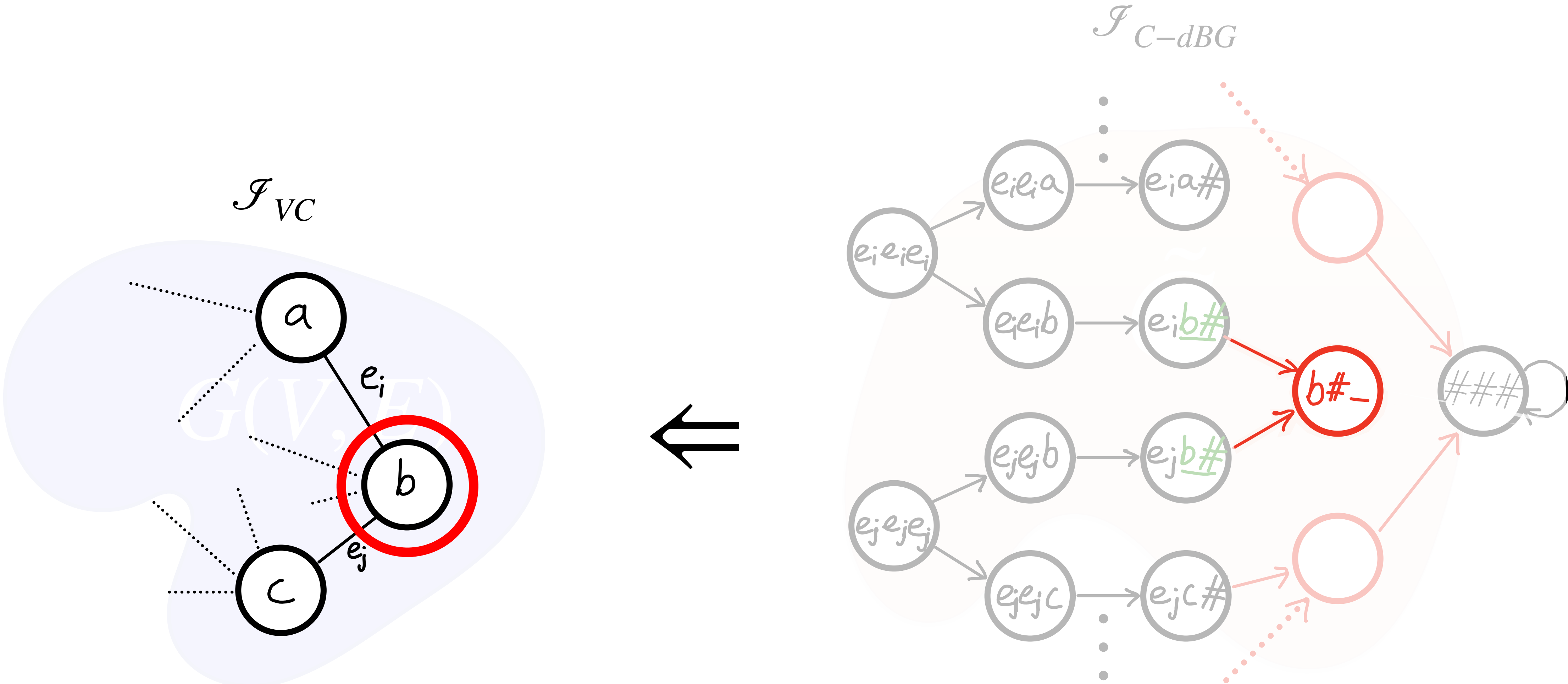
$\Leftarrow$



Hardness of Connect-DBG

For every new vertex:

- If 2 adjacent edge-gadgets → choose corresponding vertex

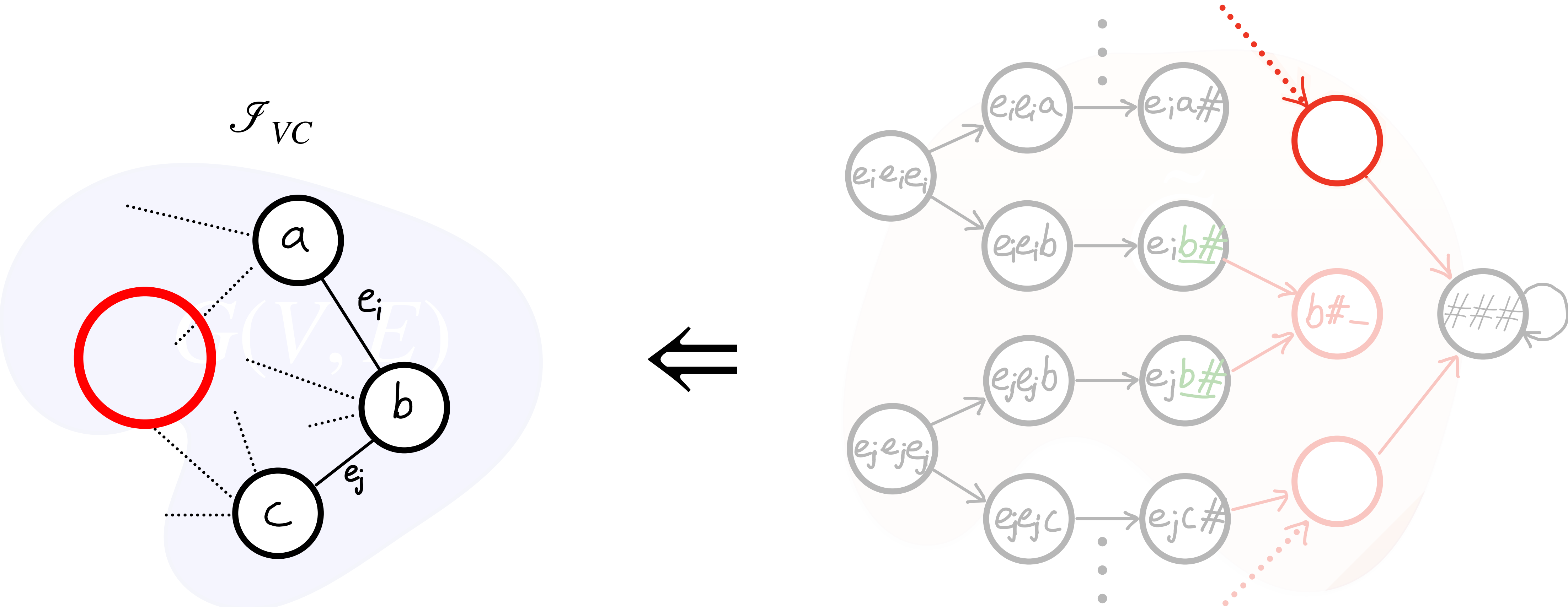


Hardness of Connect-DBG

For every new vertex:

- If 2 adjacent edge-gadgets → choose corresponding vertex
- Otherwise choose one endpoint of corresponding edge

