Updated 3/30/2023

Topic 3: Efficient query evaluation
Unit 2: Cyclic query evaluation
Lecture 21

Wolfgang Gatterbauer

CS7240 Principles of scalable data management (sp23)

https://northeastern-datalab.github.io/cs7240/sp23/

3/28/2023

Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

cycles make everything more complicated ☹

Why cyclic queries (other than social networks)

```
Likes(person, drink)
Frequents(person, bar)
Serves(bar, drink, cost)
```

2. Specify or choose a Query Supported grammar

104 Bars: Persons who frequent some bar that serves some drink they like.

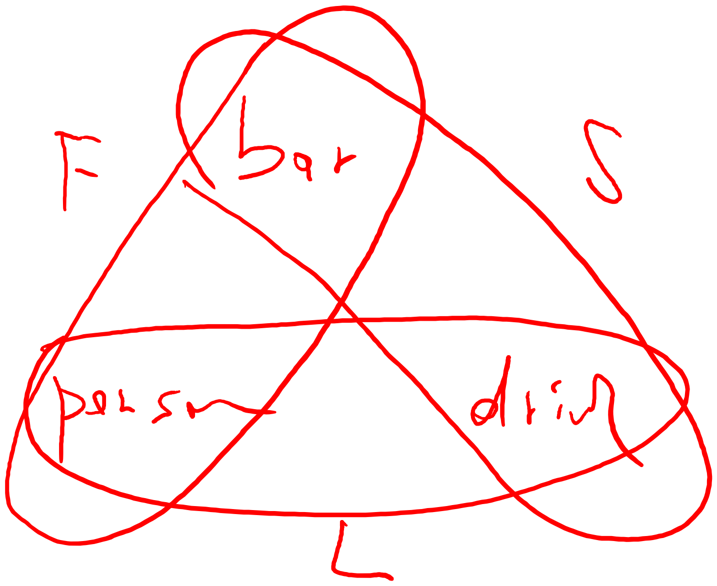
Source: http://demo.queryvis.com

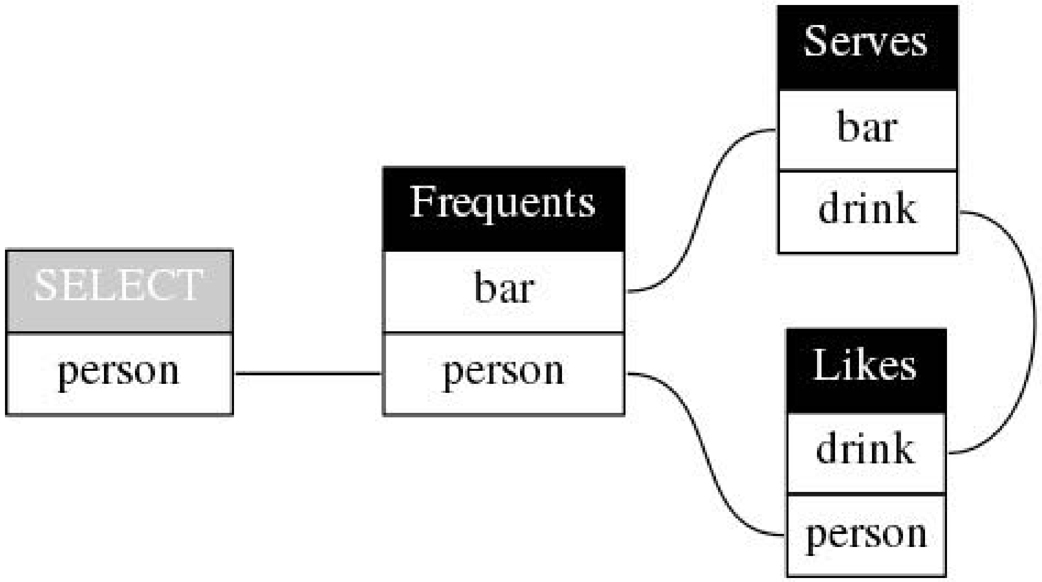
Why cyclic queries (other than social networks)

```
Likes(person, drink)
Frequents(person, bar)
Serves(bar, drink, cost)
```

2. Specify or choose a Query [Supported grammar]

104 Bars: Persons who frequent some bar that serves some drink they like.





Why cyclic queries (other than social networks)

```
Likes(person, drink)
Frequents(person, bar)
Serves(bar, drink, cost)
```

2. Specify or choose a Query — Supported grammar

104 Bars: Persons who frequent some bar that serves some drink they like.

```
SELECT   F1.person
FROM     Frequents F1
WHERE    exists
         (SELECT *
         FROM     Serves S2
         WHERE    S2.bar = F1.bar
         AND      exists
                  (SELECT *
                  FROM     Likes L3
                  WHERE    L3.person = F1.person
                  AND      S2.drink = L3.drink))
```

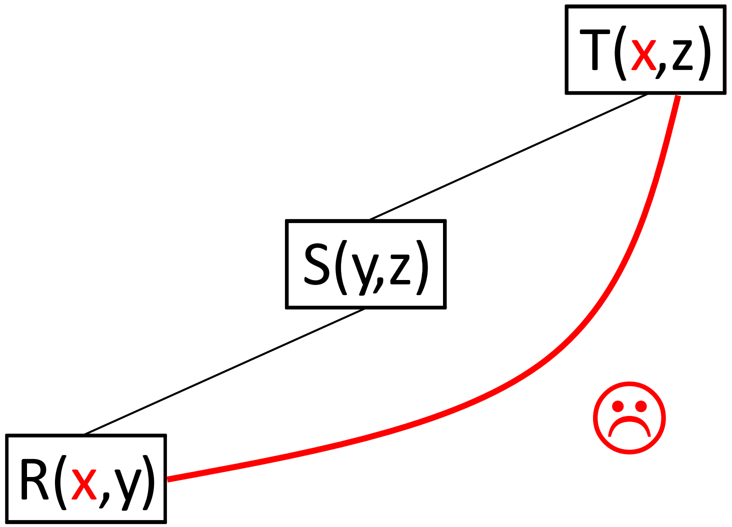
Source: http://demo.queryvis.com

Joins in databases: one-at-a-time

How can we efficiently process multi-way joins with cycles?

$Q(x,y,z)$:- $R(x,y), S(y,z), T(x,z)$.

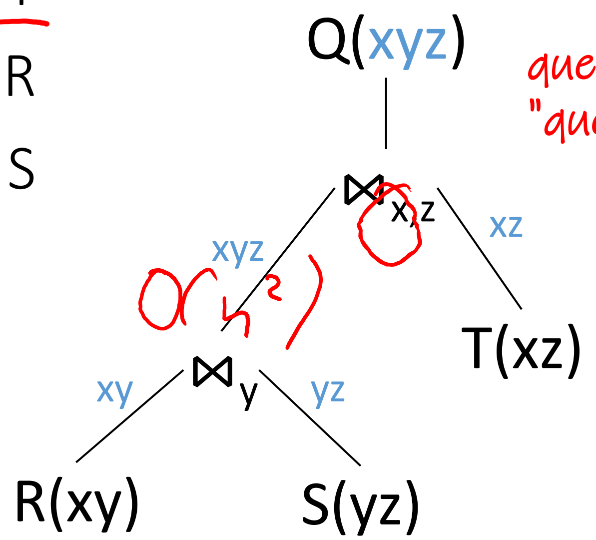
Recall:



There is no join tree! You can't fulfill the running intersection property...

Three possible plans

- $(R \bowtie S) \bowtie T$
- $(S \bowtie T) \bowtie R$
- $(T \bowtie R) \bowtie S$



query plan as "query tree"

☹

- there is no full semijoin reducer
- intermediate result size bigger than output

Can we do better for cyclic queries? ☺

Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

2SAT

$\varphi = (x \vee y) \wedge (\neg y \vee z) \wedge (\neg x \vee \neg z) \wedge (z \vee y)$

- Instance: A 2-CNF formula φ
- Problem: To decide if φ is satisfiable

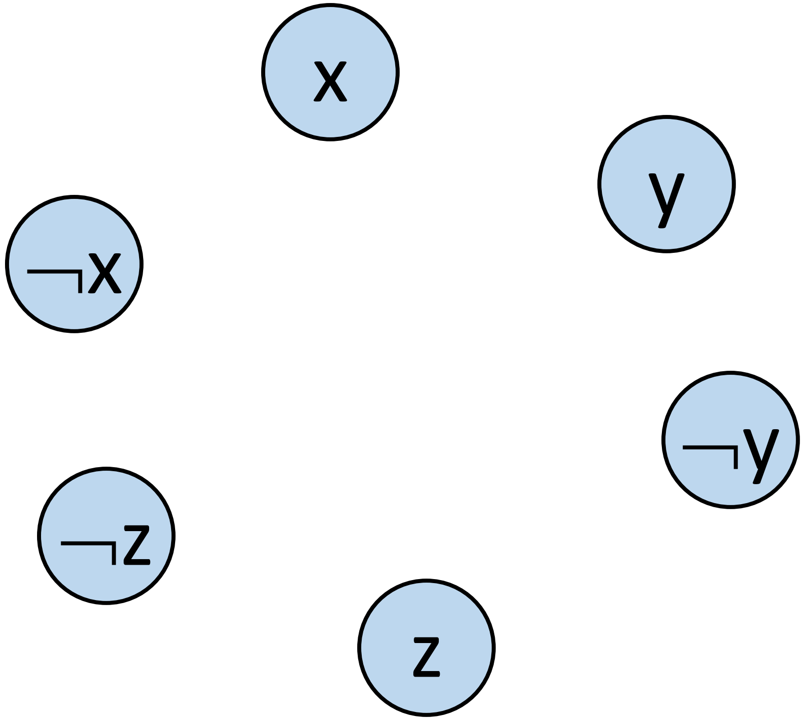
- Theorem: 2SAT is polynomial-time decidable.
  - Proof: We'll show how to solve this problem efficiently using path searches in graphs…

- Background: Given a graph $G=(V,E)$ and two vertices $s,t \in V$, finding if there is a path from s to t in G is linear-time decidable. Use some search algorithm (DFS/BFS).

2SAT: Graph Construction

$\varphi = (x \lor y) \land (\neg y \lor z) \land (\neg x \lor \neg z) \land (z \lor y)$

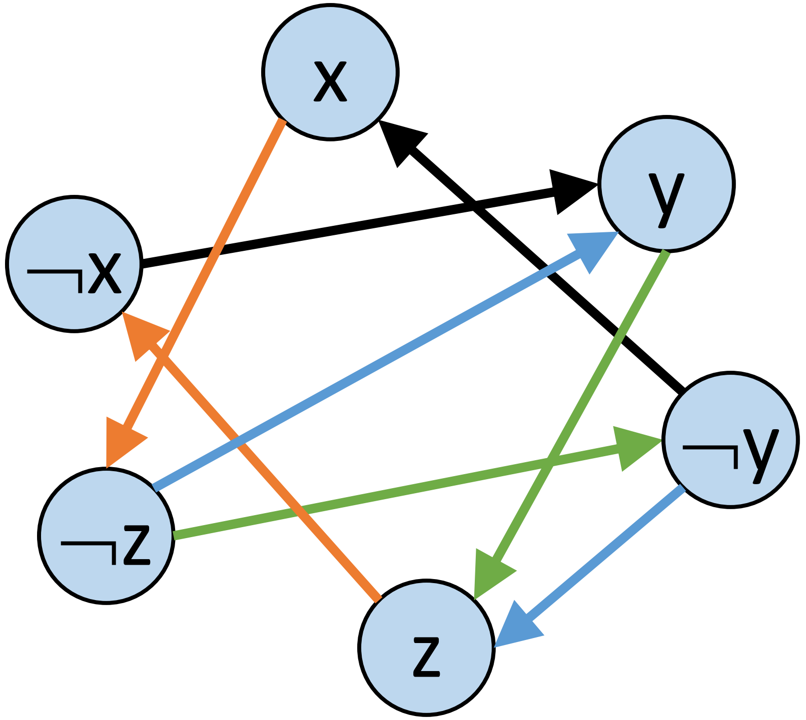
- Vertex for each variable and a negation of a variable



2SAT: Graph Construction

$\varphi = (x \vee y) \wedge (\neg y \vee z) \wedge (\neg x \vee \neg z) \wedge (z \vee y)$

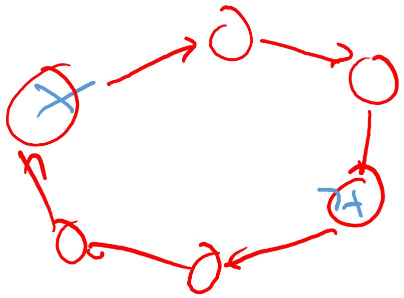
- Vertex for each variable and a negation of a variable
- Edge $(\neg x \rightarrow y)$ iff there exists a clause equivalent to $(x \vee y)$
  - Recall $(x \vee y)$ same as $(\neg x \Rightarrow y)$ and $(\neg y \Rightarrow x)$, thus also $(\neg y \rightarrow x)$

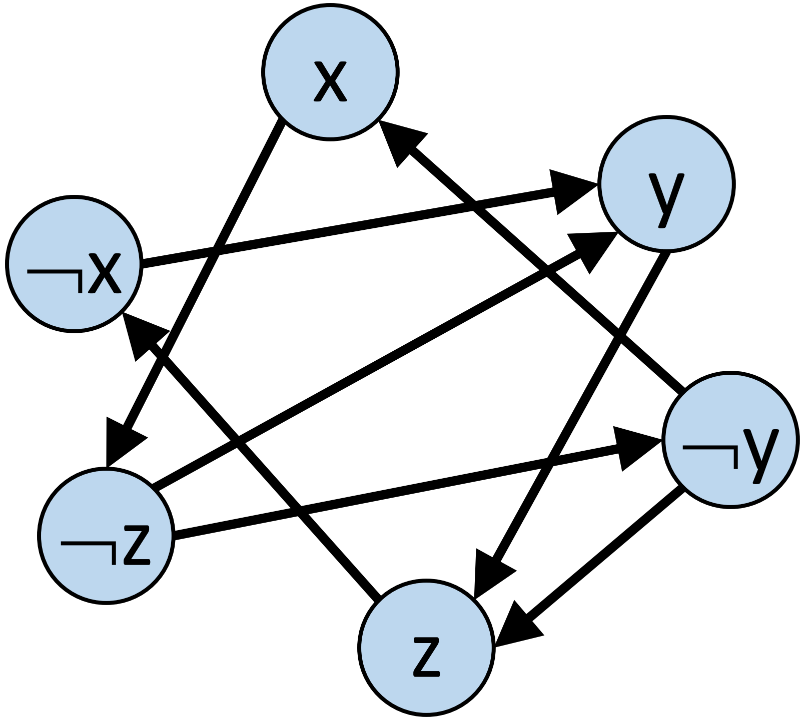


2SAT: Graph Construction

$\varphi = (x \lor y) \land (\neg y \lor z) \land (\neg x \lor \neg z) \land (z \lor y)$

- Vertex for each variable and a negation of a variable
- Edge $(\neg x \rightarrow y)$ iff there exists a clause equivalent to $(x \lor y)$
  - Recall $(x \lor y)$ same as $(\neg x \Rightarrow y)$ and $(\neg y \Rightarrow x)$, thus also $(\neg y \rightarrow x)$
- Claim: a 2-CNF formula φ is unsatisfiable iff there exists a variable x, such that:
  - there is a path from x to $\neg$x in the graph, and
  - there is a path from $\neg$x to x in the graph