$\mathcal{I}_{VC}$ $\Leftarrow$ $\mathcal{I}_{C-dBG}$

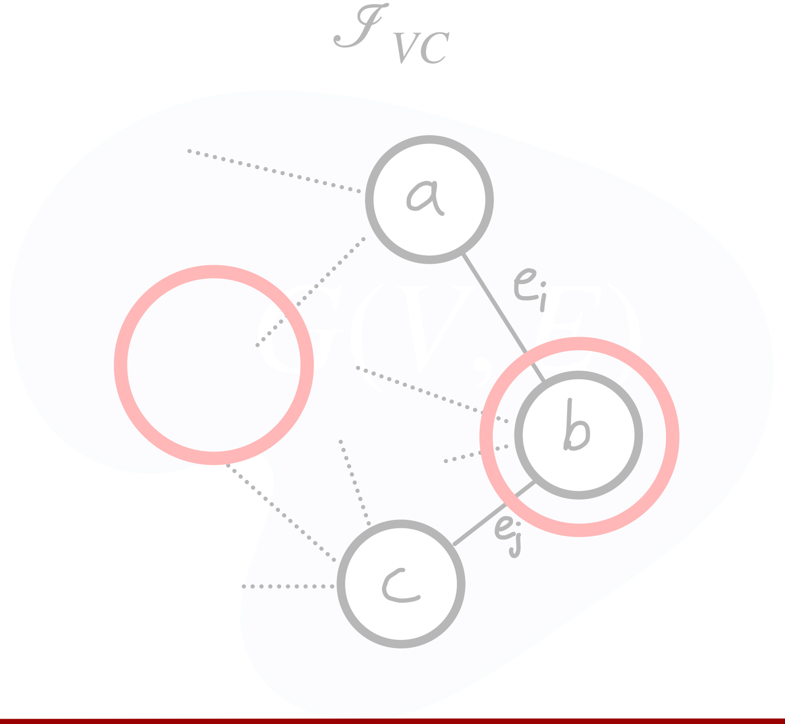


Hardness of Connect-DBG

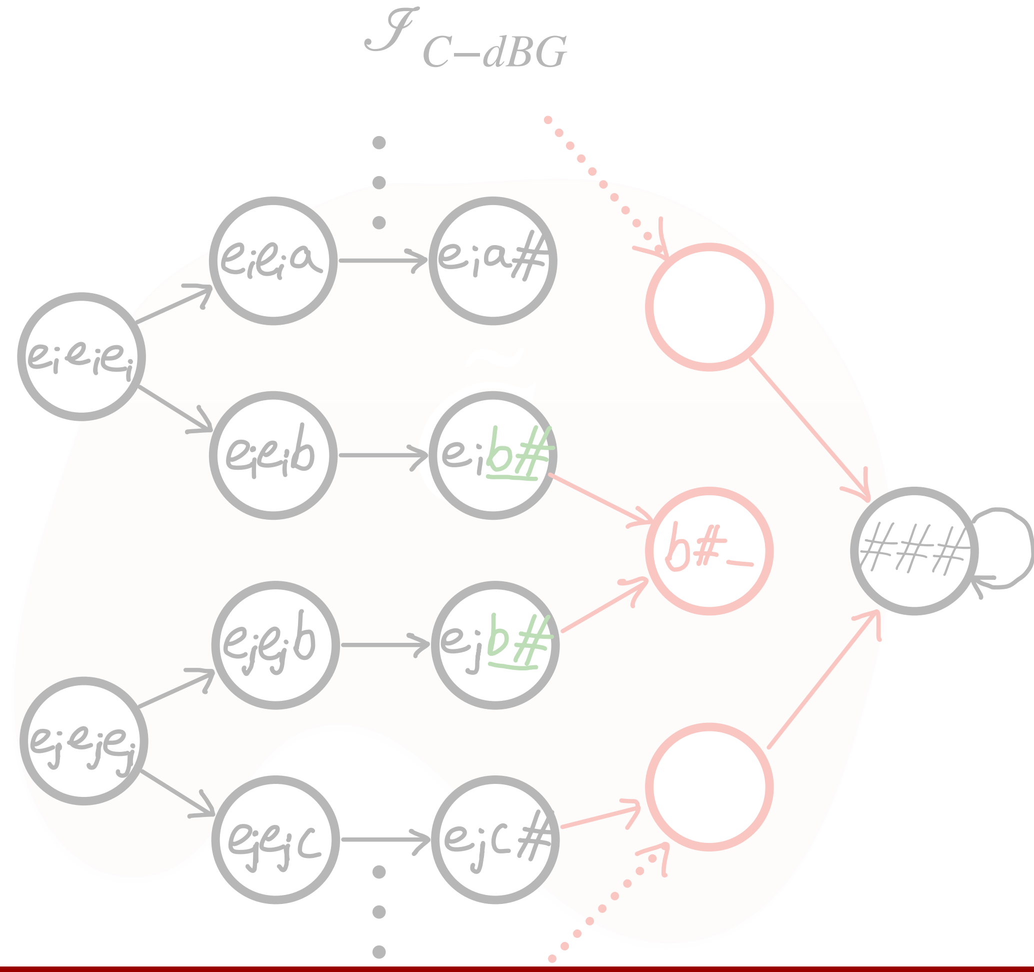
For every new vertex:

- If 2 adjacent edge-gadgets → choose corresponding vertex
- Otherwise choose one endpoint of corresponding edge

Solution to $\mathscr{I}_{VC}$ of size l



$\Leftarrow$

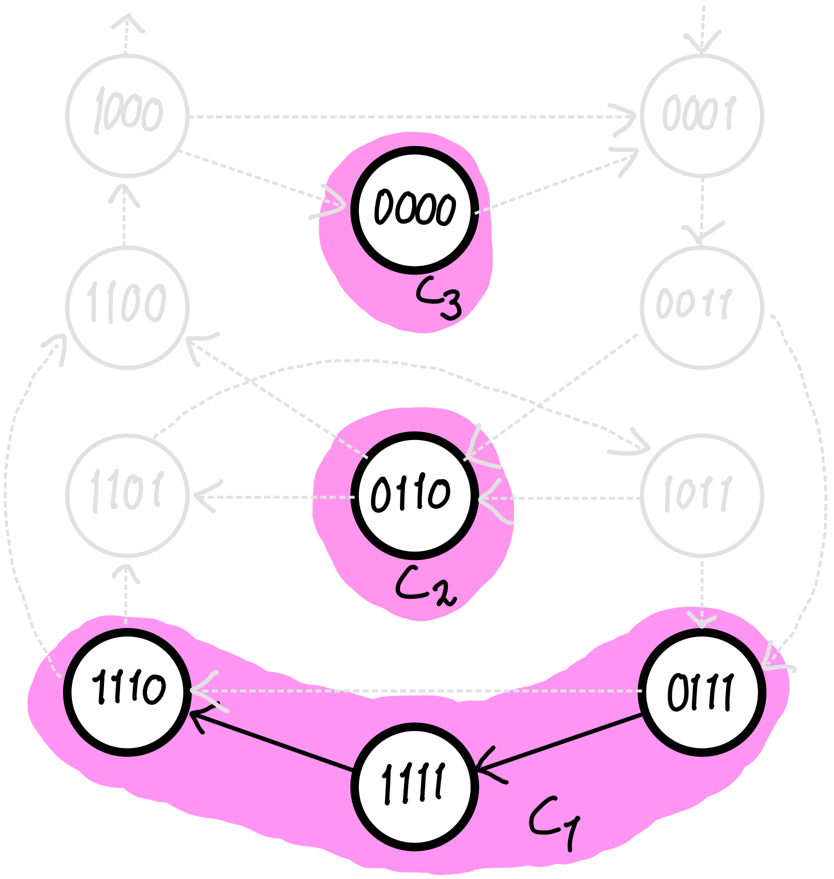


Hardness of Connect-DBG

$$OPT(\mathcal{I}_{VC}) = l \quad \Leftrightarrow \quad OPT(\mathcal{I}_{C-dBG}) = |E| + l$$

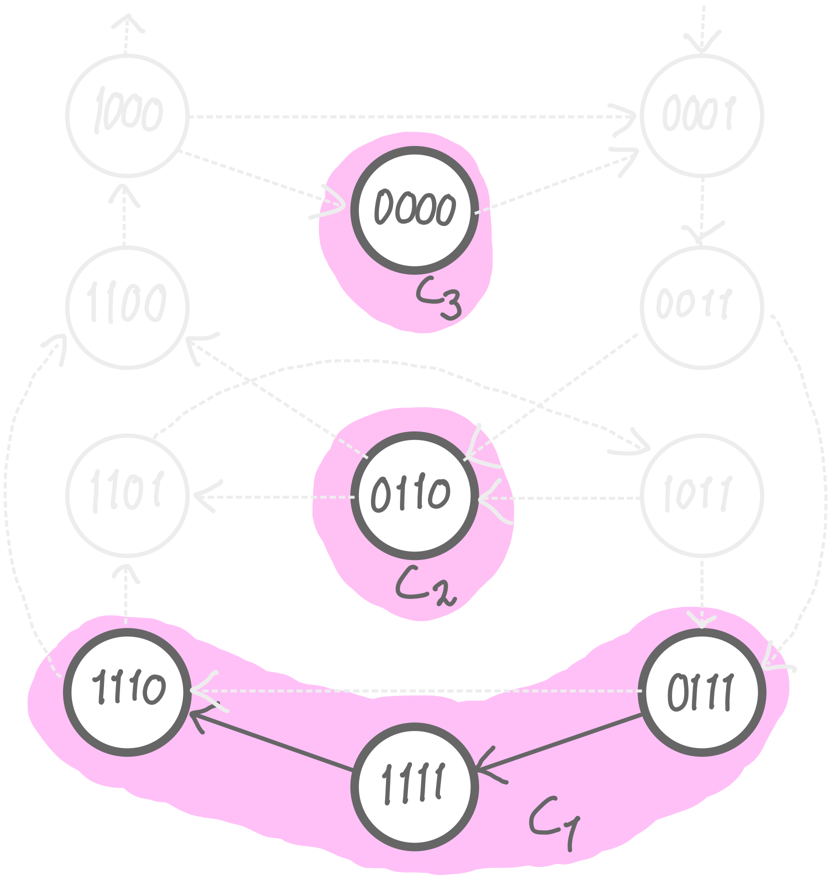
Approximation of CONNECT-DBG

Approximation of Connect-DBG



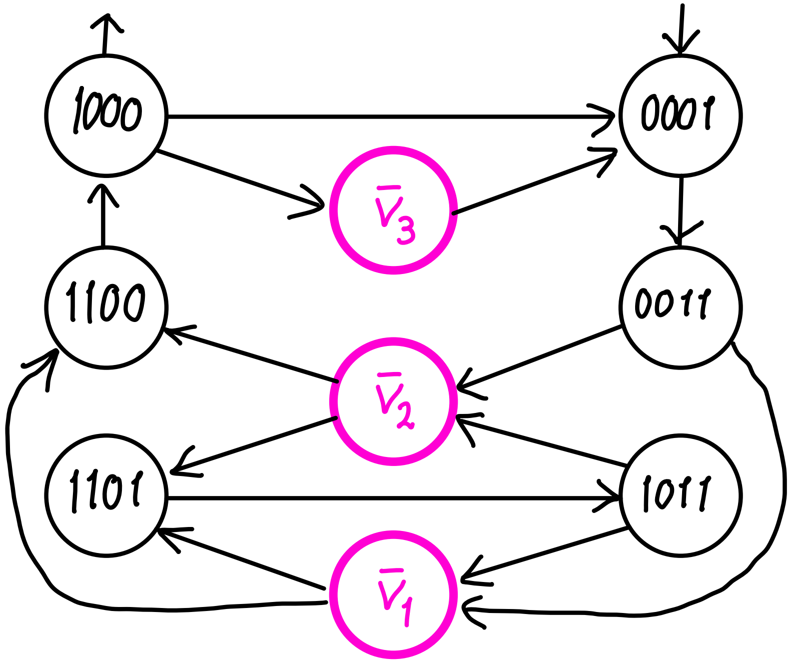
Approximation of Connect-DBG

- Collapse each connected component into a supernode in the complete dBG



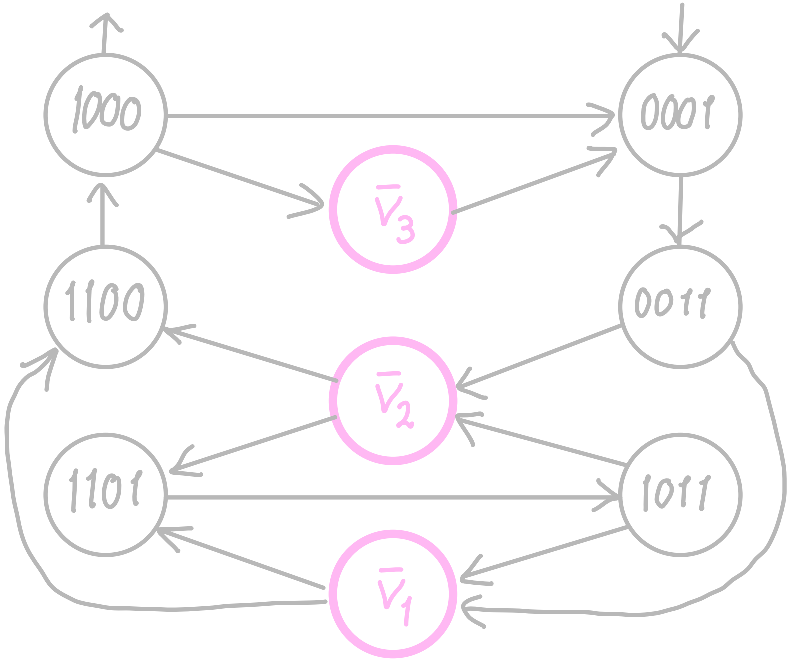
Approximation of Connect-DBG

- Collapse each connected component into a supernode in the complete dBG



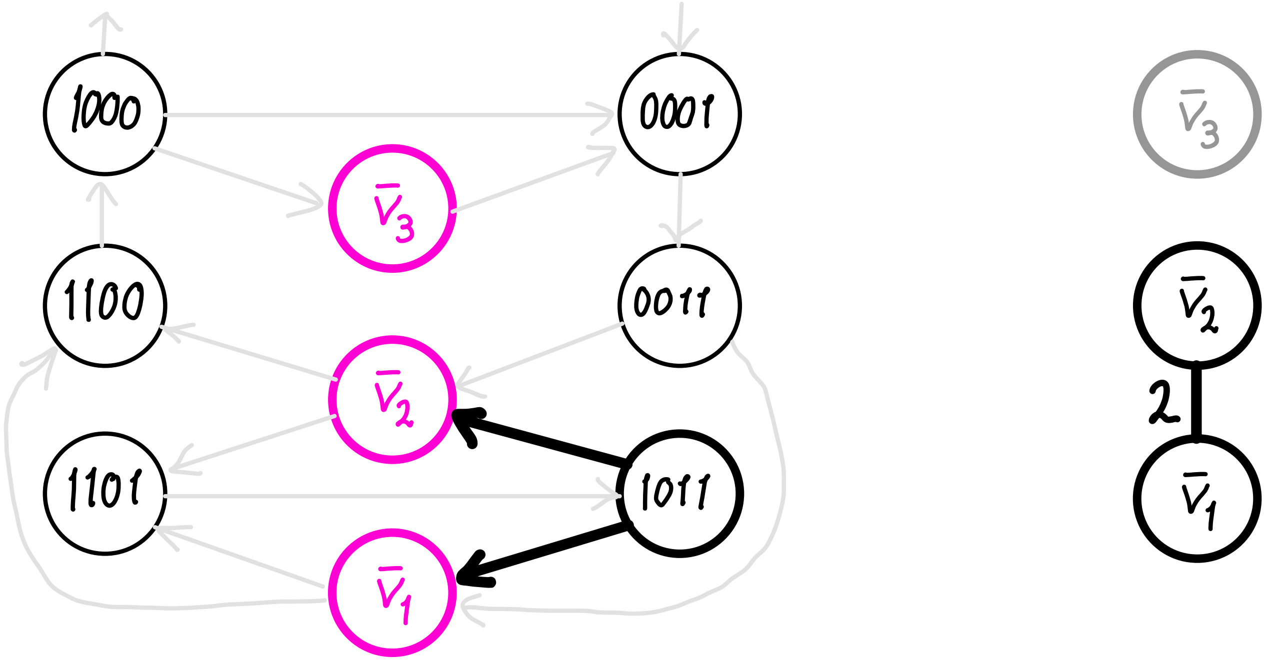
Approximation of Connect-DBG

- Construct the metric closure



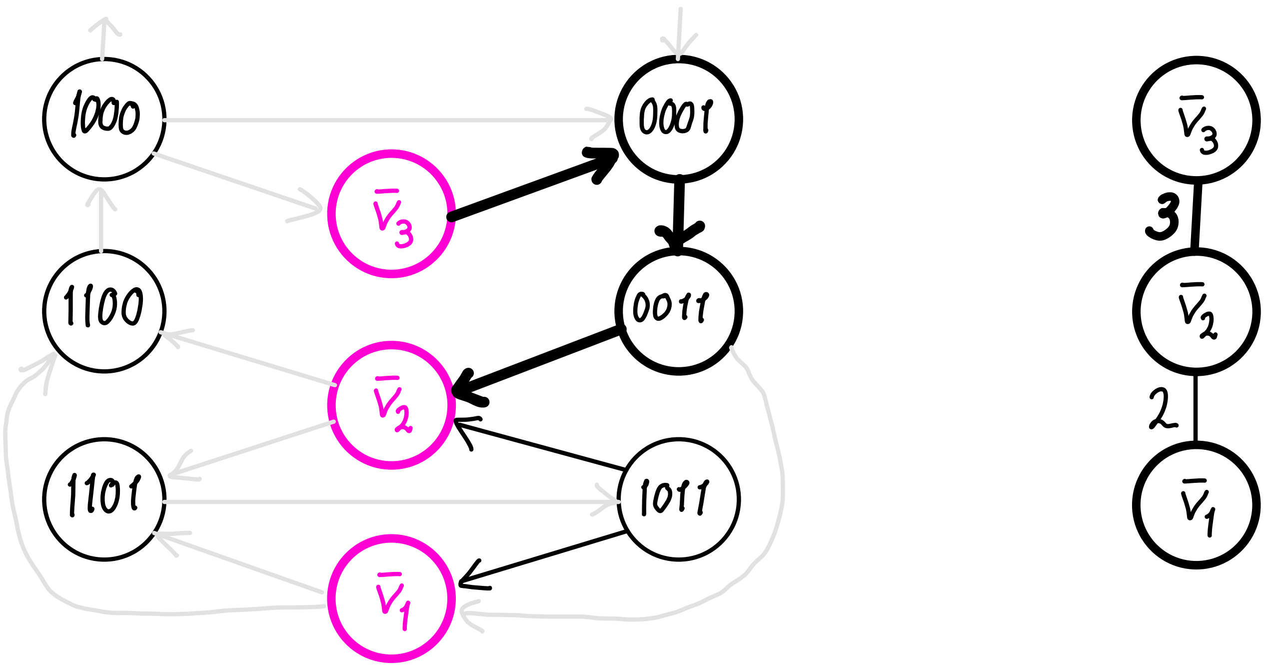
Approximation of Connect-DBG

- Construct the metric closure



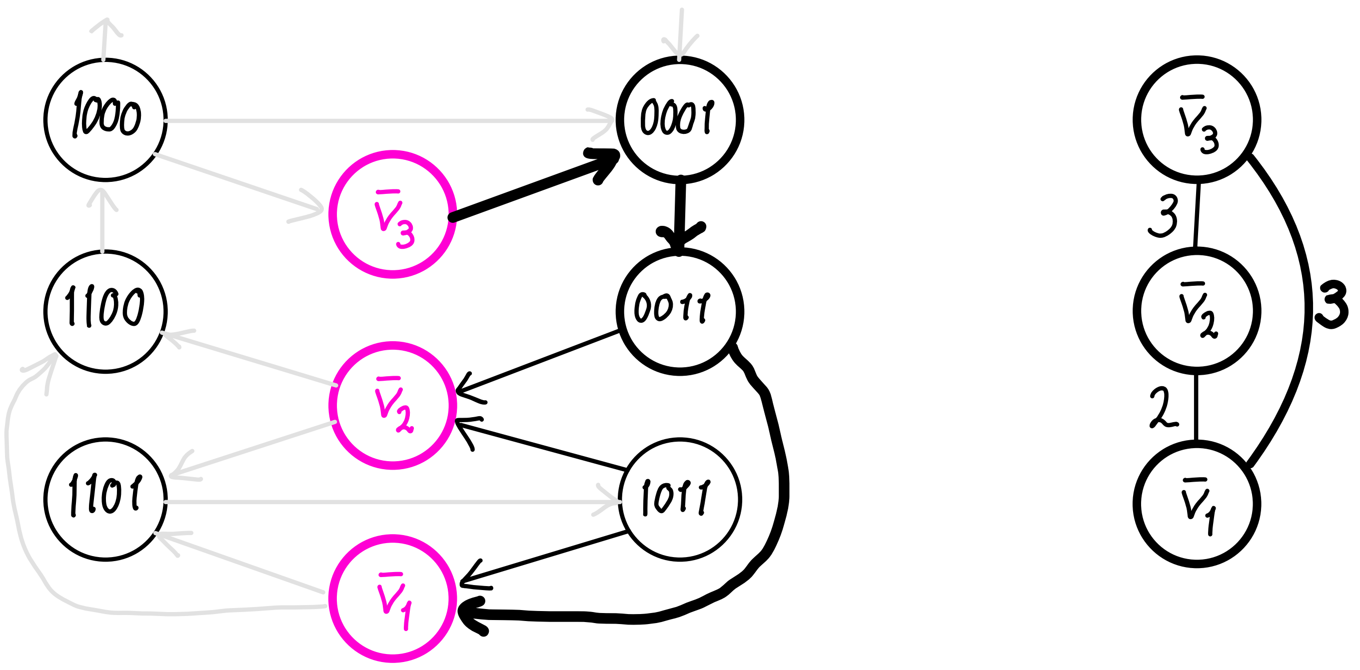
Approximation of Connect-DBG

- Construct the metric closure



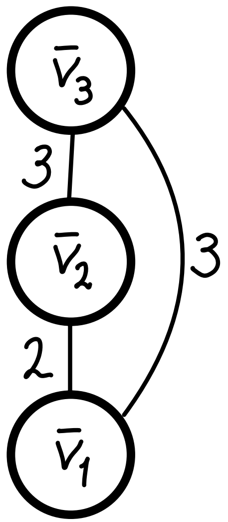