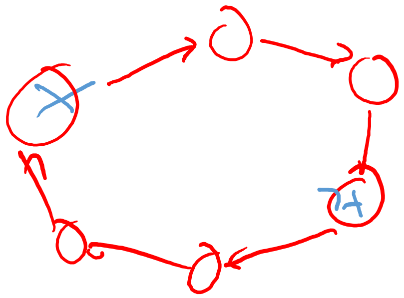


2SAT: Graph Construction

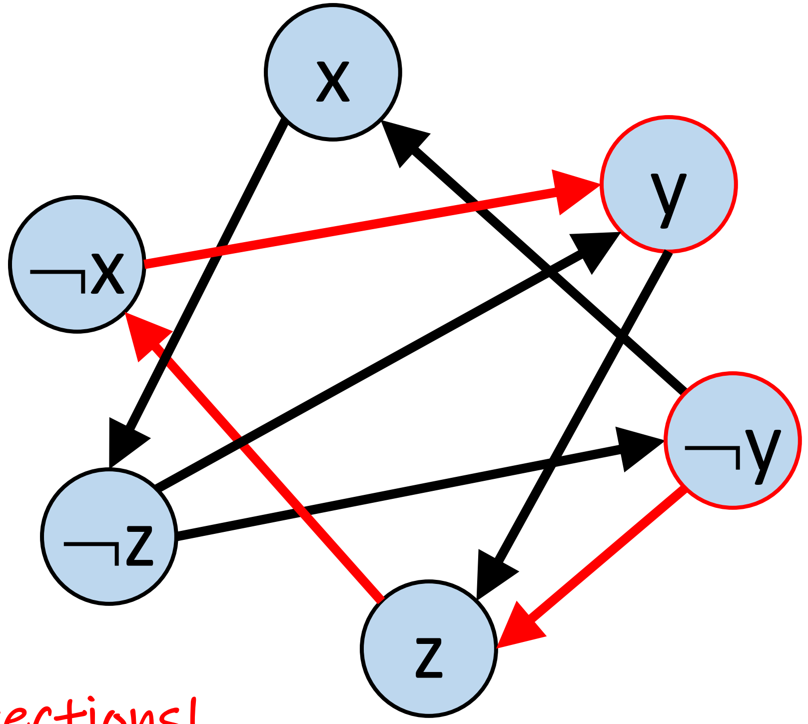
$\varphi = (x \lor y) \land (\neg y \lor z) \land (\neg x \lor \neg z) \land (z \lor y)$

- Vertex for each variable and a negation of a variable
- Edge $(\neg x \rightarrow y)$ iff there exists a clause equivalent to $(x \lor y)$
  - Recall $(x \lor y)$ same as $(\neg x \Rightarrow y)$ and $(\neg y \Rightarrow x)$, thus also $(\neg y \rightarrow x)$

- Claim: a 2-CNF formula φ is unsatisfiable iff there exists a variable x, such that:
  - there is a path from x to $\neg$x in the graph, and
  - there is a path from $\neg$x to x in the graph



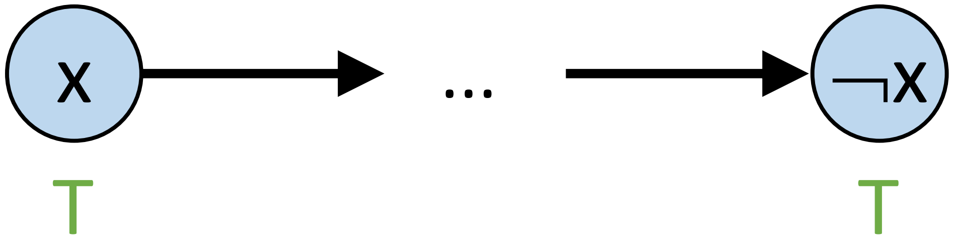
not enough,
needs both directions!



Correctness (1)

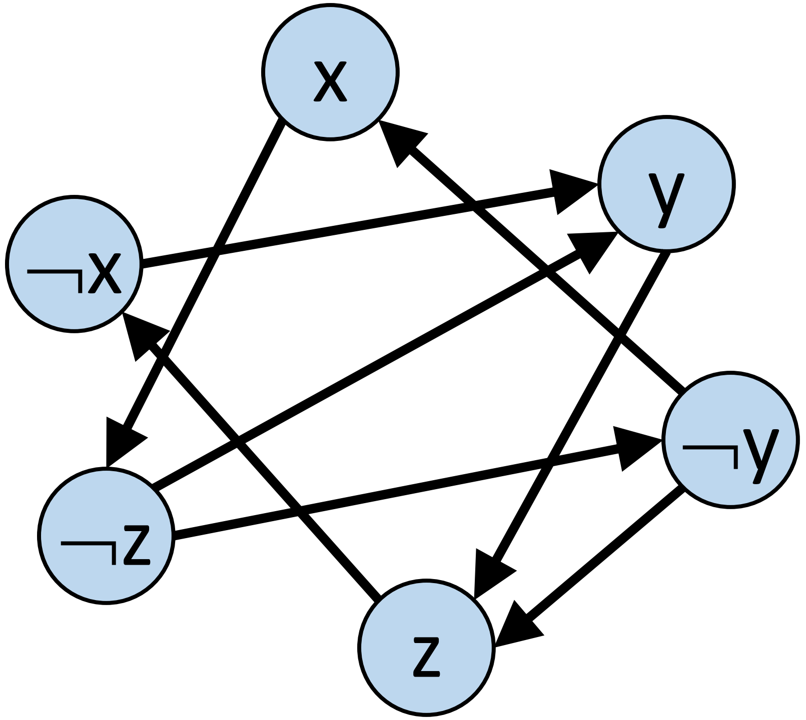
$\varphi = (x \lor y) \land (\neg y \lor z) \land (\neg x \lor \neg z) \land (z \lor y)$

- Suppose there are paths $x..\neg x$ and $\neg x..x$ for some variable x, but there's also a satisfying assignment ρ.
  - If $\rho(x)$=T:



  - Similarly for $\rho(x)$=F...

recall, needs to hold in both directions!

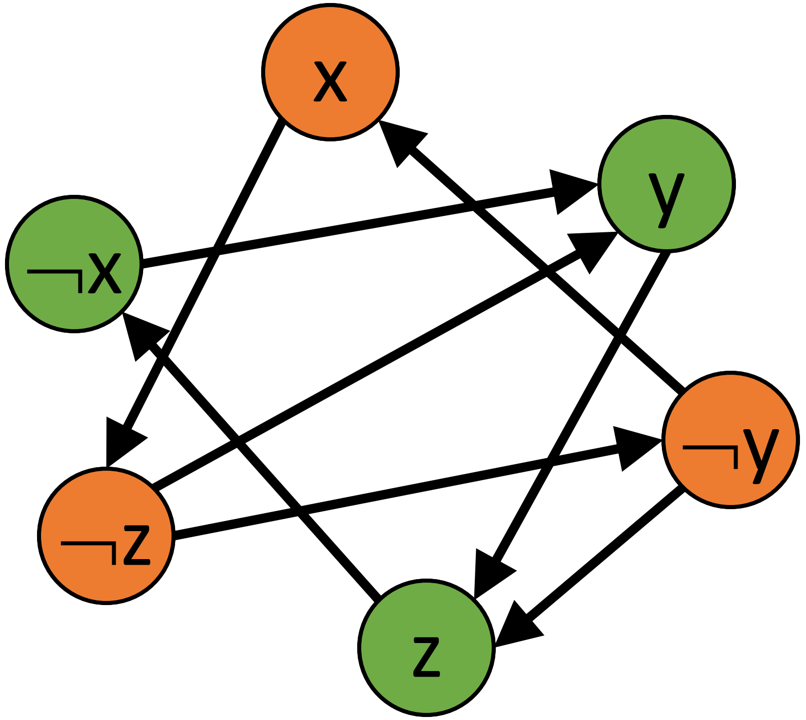


Correctness (2)

$\varphi = (x \vee y) \wedge (\neg y \vee z) \wedge (\neg x \vee \neg z) \wedge (z \vee y)$

- Suppose there are no variables with such paths.
- Construct an assignment as follows:

1. pick an unassigned literal α, with no path from α to $\neg\alpha$, and assign it T
2. assign T to all reachable vertices
3. assign F to their negations
4. Repeat until all vertices are assigned



2SAT is in P

We get the following PTIME algorithm for 2SAT:

- For each variable x find if there is a path from x to $\neg x$ and vice-versa.
- Reject if any of these tests succeeded.
- Accept otherwise.

$\Rightarrow$ 2SAT $\in$ P. ■

Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

Join Processing: two approaches

1. Cardinality-based
   - binary joins, consider the sizes of input relations as to reduce the intermediate sizes
   - commercial DBMSs: series of pairwise joins, system R (Selinger), join size estimation
2. Structural approaches (next)
   - acylicity: Yannakakis, GYO algorithm, join tree
   - bounded "width": query width, hypertree width (hw), generalized hw (ghw). All go back to notion of treewidth (work by Robertson & Seymour on graph minors)

AGM: fractional hw (fhw):

- consider both statistics on relations and query structure

Tree decomposition

In graph theory, a **tree decomposition** is a mapping of a graph into a tree that can be used to define the treewidth of the graph and speed up solving certain computational problems on the graph.

Tree decompositions are also called **junction trees**, **clique trees**, or **join trees**. They play an important role in problems like probabilistic inference, constraint satisfaction, query optimization, [*citation needed*] and matrix decomposition.

The concept of tree decomposition was originally introduced by Rudolf Halin (1976). Later it was rediscovered by Neil Robertson and Paul Seymour (1984) and has since been studied by many other authors.[1]

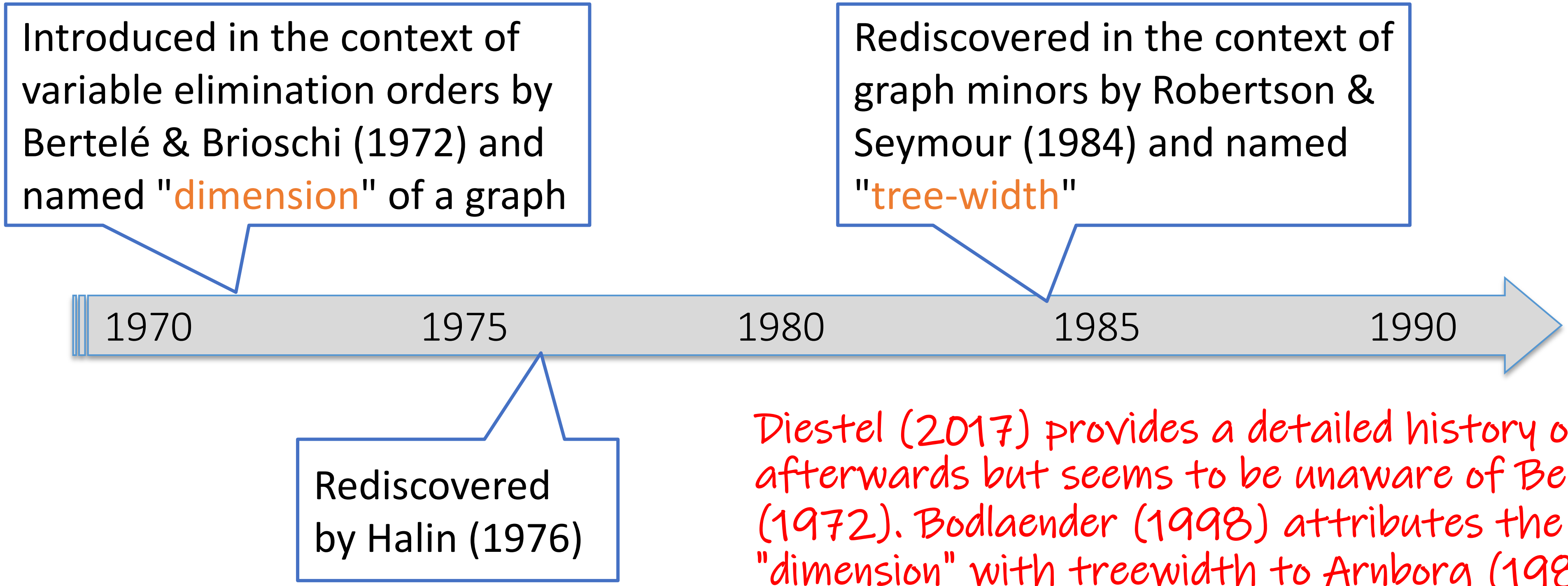
Dynamic programming [edit]

At the beginning of the 1970s, it was observed that a large class of combinatorial optimization problems defined on graphs could be efficiently solved by non-serial dynamic programming as long as the graph had a bounded *dimension*,[5] a parameter related to treewidth. Later, several authors independently observed, at the end of the 1980s,[6] that many algorithmic problems that are NP-complete for arbitrary graphs may be solved efficiently by dynamic programming for graphs of bounded treewidth, using the tree-decompositions of these graphs.

- Robertson, Neil; Seymour, Paul D. (1984), "Graph minors III: Planar tree-width", *Journal of Combinatorial Theory*, Series B, **36** (1): 49–64, doi:10.1016/0095-8956(84)90013-3.

Very incomplete history of treewdith

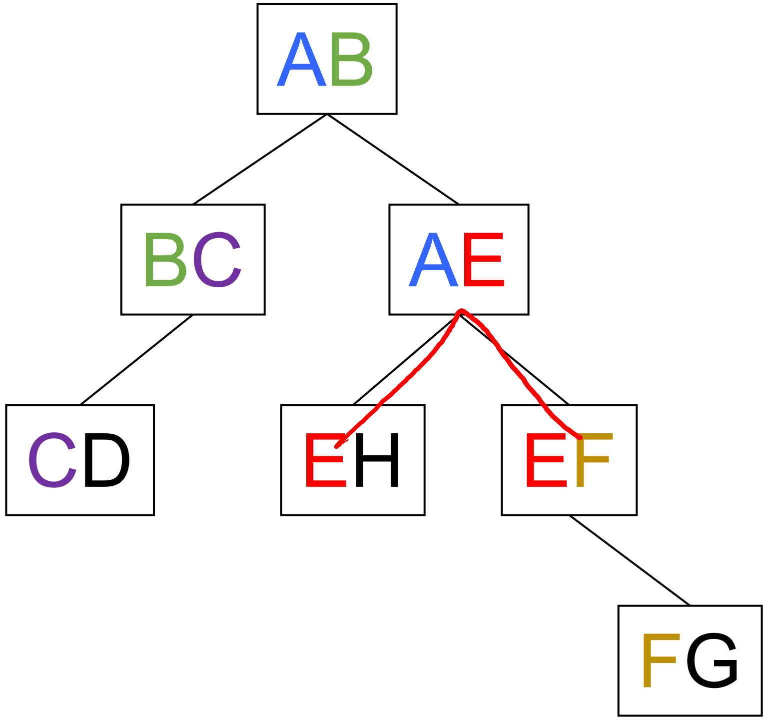
The treewidth of a graph is an important graph complexity parameter that determines the runtime of practical algorithms. Intuitively measures how close a graph is to being a tree.