Approximation of Connect-DBG

- Construct the metric closure



Approximation of Connect-DBG

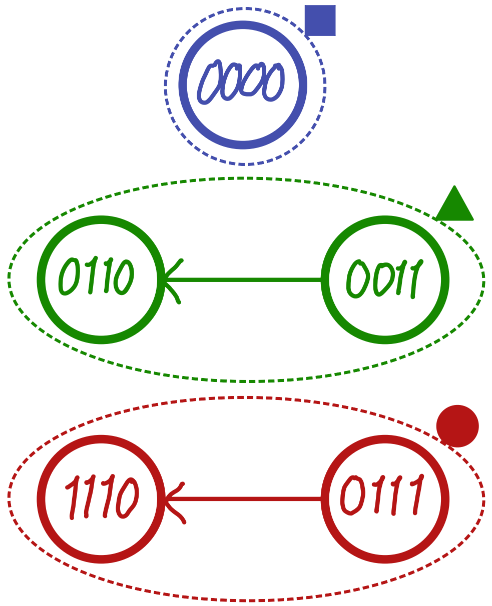
- Use 2-approximation for metric closure of Steiner Tree Problem by Kou et al. (1981)[1]



[1] Lawrence T. Kou, George Markowsky, and Leonard Herman. A fast algorithm for Steiner trees. *Acta Informatica,* 15:141-145, 1981. doi:10.1007/BF00288961

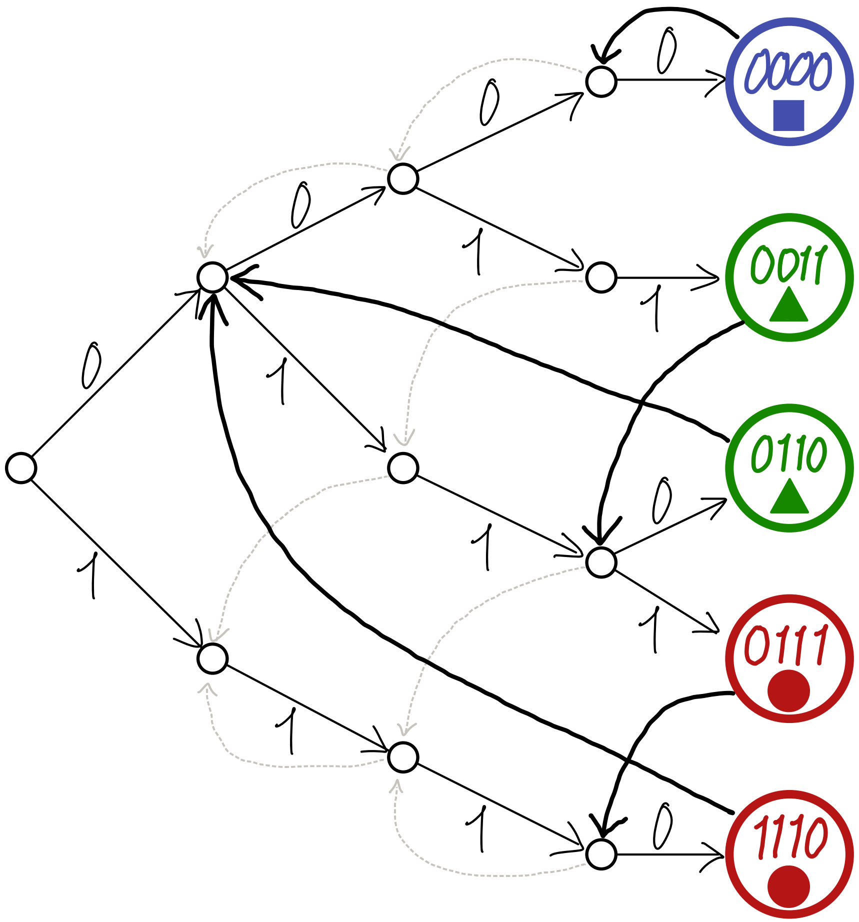
Improvement of CONNECT-DBG-P

Improvement of Connect-DBG-P



- Aho-Corasick (AC) Machine (KMP generalization)

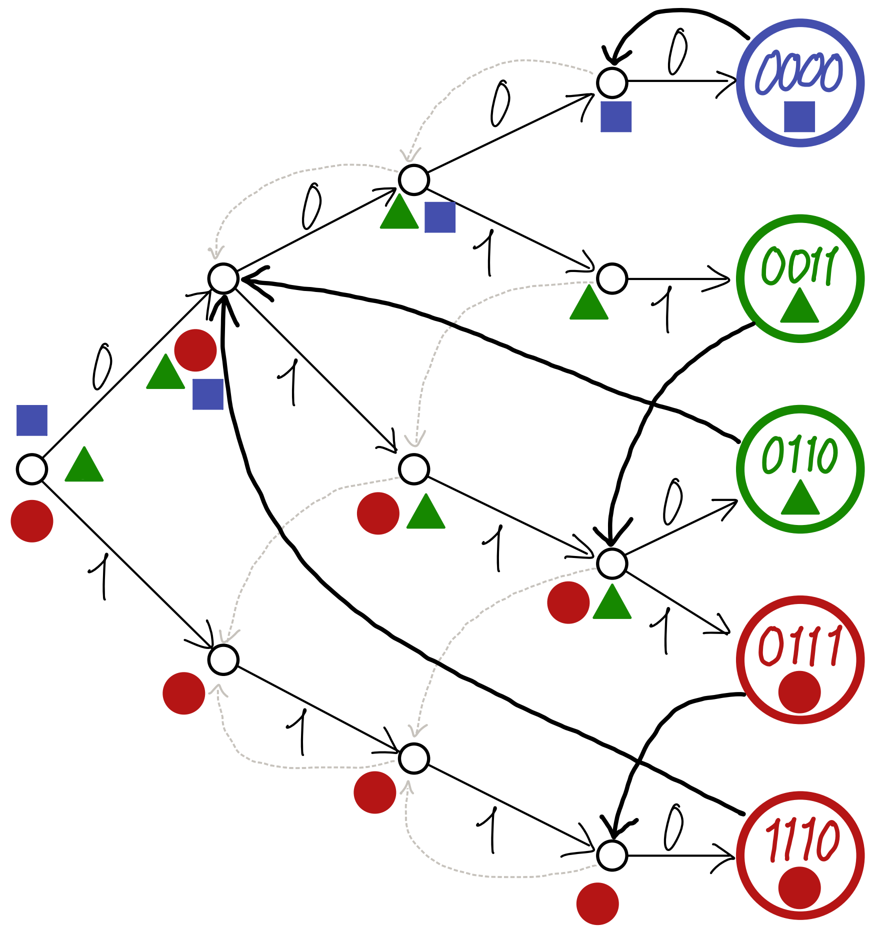
- Aho-Corasick (AC) Machine (KMP generalization)



Improvement of Connect-DBG-P

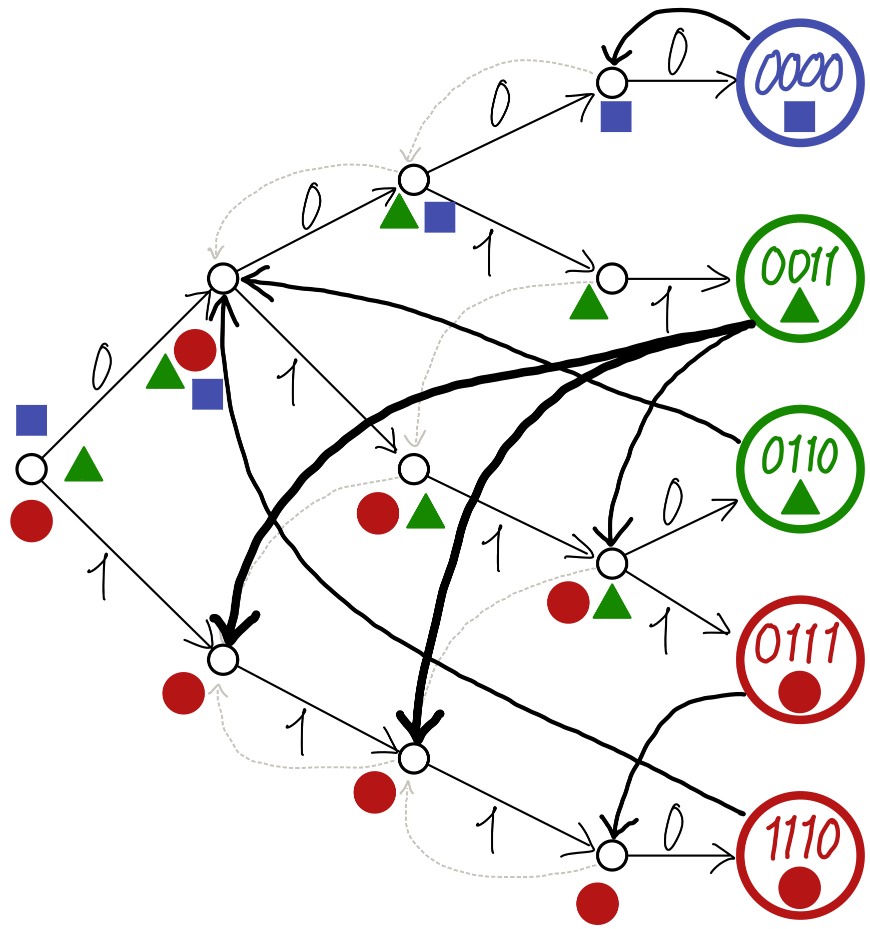
- Aho-Corasick (AC) Machine (KMP generalization)

- Add colors



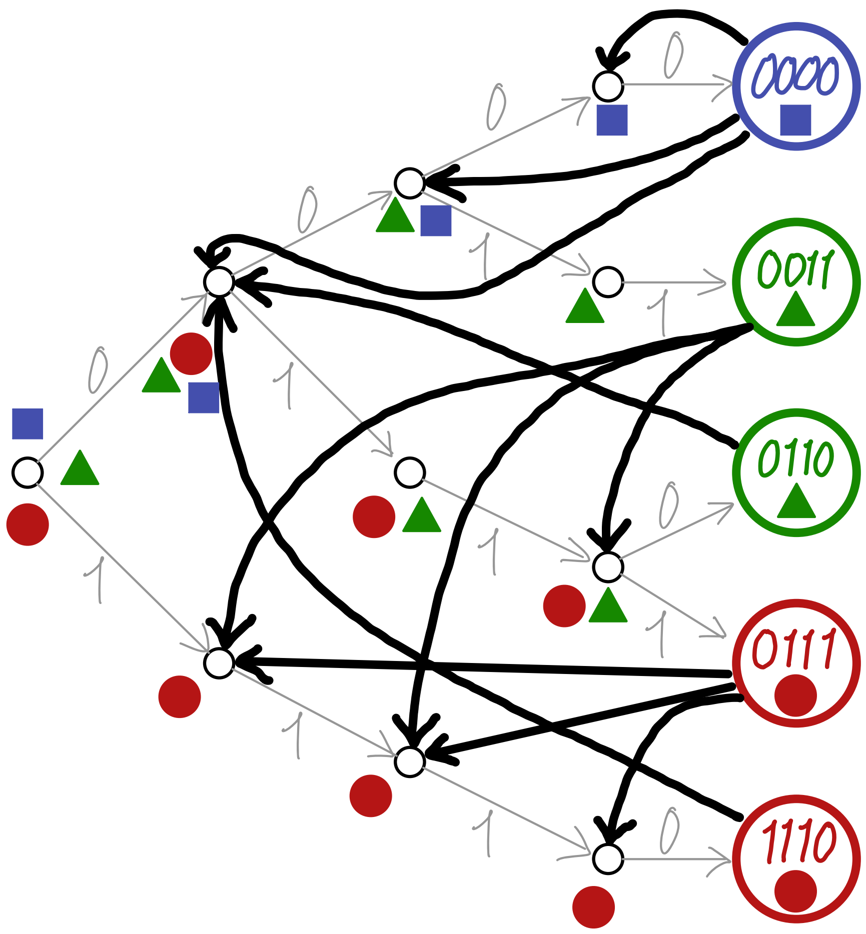
Improvement of Connect-DBG-P

- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- **Add backward edges**



Improvement of Connect-DBG-P

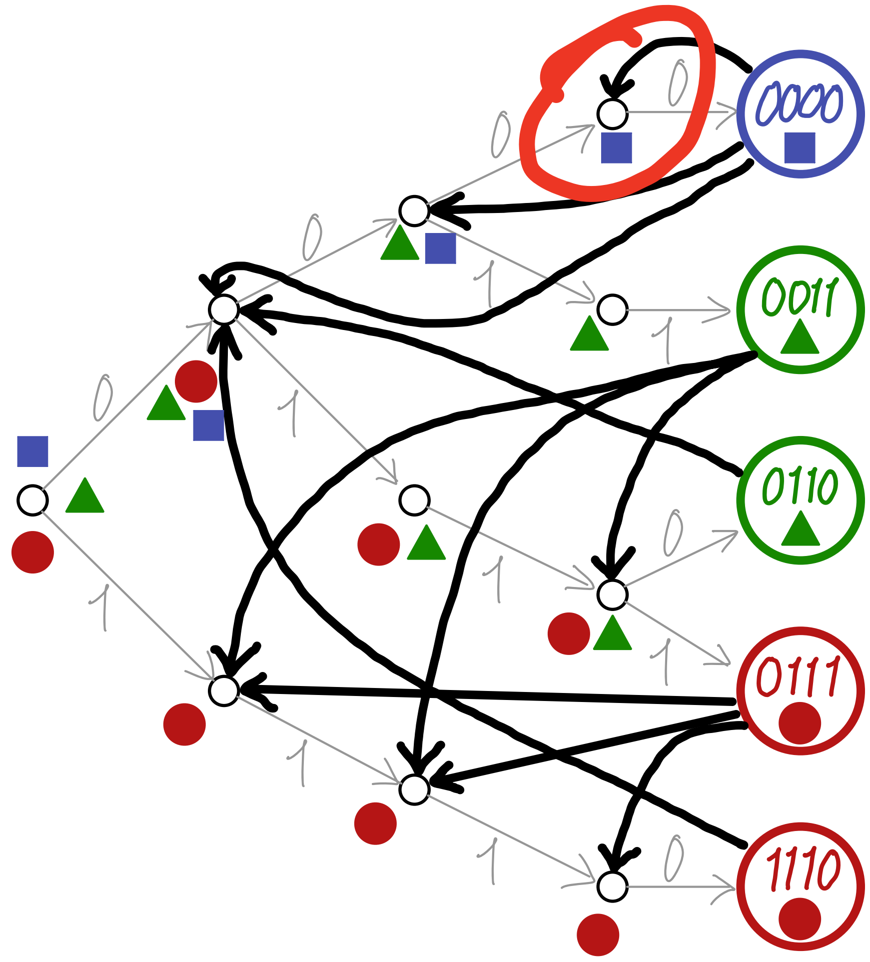
- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- **Add backward edges**



Improvement of Connect-DBG-P

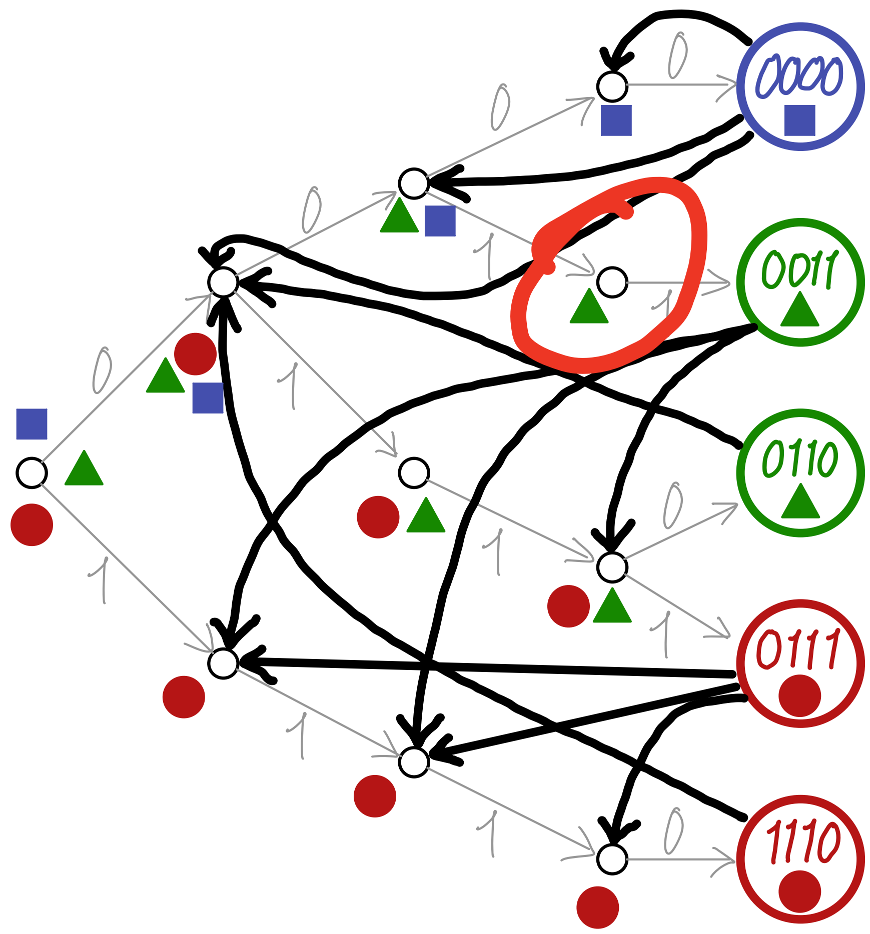
- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges

- **Reverse BFS from root:**
  - Compare backward edges to colors



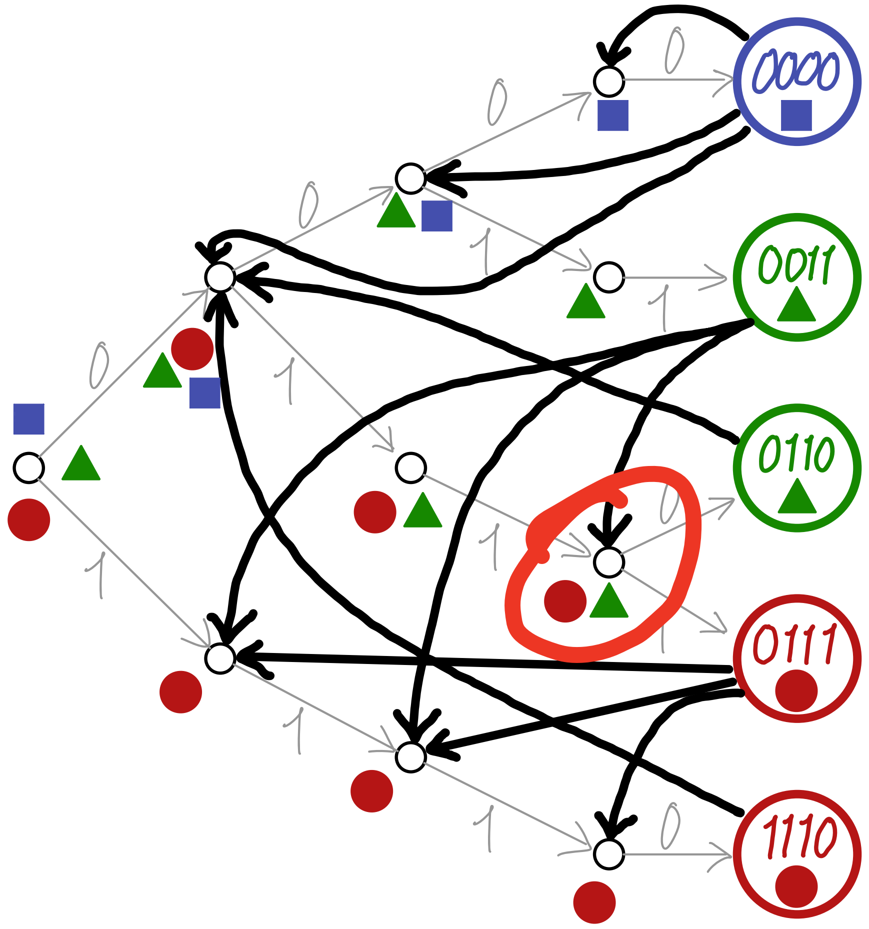
Improvement of Connect-DBG-P

- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges
- Reverse BFS from root:
  - Compare backward edges to colors



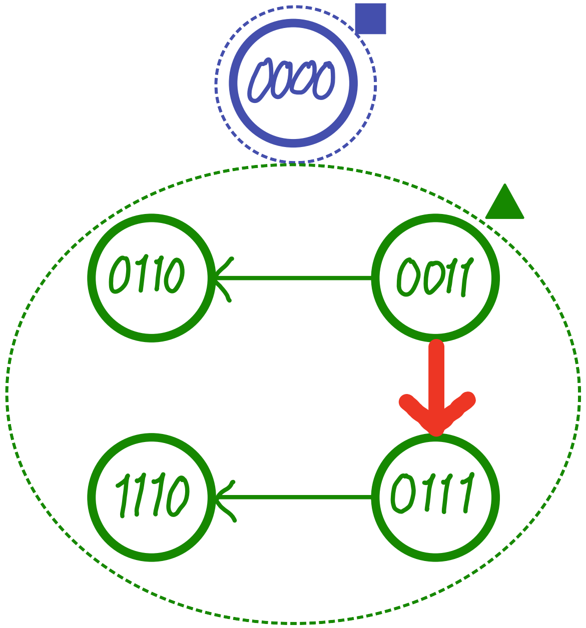
Improvement of Connect-DBG-P

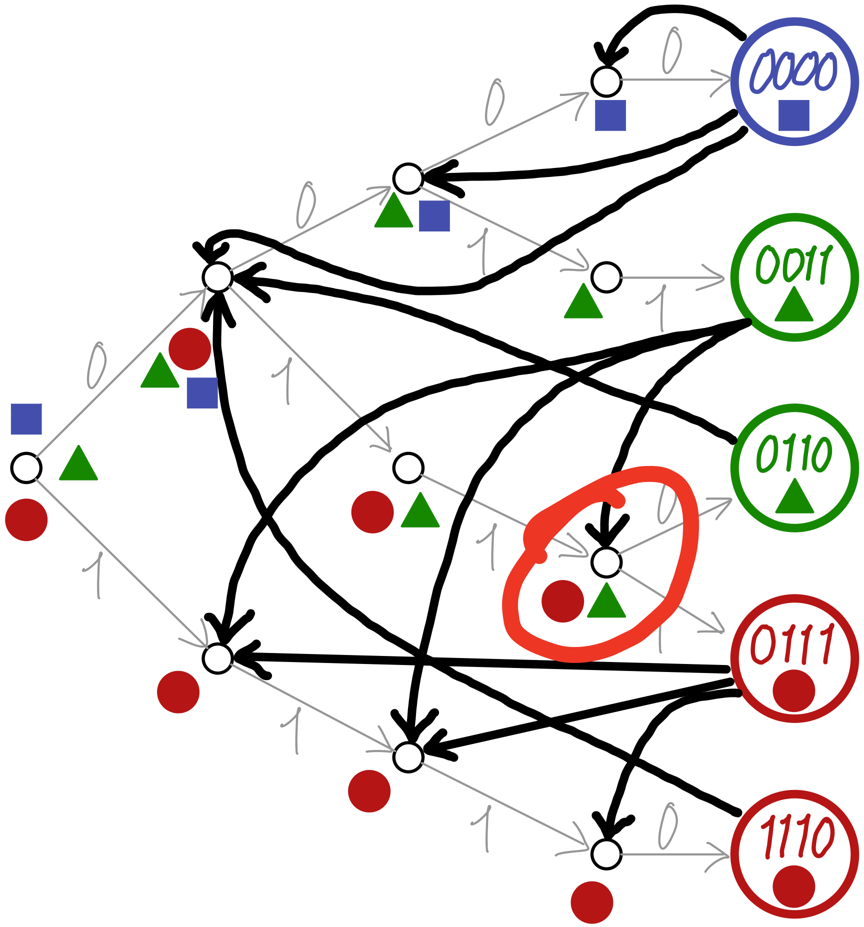
- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges
- Reverse BFS from root:
    - Compare backward edges to colors
    - Connect with paths



Improvement of Connect-DBG-P

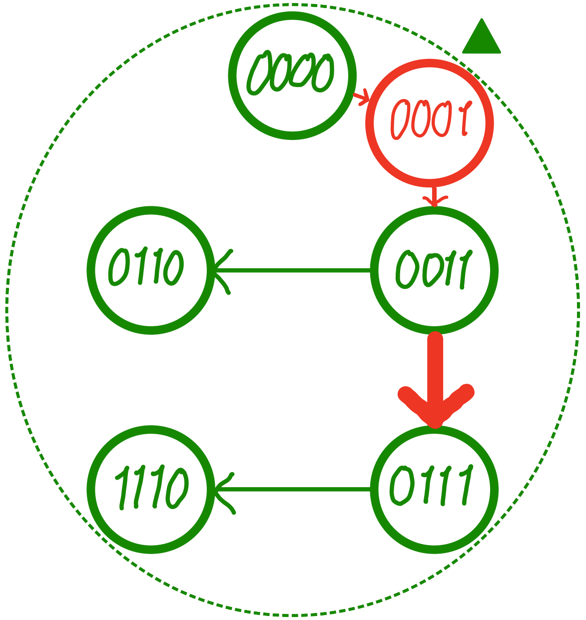
- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges
- Reverse BFS from root:
  - Compare backward edges to colors
  - Connect with paths

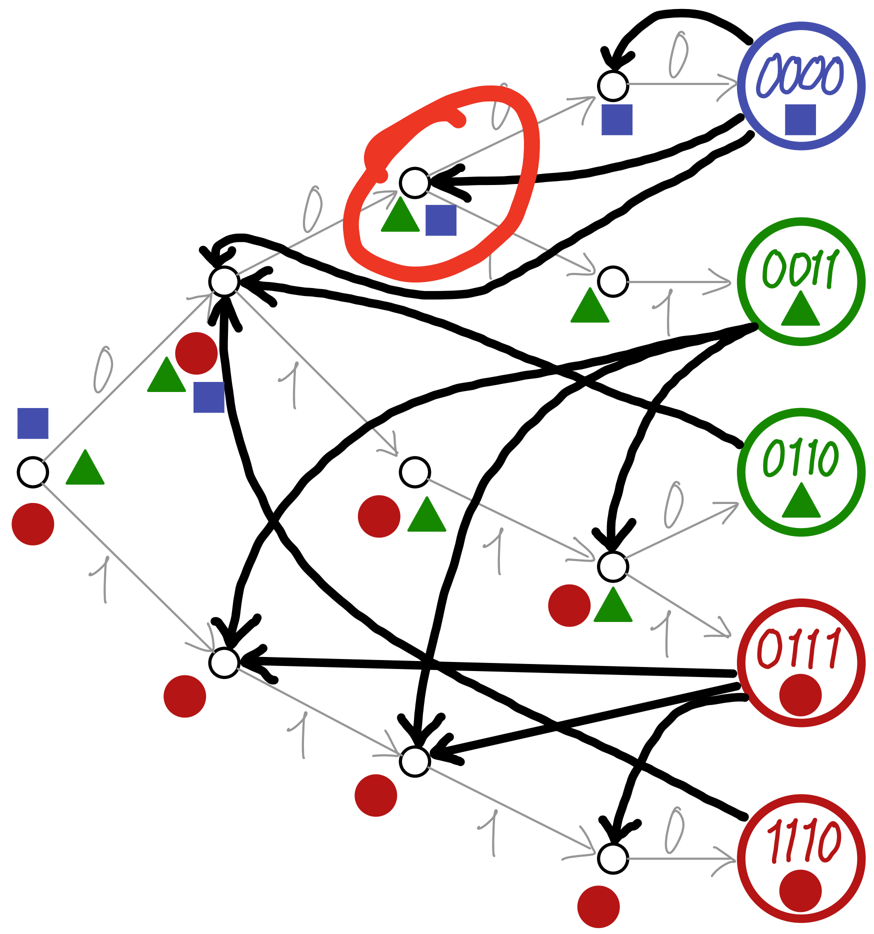




Improvement of Connect-DBG-P

- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges
- Reverse BFS from root:
  - Compare backward edges to colors
  - Connect with paths