Diestel (2017) provides a detailed history of what happened afterwards but seems to be unaware of Bertelé & Brioschi (1972). Bodlaender (1998) attributes the connection of "dimension" with treewidth to Arnborg (1985) who actually never uses the word "treewidth" nor references R&S (1984)...

Bertelè, Brioschi. Nonserial Dynamic Programming, 1972 (definition 2.7.8). https://dl.acm.org/doi/10.5555/578817 , Halin. S-functions for graphs, Journal of Geometry, 1976. https://doi.org/10.1007%2FBF01917434 , Robertson, Seymour. Graph minors III: Planar tree-width, Journal of Combinatorial Theory, 1984 https://doi.org/10.1016%2F0095-8956%2884%2990013-3 , Diestel. Graph theory, 5th ed, 2017 (section 12). https://doi.org/10.1007/978-3-662-53622-3 , Bodlaender. A partial k-arboretum of graphs with bounded treewidth (tutorial), Theoretical Computer Science, 1998. https://doi.org/10.1016/S0304-3975(97)00228-4 , Arnborg. Efficient algorithms for combinatorial problems on graphs with bounded decomposability -- a survey, BIT, 1985. https://dl.acm.org/doi/abs/10.5555/3765.3773

Definition of an attribute-connected tree



DEFINITION: A tree is attribute-connected if the subtree induced by each attribute is connected

Same as the running intersection property from join trees (also known as junction tree)

Also called "coherence"

Tree decomposition

A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one

Tree decomposition example 1: a tree

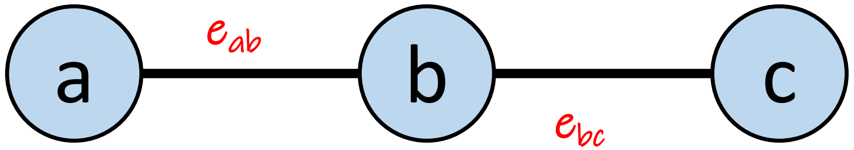
A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v \subseteq N}$ assigned to each vertex (or "supernode") $\mathrm{v \in V}$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



tree decomposition

?

Tree decomposition example 1: a tree

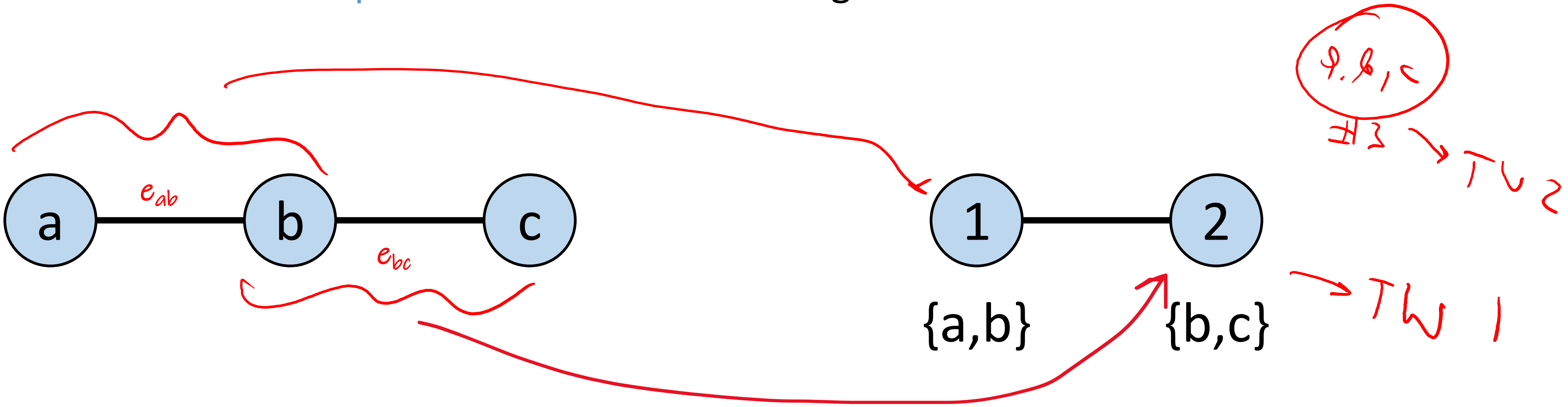
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



That's why treewidth defined as max cardinality - 1

Tree decomposition example 2

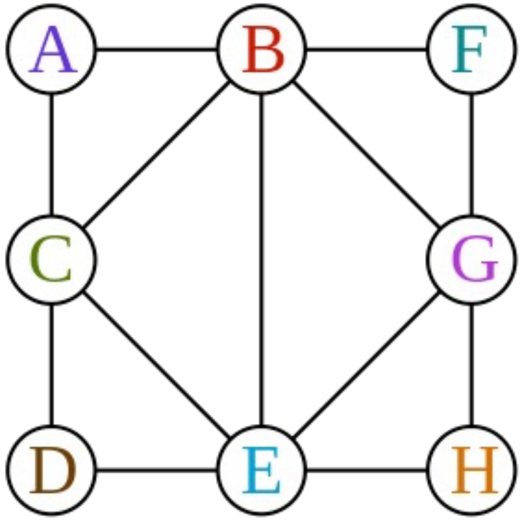
A tree decomposition of graph $\mathrm{G}(\mathrm{N}, \mathrm{E})$ is a tree $\mathrm{T}(\mathrm{V}, \mathrm{F})$ and a subset $\mathrm{N}_{\mathrm{v}} \subseteq \mathrm{N}$ assigned to each vertex (or "supernode") $\mathrm{v} \in \mathrm{V}$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



tree decomposition

?

Tree decomposition example 2

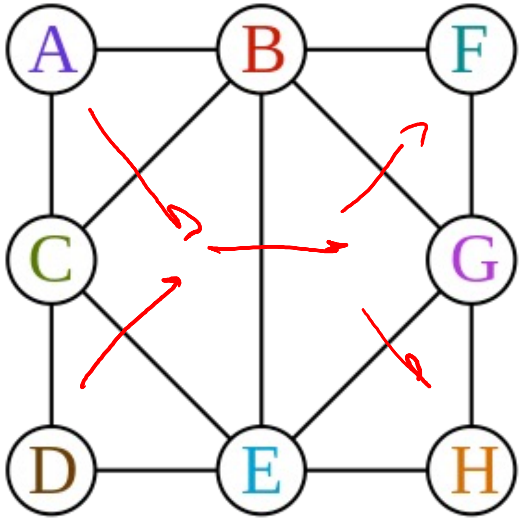
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

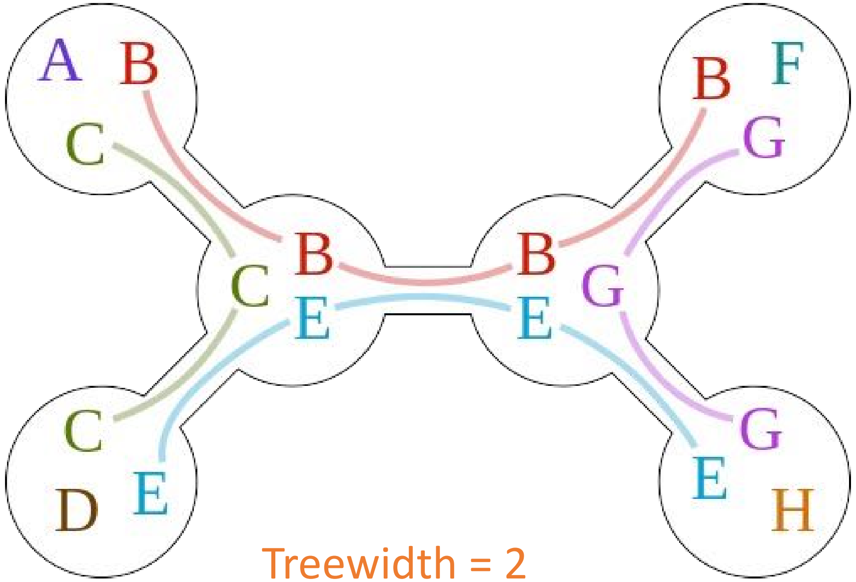
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





Treewidth = 2

Notice running intersection property

Tree decomposition example 3

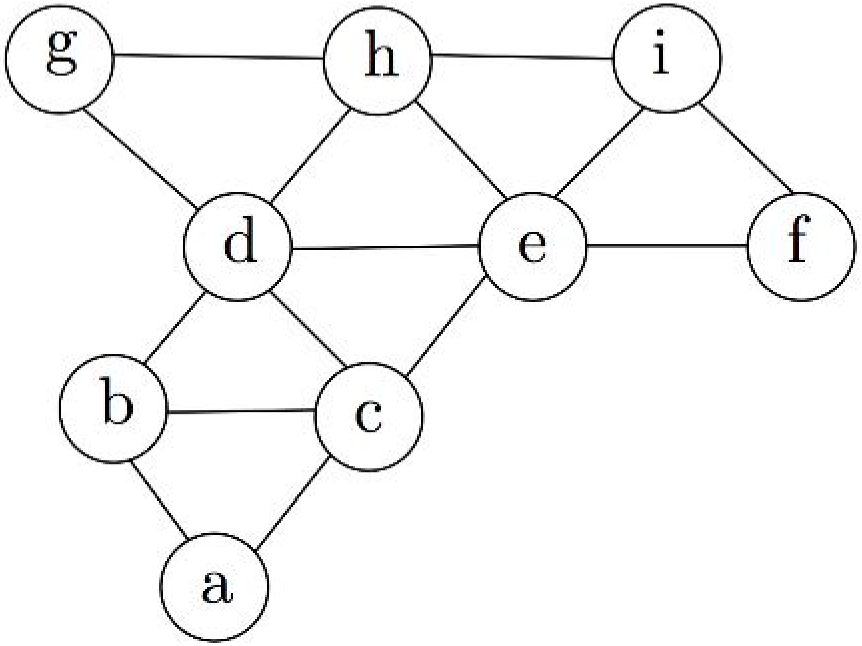
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



tree decomposition

?

Example from: https://www.mi.fu-berlin.de/en/inf/groups/abi/teaching/lectures/lectures_past/WS0910/V____Discrete_Mathematics_for_Bioinformatics__P1/material/scripts/treedecomposition1.pdf

Tree decomposition example 3

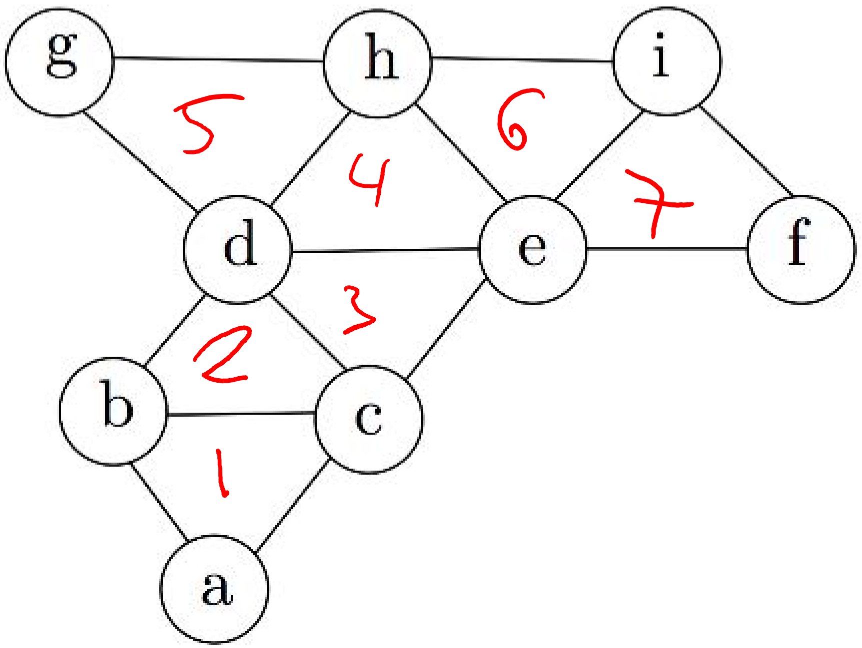
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

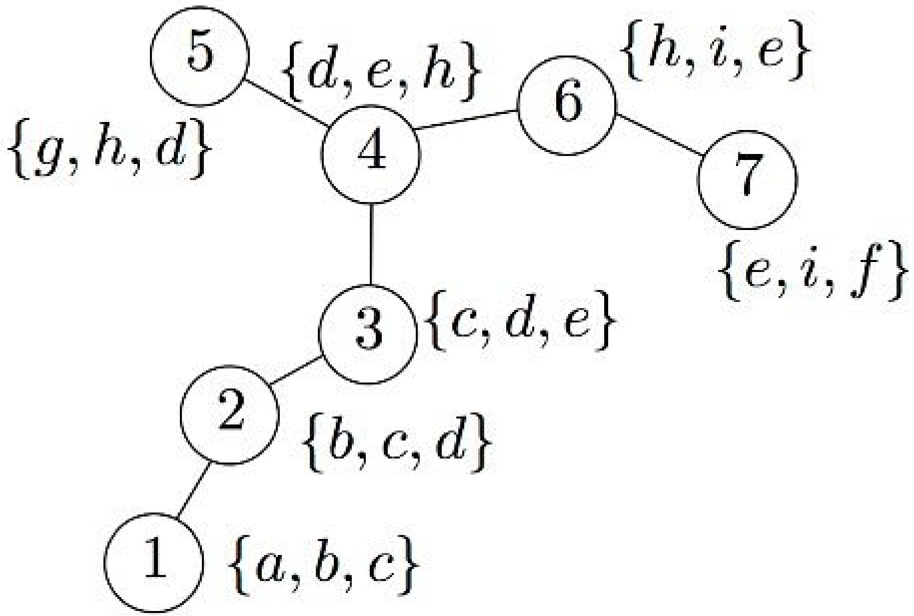
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





Tree decomposition example 4: a cycle

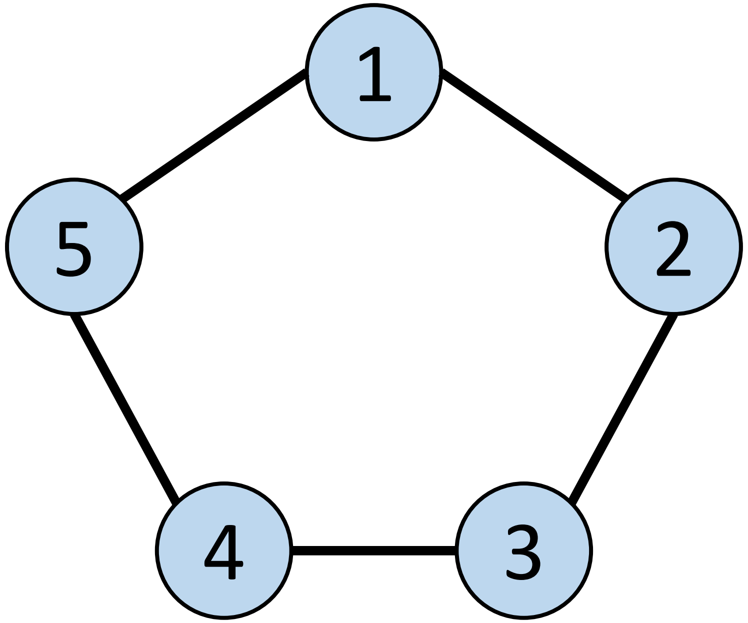
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



tree decomposition

?