Improvement of Connect-DBG-P

- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges
- Reverse BFS from root:
  - Compare backward edges to colors
  - Connect with paths

$$\mathcal{O}(k|V|\alpha(|V|) + |E|) \ \text{time}$$

Improvement of Connect-DBG-P

- Aho-Corasick (AC) Machine (KMP generalization)
- Add colors
- Add backward edges
- Reverse BFS from root:
  - Compare backward edges to colors
  - Connect with paths

$$\mathcal{O}(k\,|V|\,\alpha(|V|) + |E|) \ \text{time}$$

$\alpha(\cdot)$ is the inverse Ackermann function

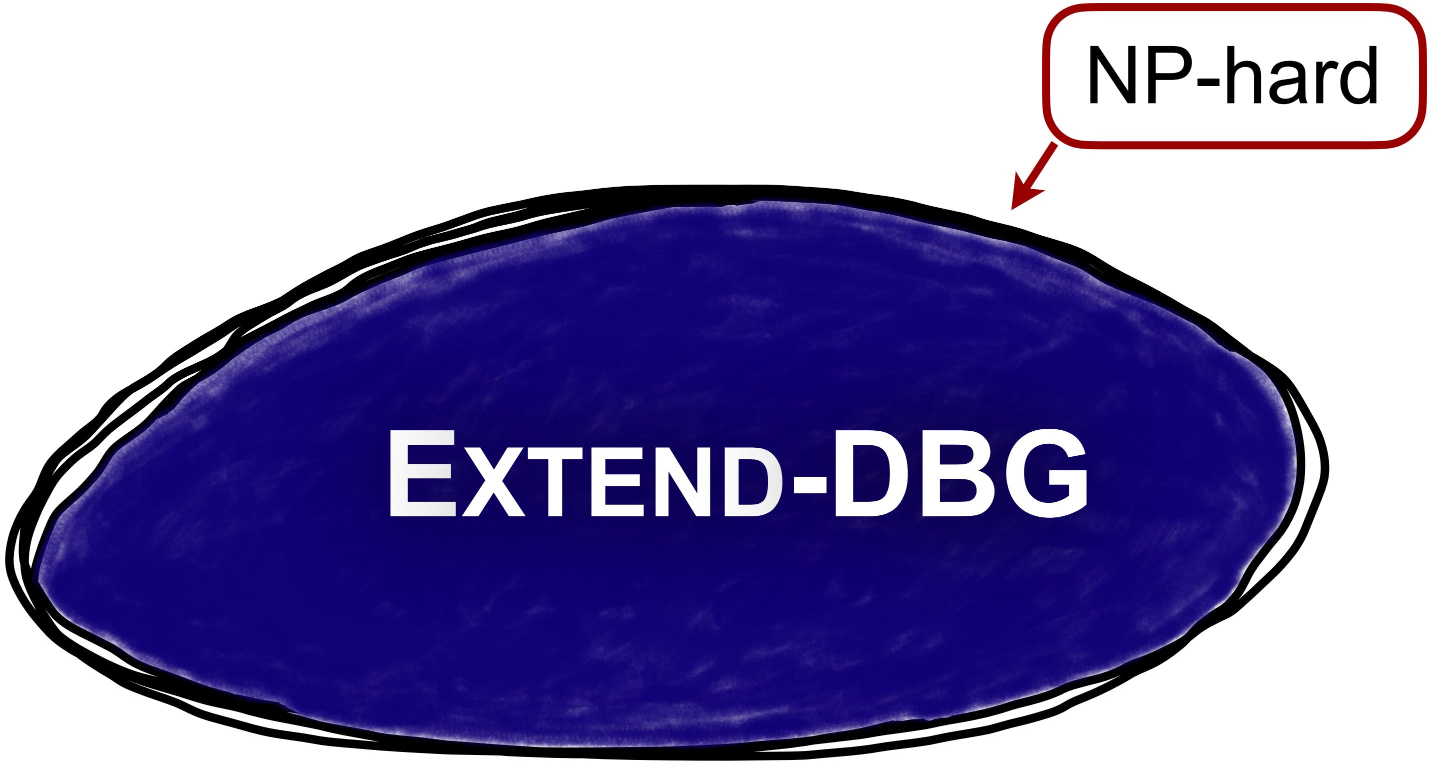
$\alpha(n)$ grows slower than $\log^* n$