Tree decomposition example 4: a cycle

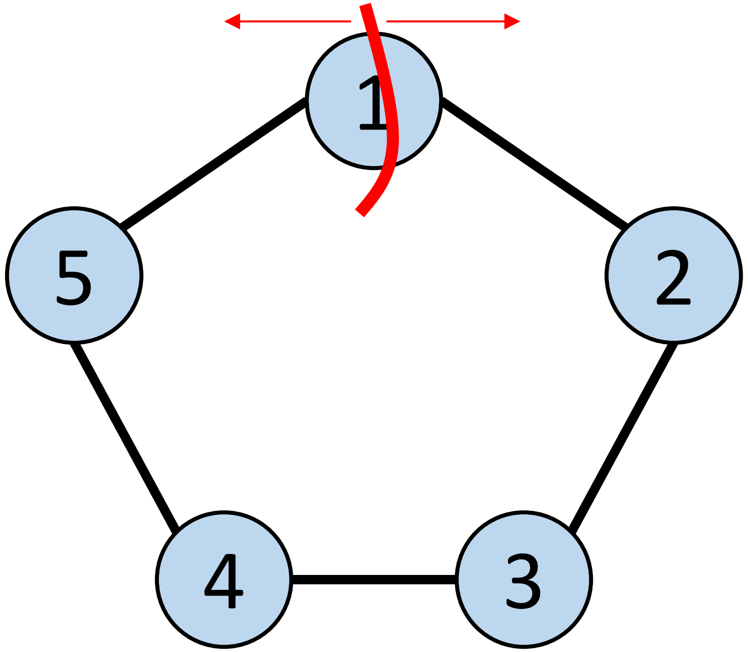
A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v} \subseteq \mathrm{N}$ assigned to each vertex (or "supernode") $\mathrm{v} \in \mathrm{V}$ s.t.:

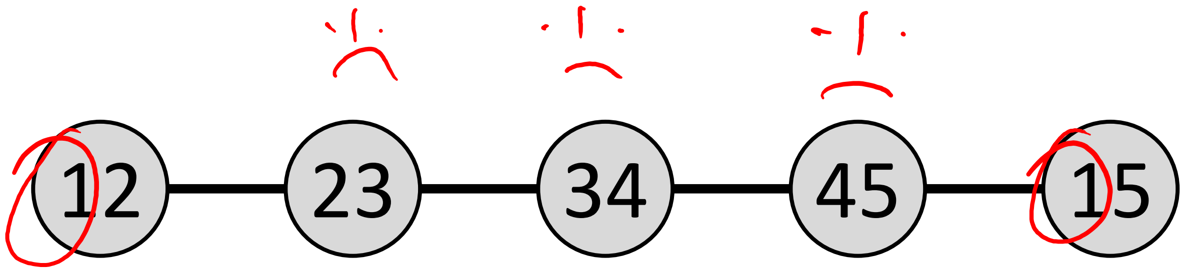
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





What about coherence?

Tree decomposition example 4: a cycle

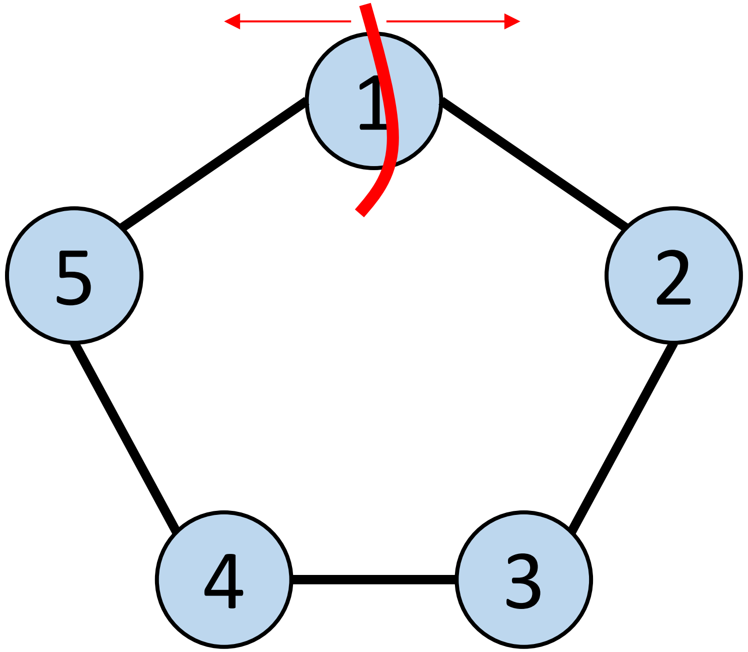
A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v} \subseteq \mathrm{N}$ assigned to each vertex (or "supernode") $\mathrm{v} \in \mathrm{V}$ s.t.:

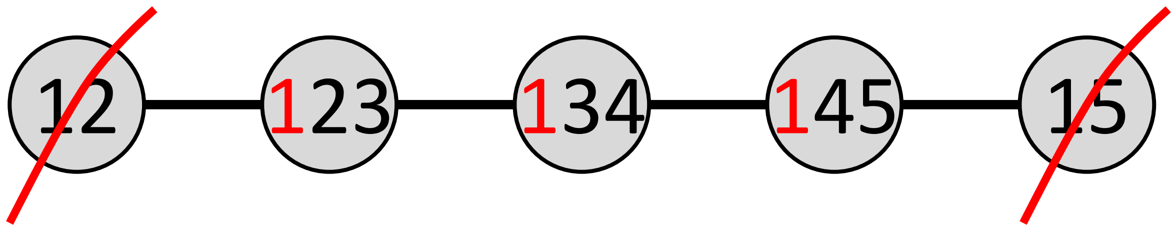
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





Tree decomposition example 4: a cycle

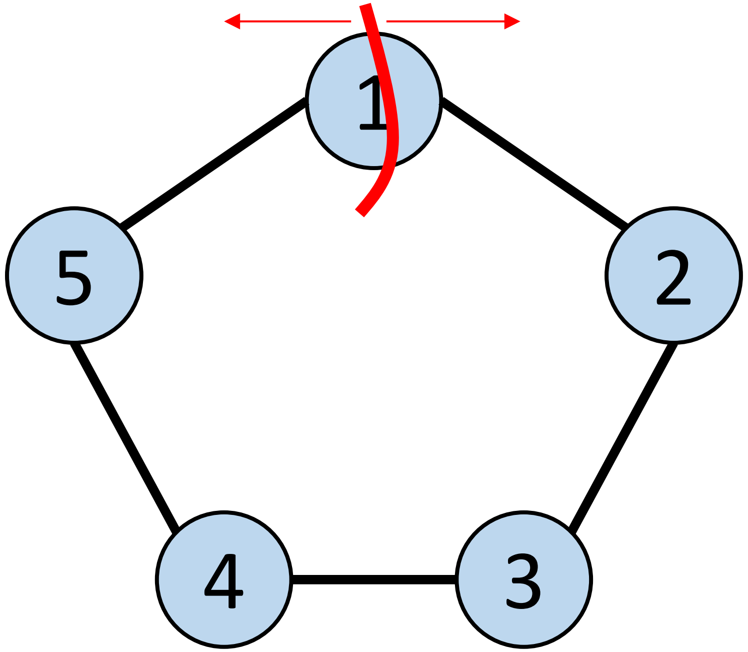
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

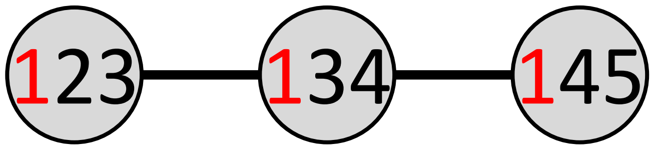
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





Tree decomposition example 5: the triangle

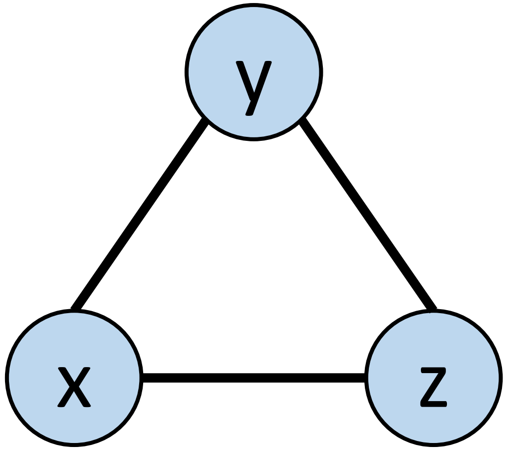
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



tree decomposition

?

Tree decomposition example 5: the triangle

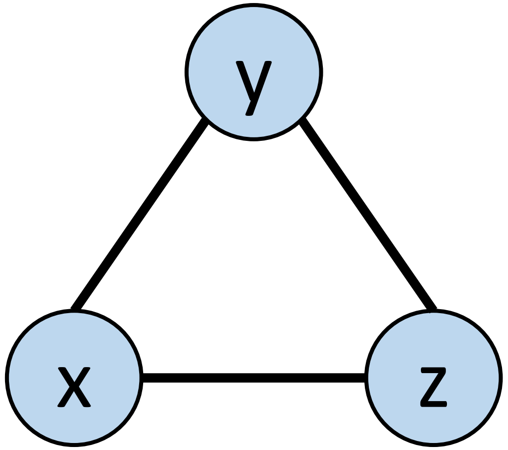
A tree decomposition of graph $G(N, E)$ is a tree $T(V, F)$ and a subset $N_v \subseteq N$ assigned to each vertex (or "supernode") $v \in V$ s.t.:

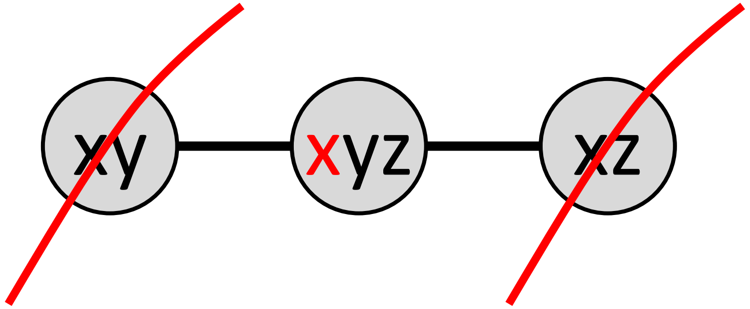
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





Tree decomposition example 5: the triangle

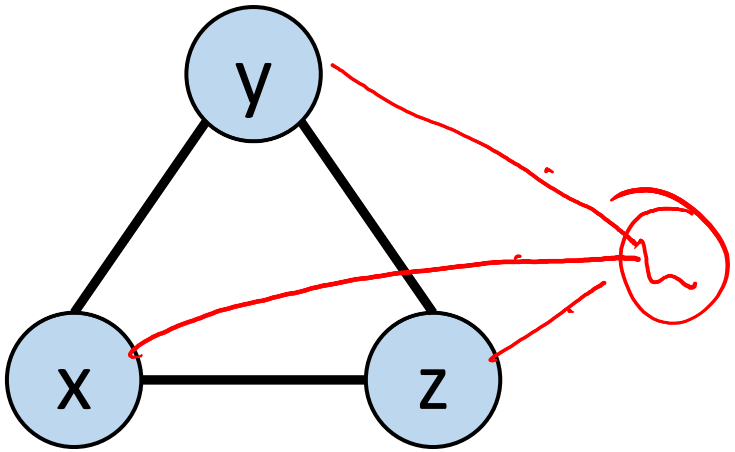
A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v \subseteq N}$ assigned to each vertex (or "supernode") $\mathrm{v \in V}$ s.t.:

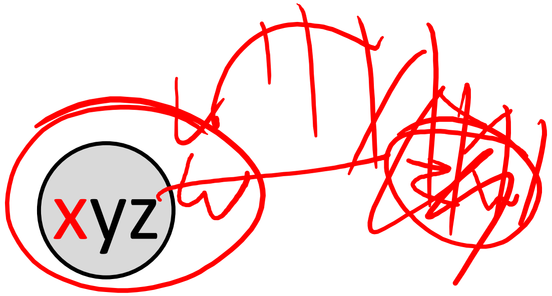
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one





More generally, a K_d (d-clique) has a minimal treewidth of d-1

Tree decomposition example 6: a longer tree

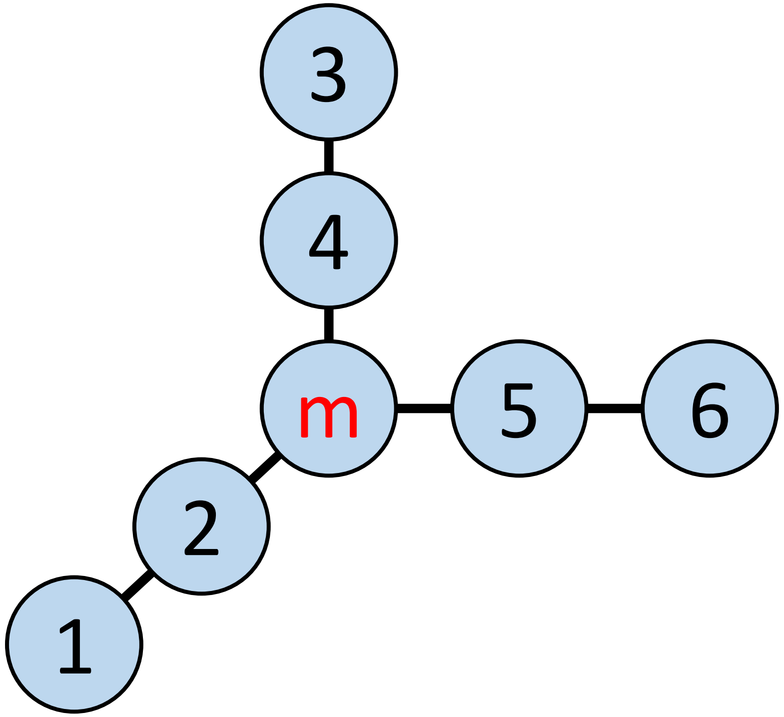
A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v} \subseteq \mathrm{N}$ assigned to each vertex (or "supernode") $\mathrm{v} \in \mathrm{V}$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one



tree decomposition

?