Summary

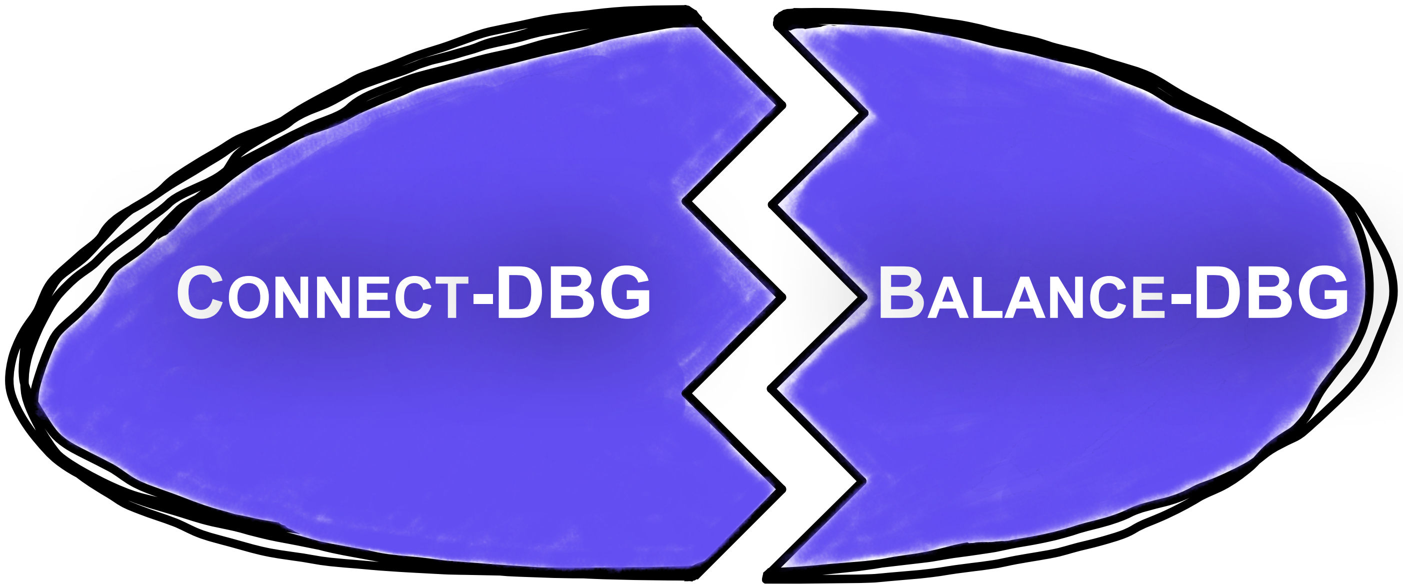
Summary

EXTEND-DBG

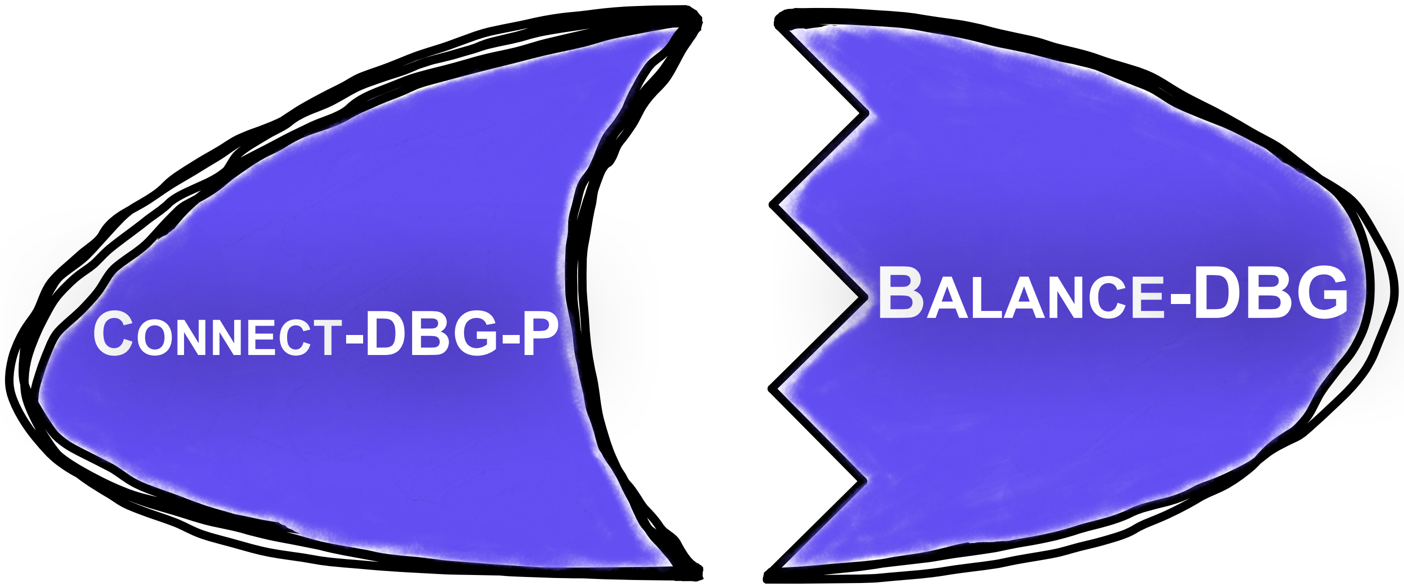
Summary



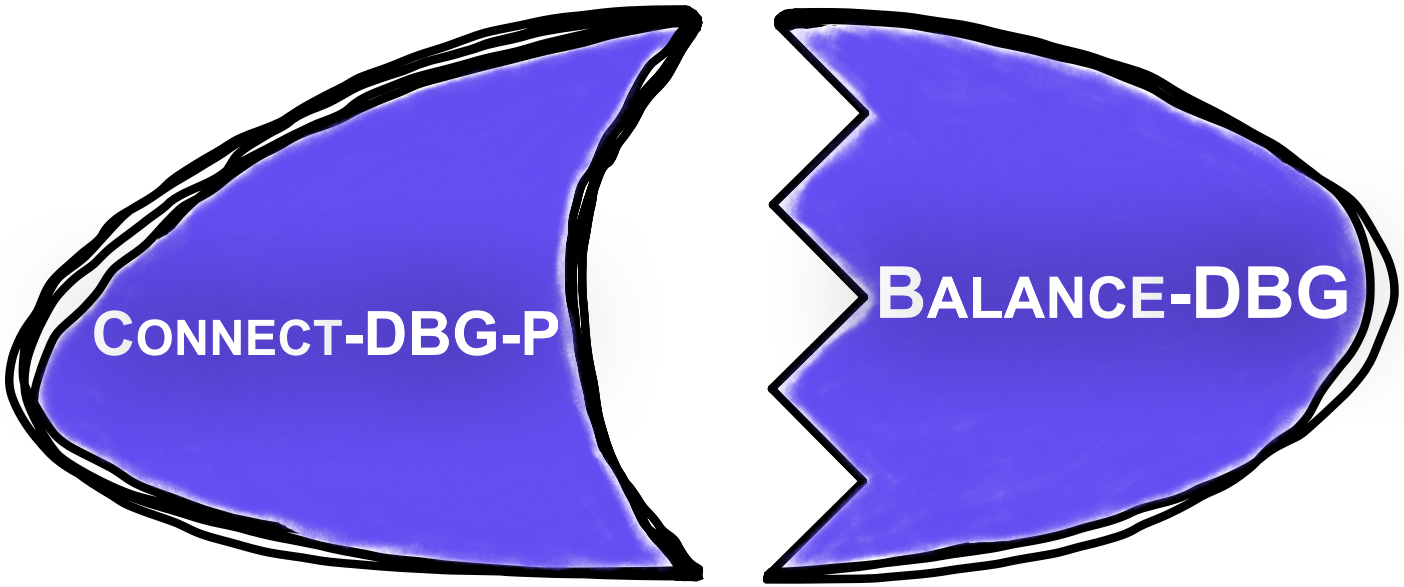
Summary



Summary



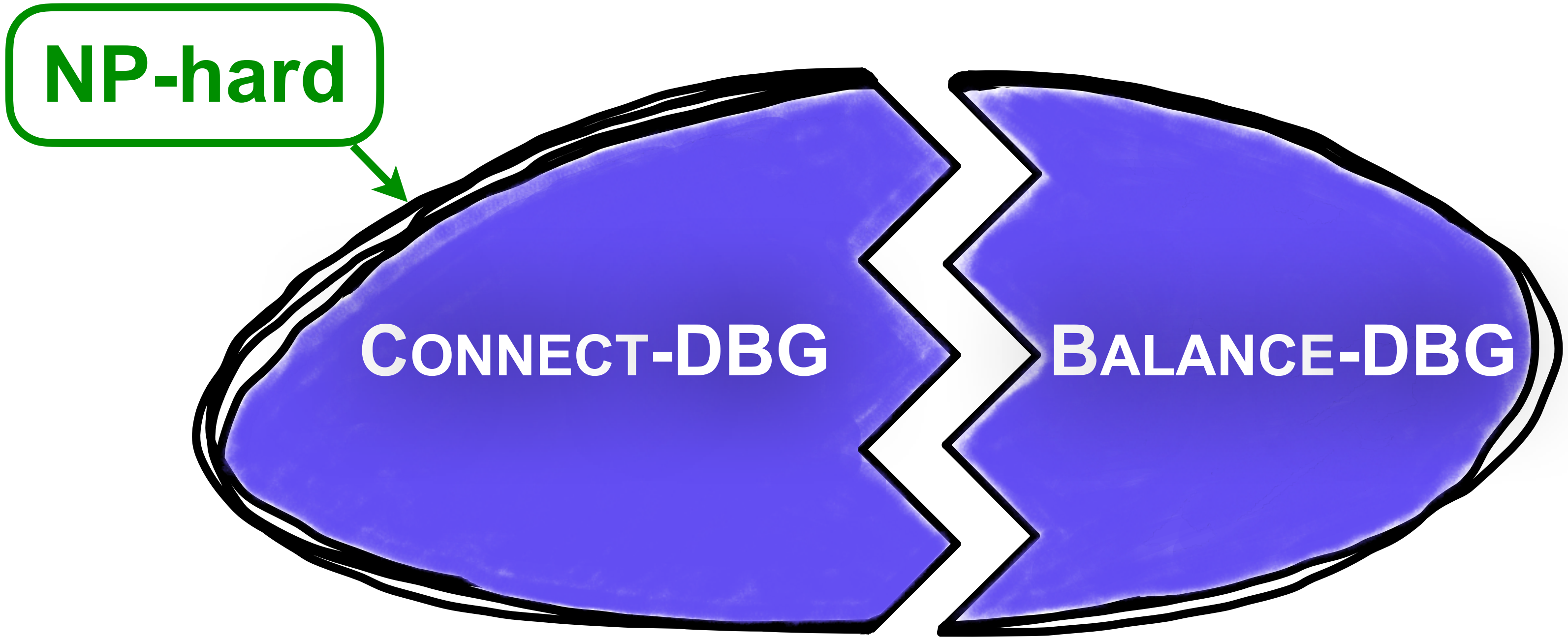
Summary



$\mathcal{O}(k|V|\alpha(|V|) + |E|)$ time

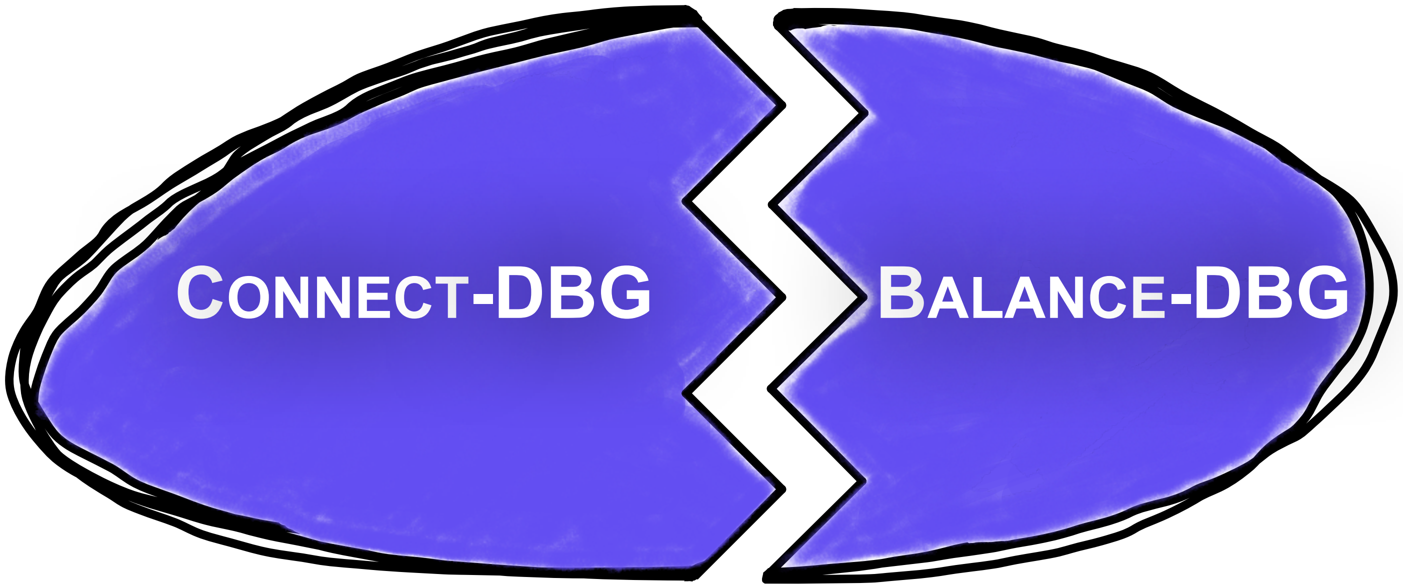
$\mathcal{O}(k|V| + |E| + |A|)$ time

Summary



$\mathcal{O}(k|V| + |E| + |A|)$ time

Summary



2-approximation $\mathcal{O}(k|V| + |E| + |A|)$ time

Summary

Extend-DBG

Approximation?