Tree decomposition example 6: a longer tree

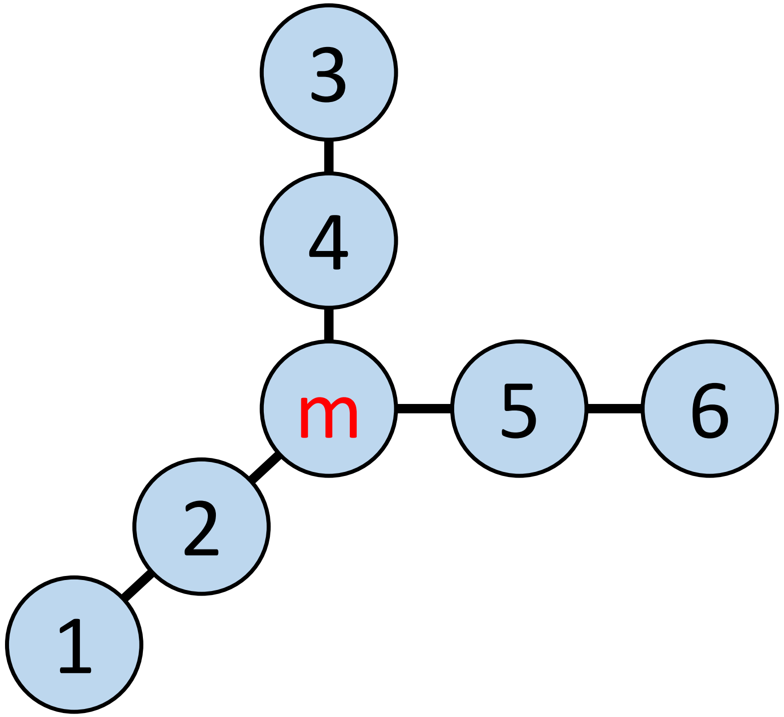
A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v \subseteq N}$ assigned to each vertex (or "supernode") $\mathrm{v \in V}$ s.t.:

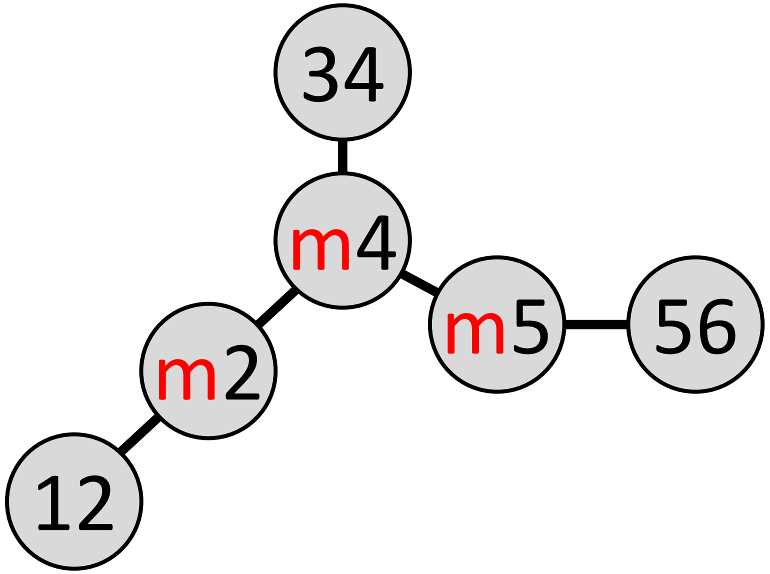
(1) Node coverage: Every vertex of G is assigned at least one vertex in T

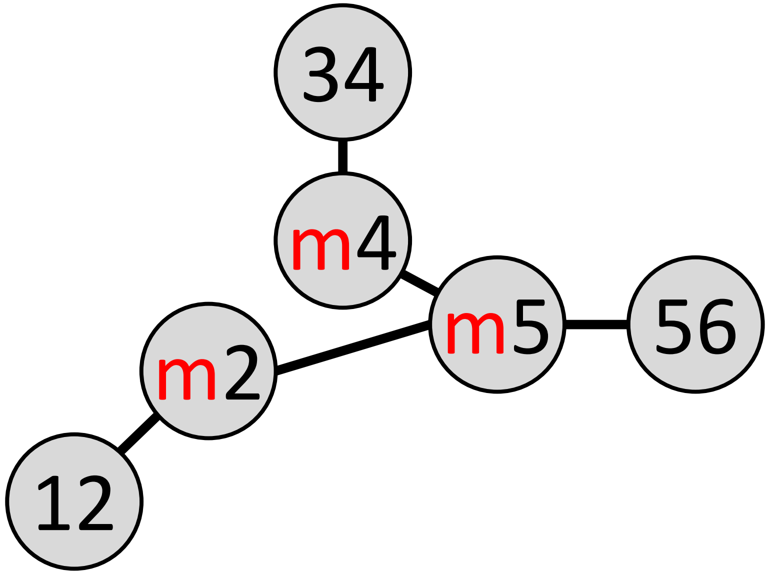
(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

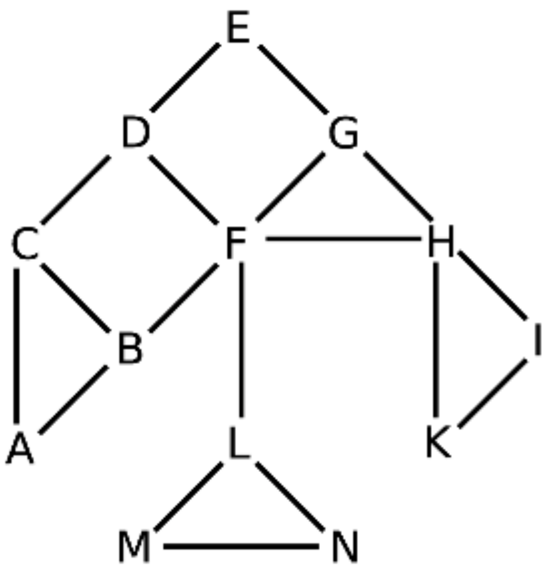
The width of a tree decomposition is the size of its largest set minus one





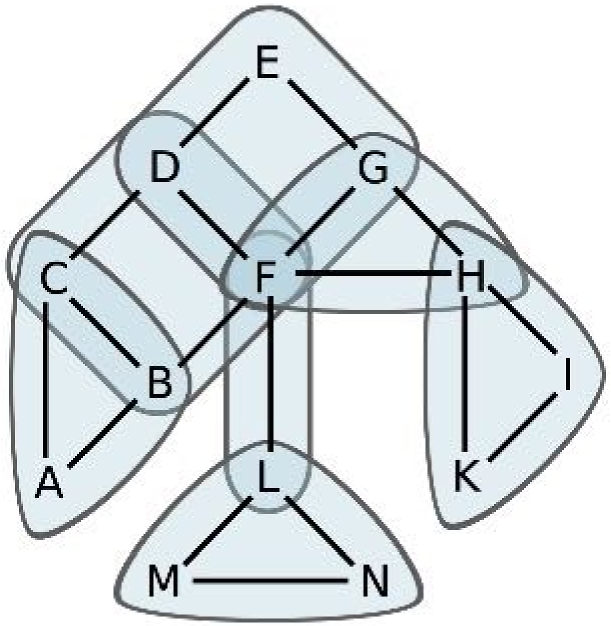


Tree decomposition example 7



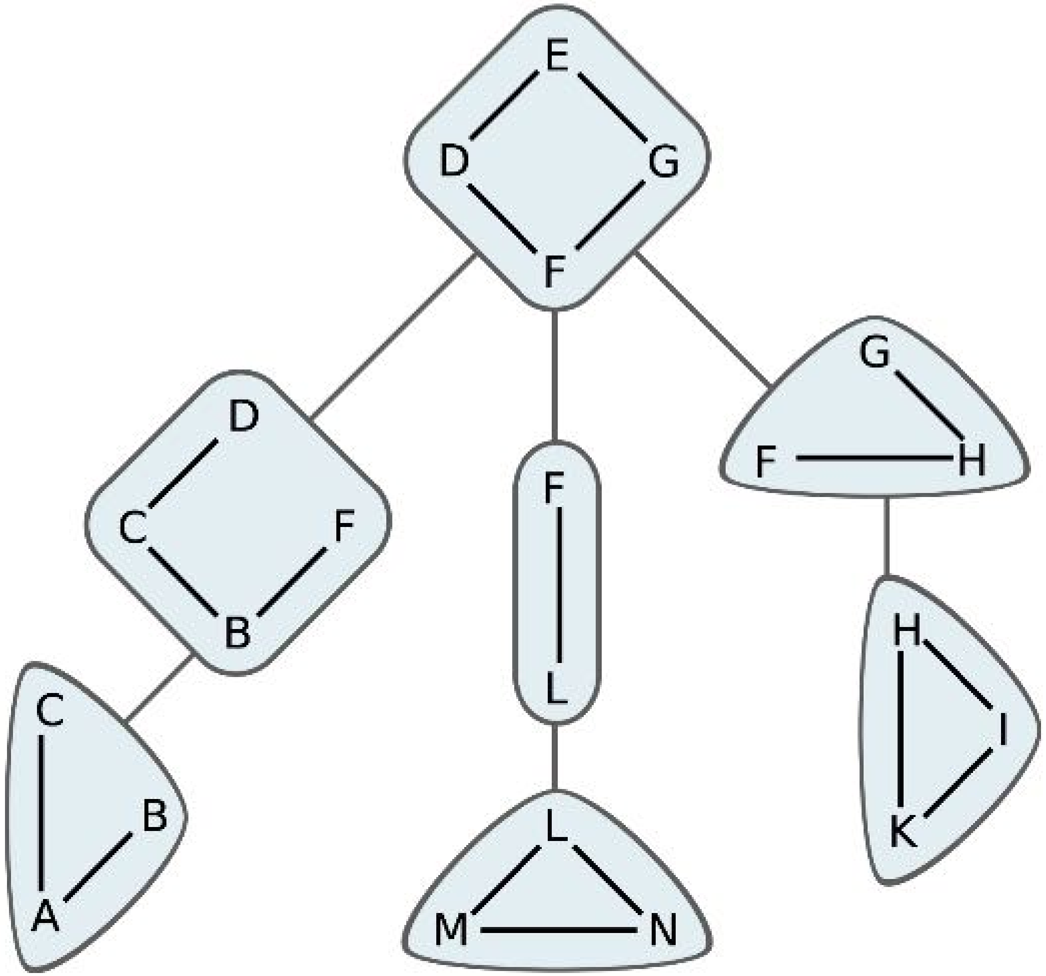
Example by: Markus Krötzsch. "Database theory: Lecture 6: Tree-like Conjunctive Queries." 2016. https://iccl.inf.tu-dresden.de/web/Database_Theory_(SS2016)/en

Tree decomposition example 7



Example by: Markus Krötzsch. "Database theory: Lecture 6: Tree-like Conjunctive Queries." 2016. https://iccl.inf.tu-dresden.de/web/Database_Theory_(SS2016)/en

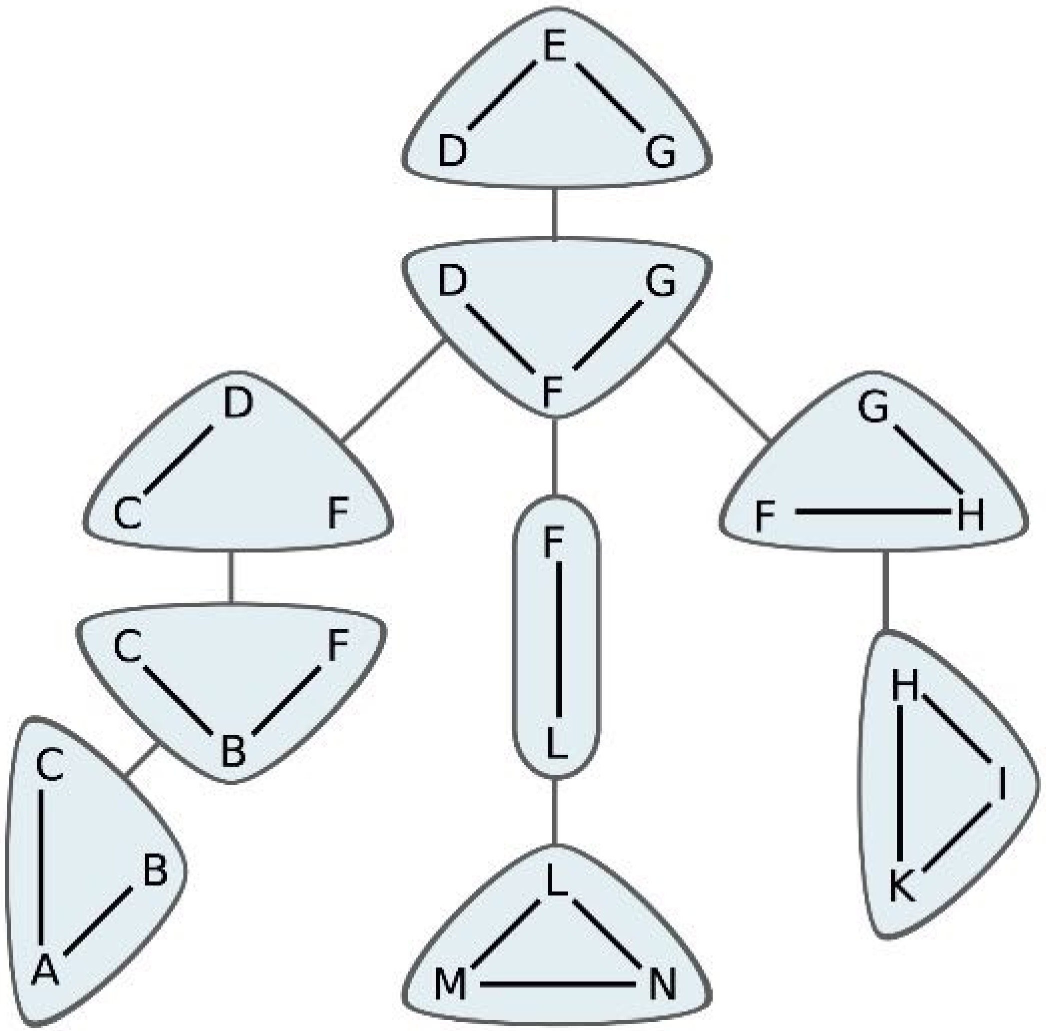
Tree decomposition example 7



⤳ tree decomposition of width 3

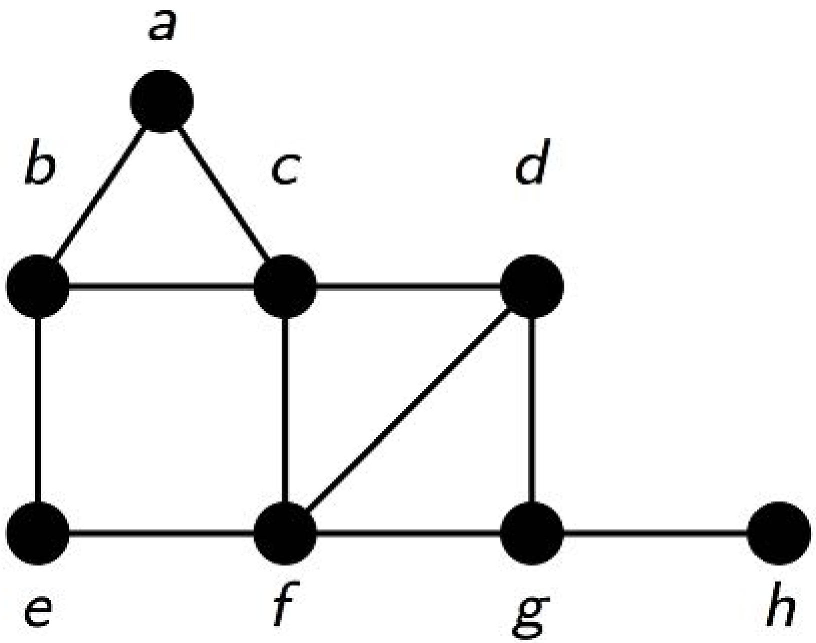
Example by: Markus Krötzsch. "Database theory: Lecture 6: Tree-like Conjunctive Queries." 2016. https://iccl.inf.tu-dresden.de/web/Database_Theory_(SS2016)/en

Tree decomposition example 7

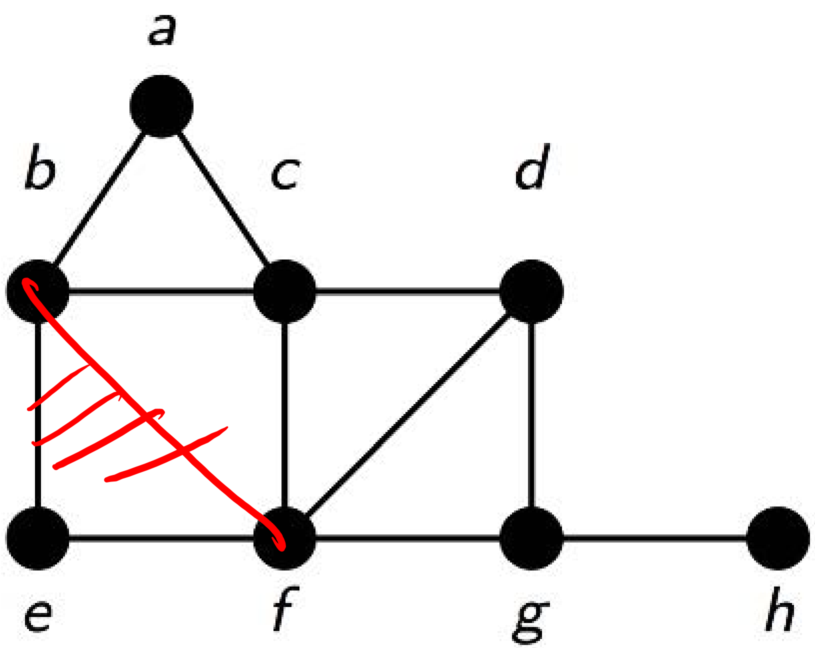


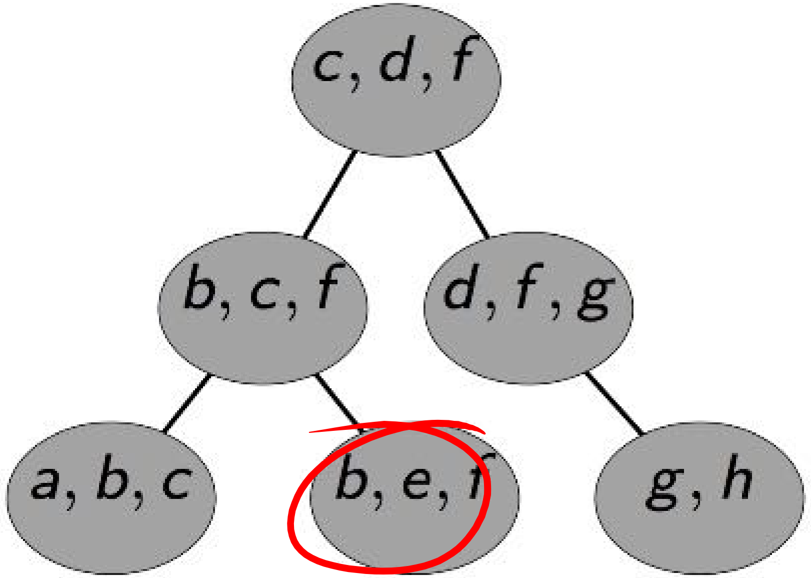
⇝ tree decomposition of width 2 = treewidth of the example graph

Tree decomposition example 8



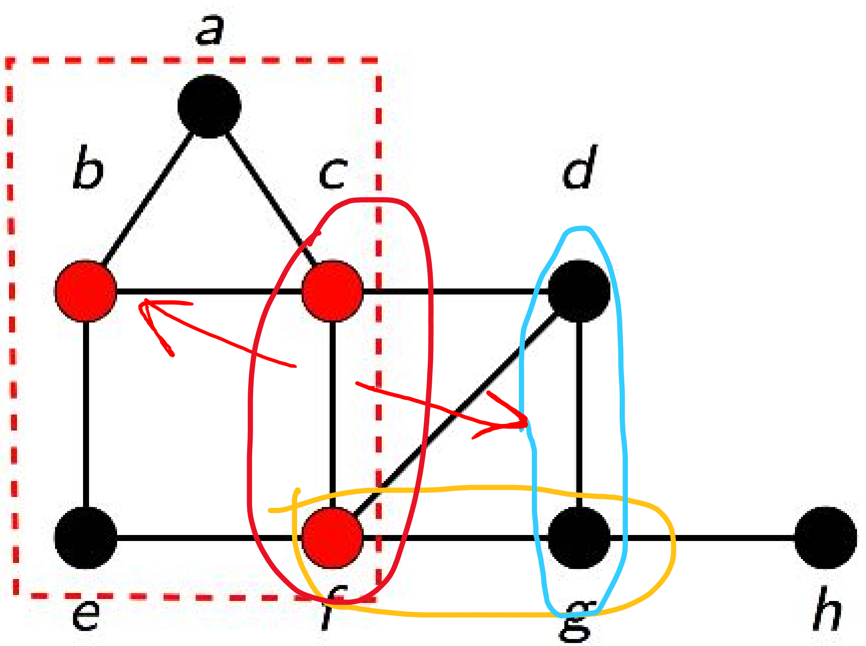
Tree decomposition example 8

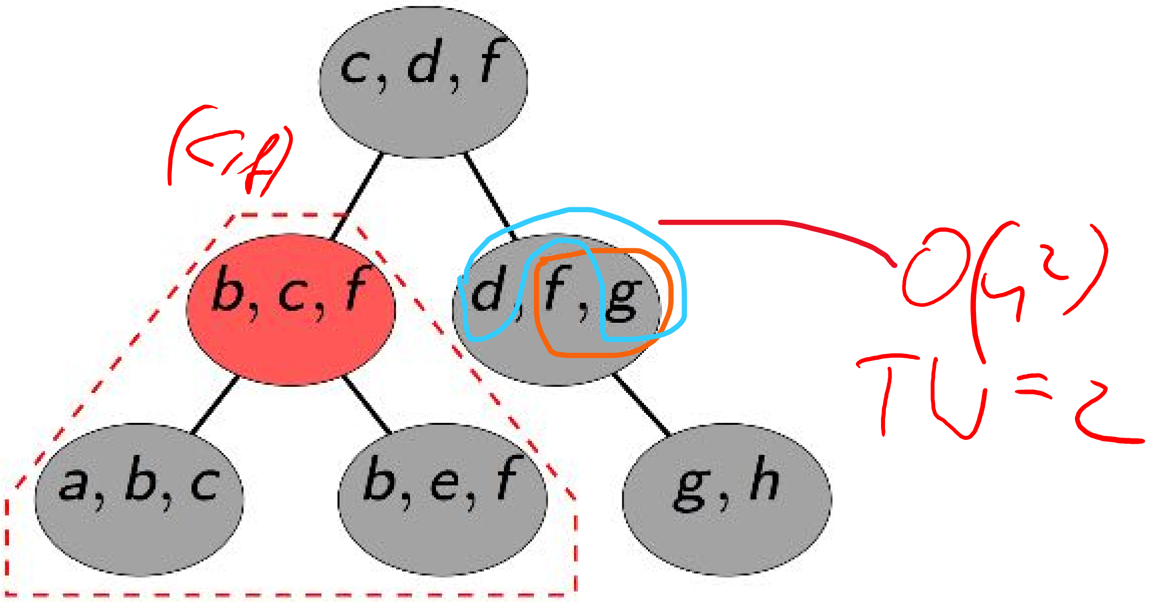




Example by: Marx. "Graphs, hypergraphs, and the complexity of conjunctive database queries", ICDT 2017. http://edbticdt2017.unive.it/marx-icdt2017-talk.pdf

Tree decomposition example 8

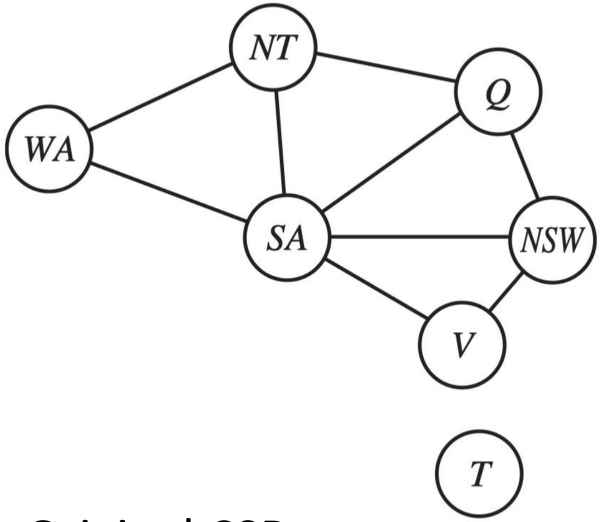




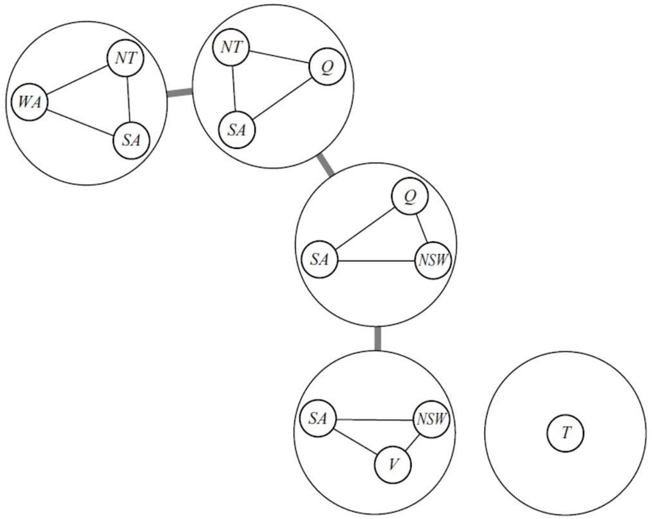
A subtree communicates with the outside world only via the root of the subtree.

Tree Decompositions (TDs) for CSPs

Notice here each node is a variable with domain of size d (e.g. 3 colors)



Original CSP:
Map-coloring of Australia



Tree decomposition with supernodes (sets of variables)

TD:

- If two variables are connected in the original problem, they must appear together (along with the constraint) in at least one supernode
- If a variable occurs in two supernodes in the TD, it must appear in every supernode on the path that connects the two (coherence)
- The only constraints between the supernodes are that the variables take on the same values across supernodes (like semi-join messages from Yannakakis)

Translates into $\boxed{O(n^{tw})}$ where n is size of constraints per edge

- Solving CSP on a tree with k variables and domain size m is $O(km^2)$
- TD algorithm: find all solutions within each supernode, which is $\boxed{O(m^{tw+1})}$ where tw is the treewidth (= one less than size of largest supernode). Recall treewidth of tree is 1, thus complexity $O(m^2)$
- Then, use the tree-structured Yannakakis algorithm, treating the supernodes as new variables...
- Finding a tree decomposition of smallest treewidth is NP-complete, but good heuristic methods exist.

Alternative definition of Tree decomposition (TD)

A tree decomposition of graph $\mathrm{G(N, E)}$ is a tree $\mathrm{T(V, F)}$ and a subset $\mathrm{N_v} \subseteq \mathrm{N}$ assigned to each vertex (or "supernode") $\mathrm{v} \in \mathrm{V}$ s.t.:

(1) Node coverage: Every vertex of G is assigned at least one vertex in T

(2) Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

(3) Coherence: The tree is "attribute-connected"

The width of a tree decomposition is the size of its largest set minus one

ALTERNATIVE DEFINITION:

A tree decomposition of graph $G(N, E)$ is a pair $\langle T, \chi \rangle$ where $T(V, F)$ is a tree, and χ is a labeling function assigning to each vertex $v \in V$ a set of vertices $\chi(v) \subseteq N$, s.t. above conditions (2) and (3) are satisfied.

Small decompositions allow to "compress" the search space

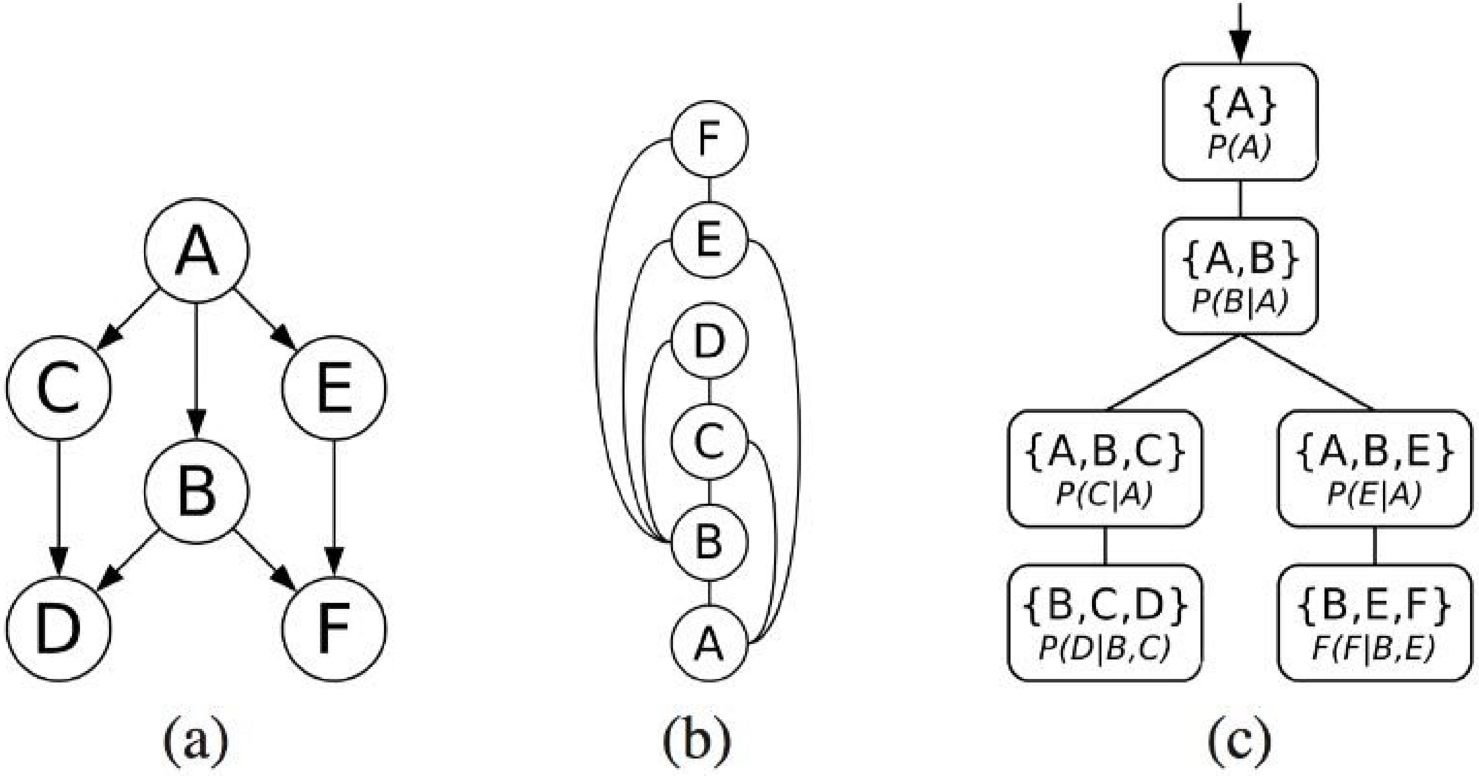


Figure 1: Example belief network, its triangulated primal graph along ordering $d = A, B, C, D, E, F$, and the corresponding bucket tree decomposition.

Figure from: Otten, Dechter. Bounding Search Space Size via (Hyper)tree Decompositions. UAI 2008. https://arxiv.org/abs/1206.3284

Explaining Treewidth with cops & robbers

Pursuit-evasion games

- Pursuit-evasion (sometimes called "cops and robber") is a family of problems in which one group (cops) attempts to track down members of another group (robbers) in some structured environment, usually graphs.
- Related to pebble games and Ehrenfeucht–Fraïssé games
- Next: A variations of "Cops and Robber" can be used to describe the treewidth of a graph

For more details see: https://en.wikipedia.org/wiki/Pursuit%E2%80%93evasion, https://en.wikipedia.org/wiki/Pebble_game, https://en.wikipedia.org/wiki/Ehrenfeucht%E2%80%93Fra%C3%AFss%C3%A9_game, https://en.wikipedia.org/wiki/Cop_number#Special_classes_of_graphs

Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

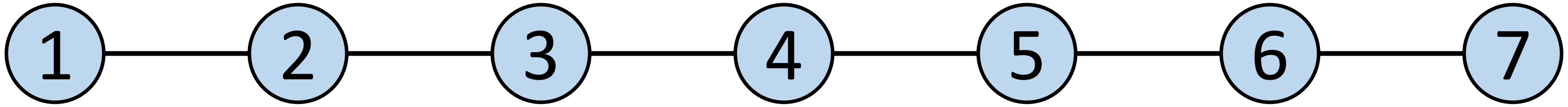
(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

THEOREM [Seymour & Thomas (1993)]

You have a winning strategy with k cops iff the tree-width of the graph is at most k−1.



Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

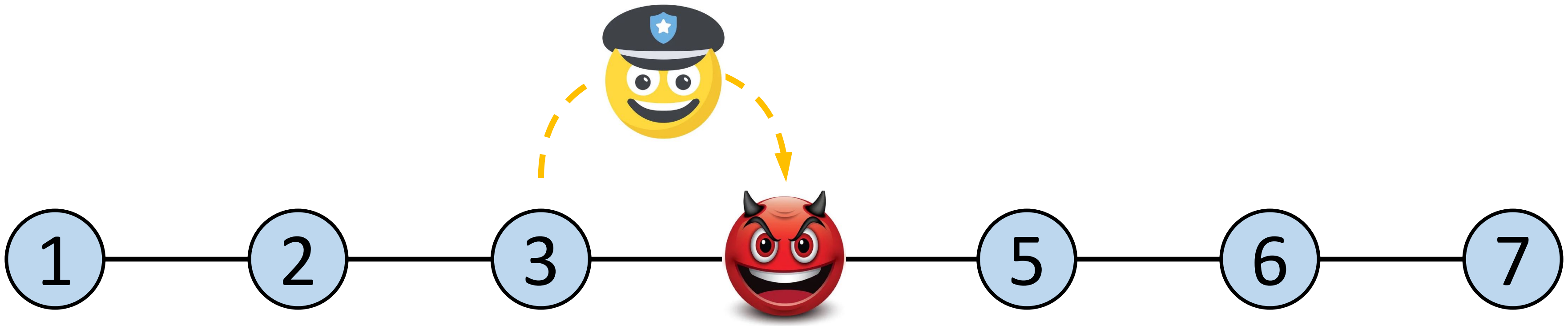
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027
Wolfgang Gatterbauer. Principles of scalable data management: https://northeastern-datalab.github.io/cs7240/

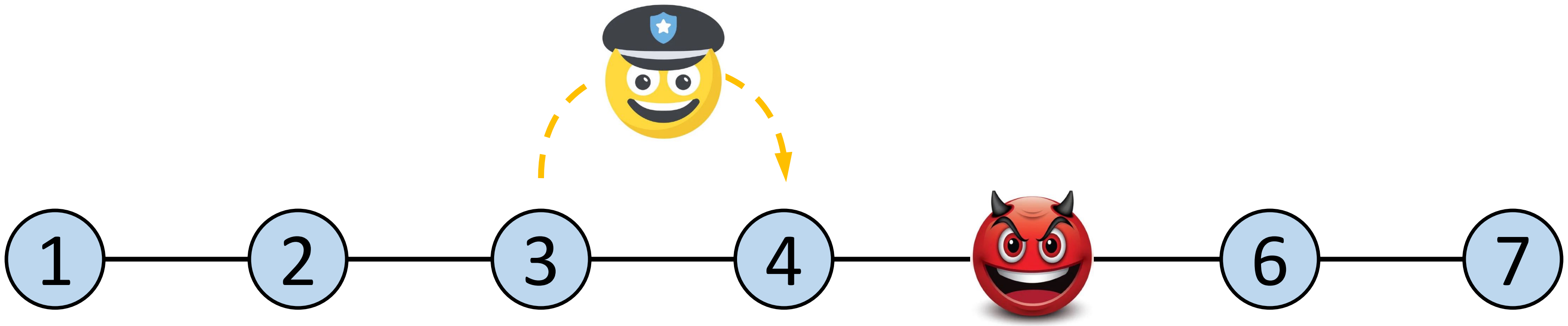
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027
Wolfgang Gatterbauer. Principles of scalable data management: https://northeastern-datalab.github.io/cs7240/

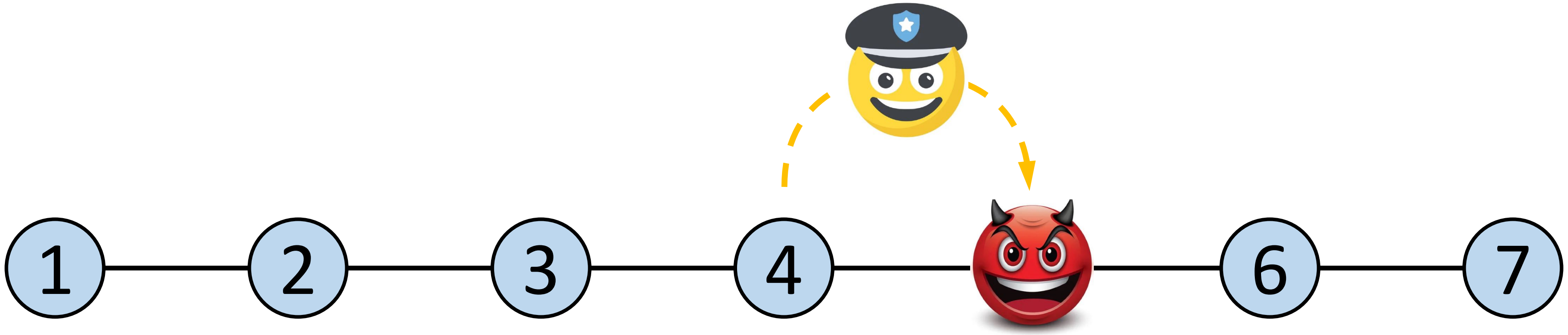
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

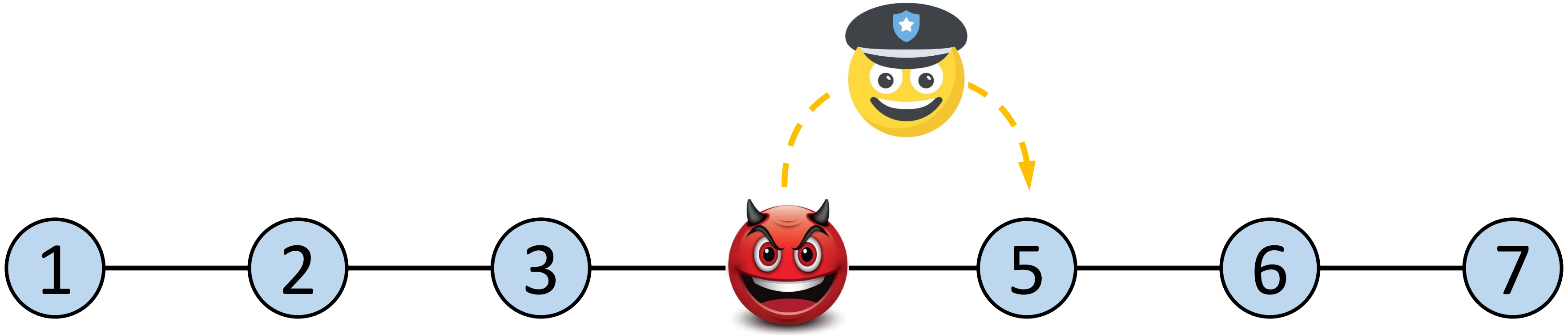
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

You can never catch the robber with only one cop ☹

Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

Treewidth with Cops and robber

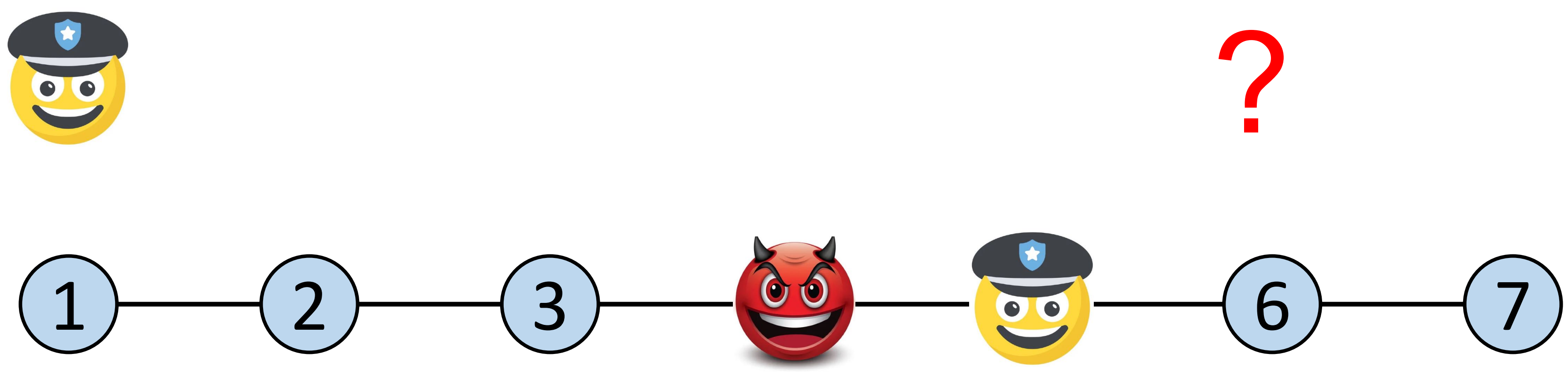
k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

What is the best move with a 2nd cop



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

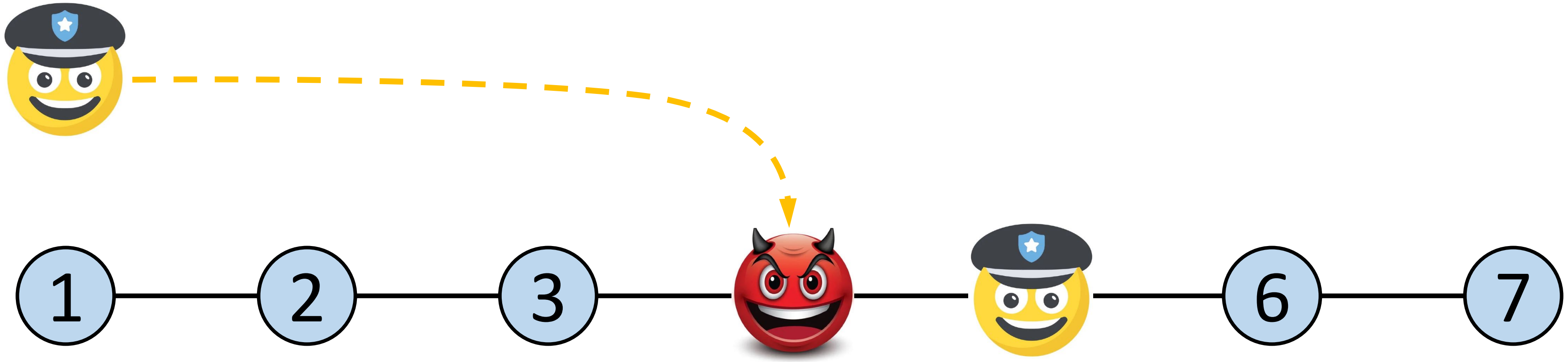
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

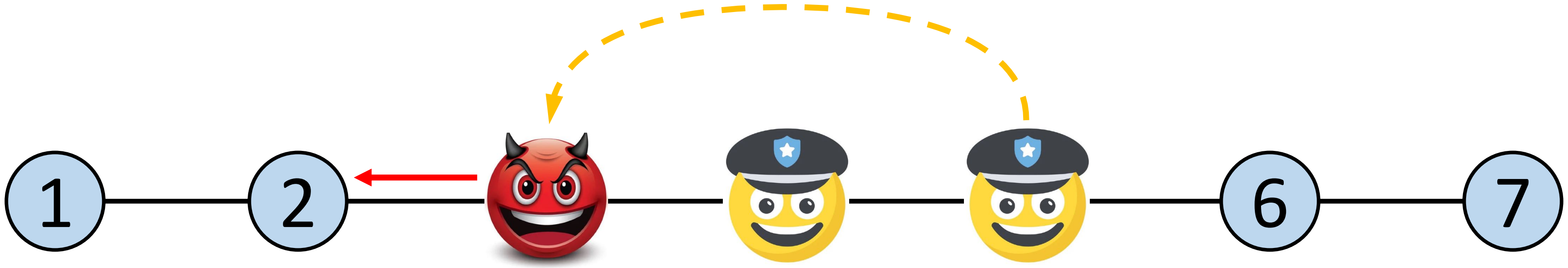
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.

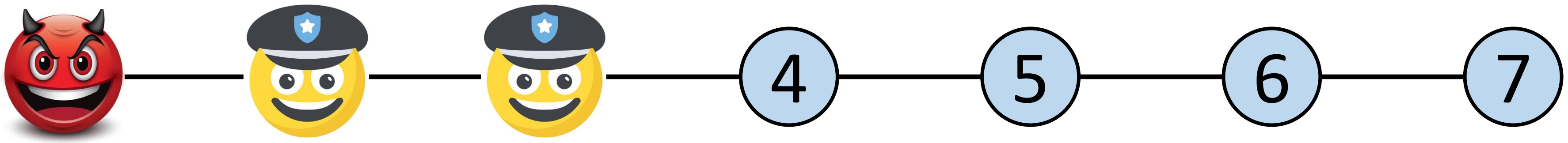
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



Seymour, Thomas. Graph searching and a min-max theorem for tree-width, Journal of Combinatorial Theory, Series B, 1993. https://doi.org/10.1006/jctb.1993.1027

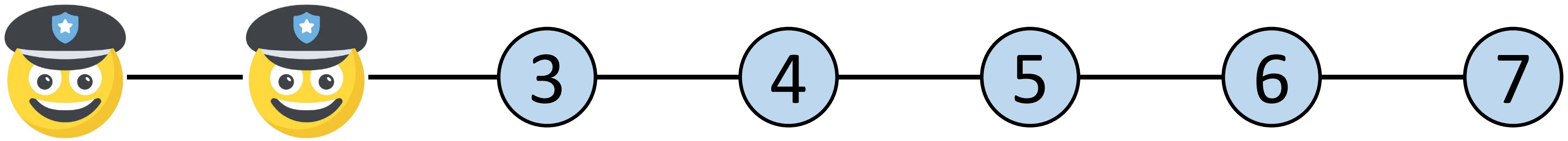
Treewidth with Cops and robber

k cops and 1 robber move on vertices of a graph. The robber can move quickly along paths that are not blocked by cops. Cops can fly via helicopters to new nodes. You control the cops and want to catch the robber (catch = occupy the same node). A single move consists of:

(1) A cop flies off the graph in a helicopter and announces a new landing vertex.

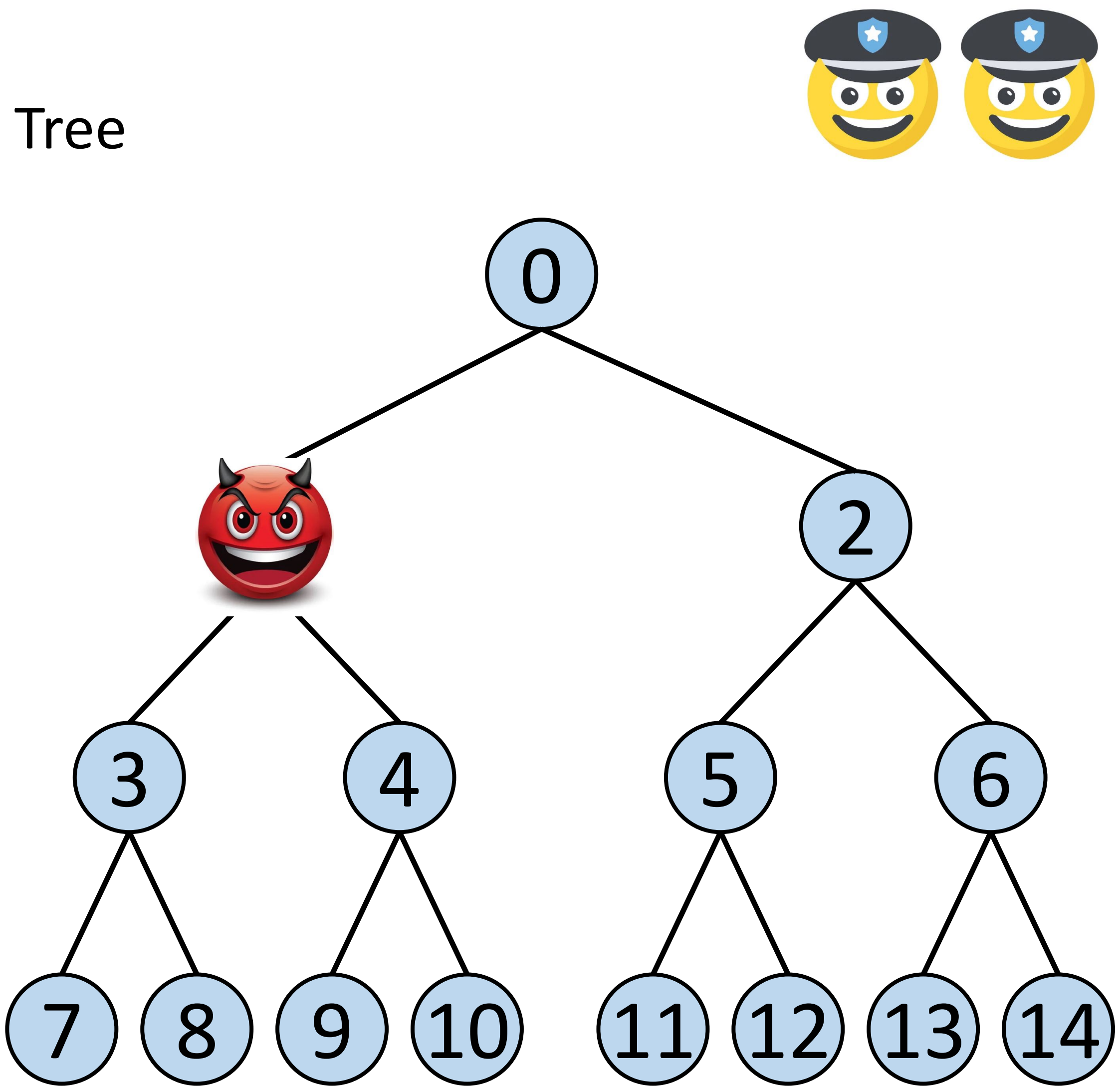
(2) While the cop flies, the robber can move quickly along the edges and escape.

(3) Then the cop lands.



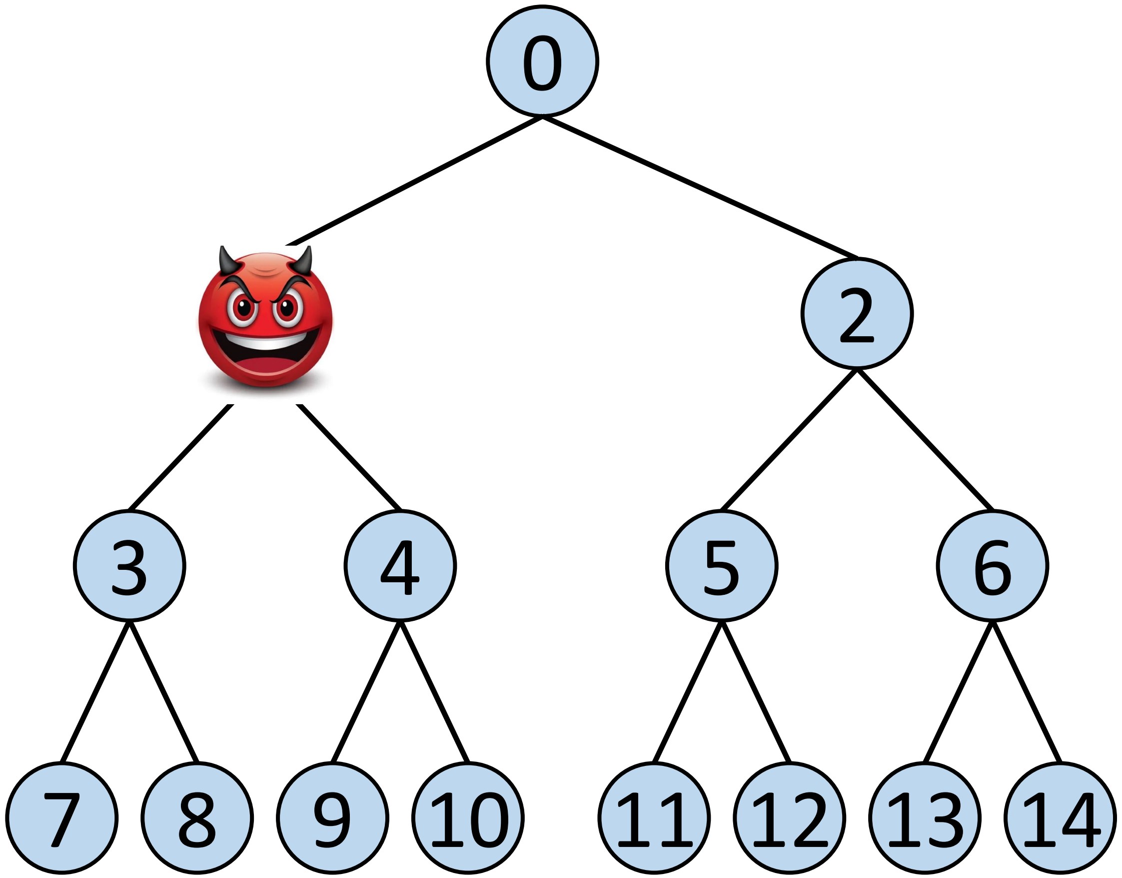
Robbers cannot hide on trees with 2 cops

Tree

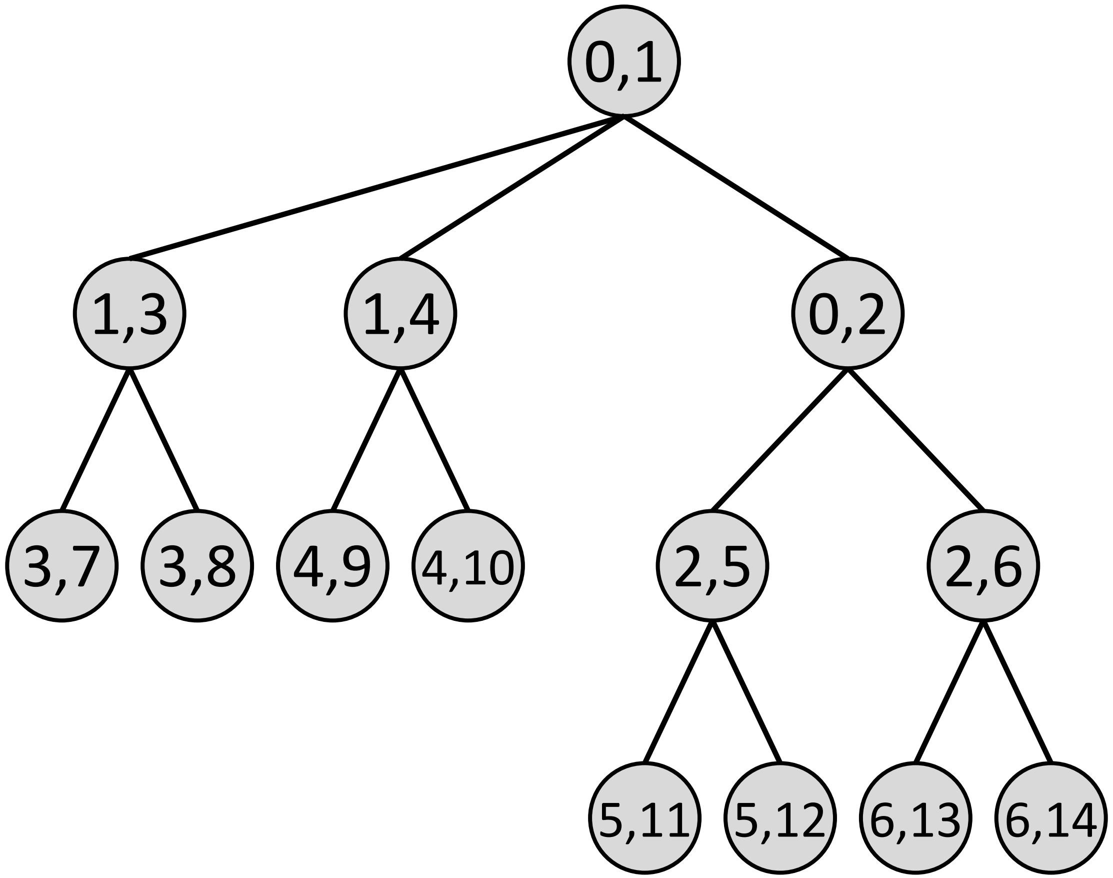


Robbers cannot hide on trees with 2 cops

Tree

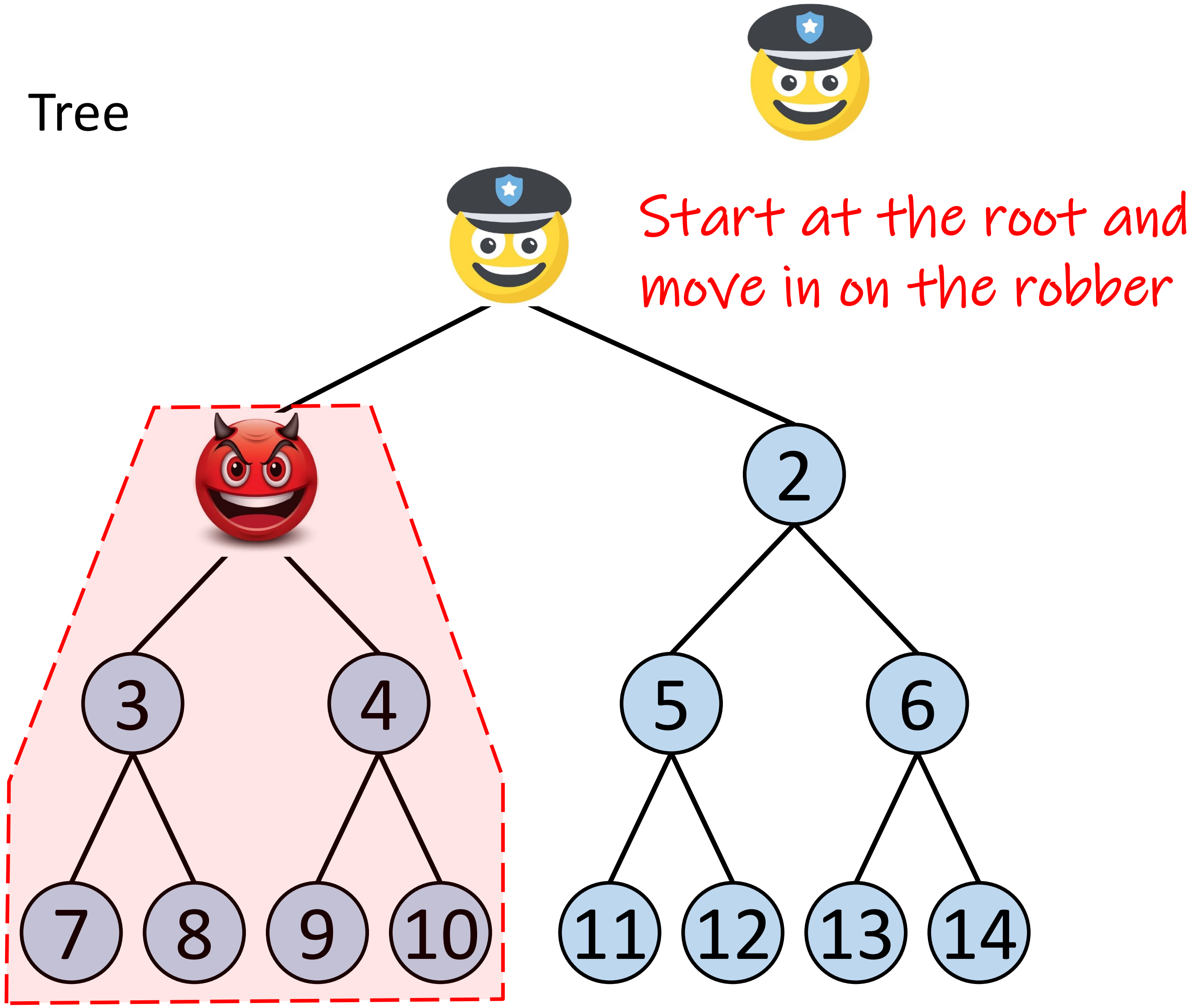


Tree decomposition

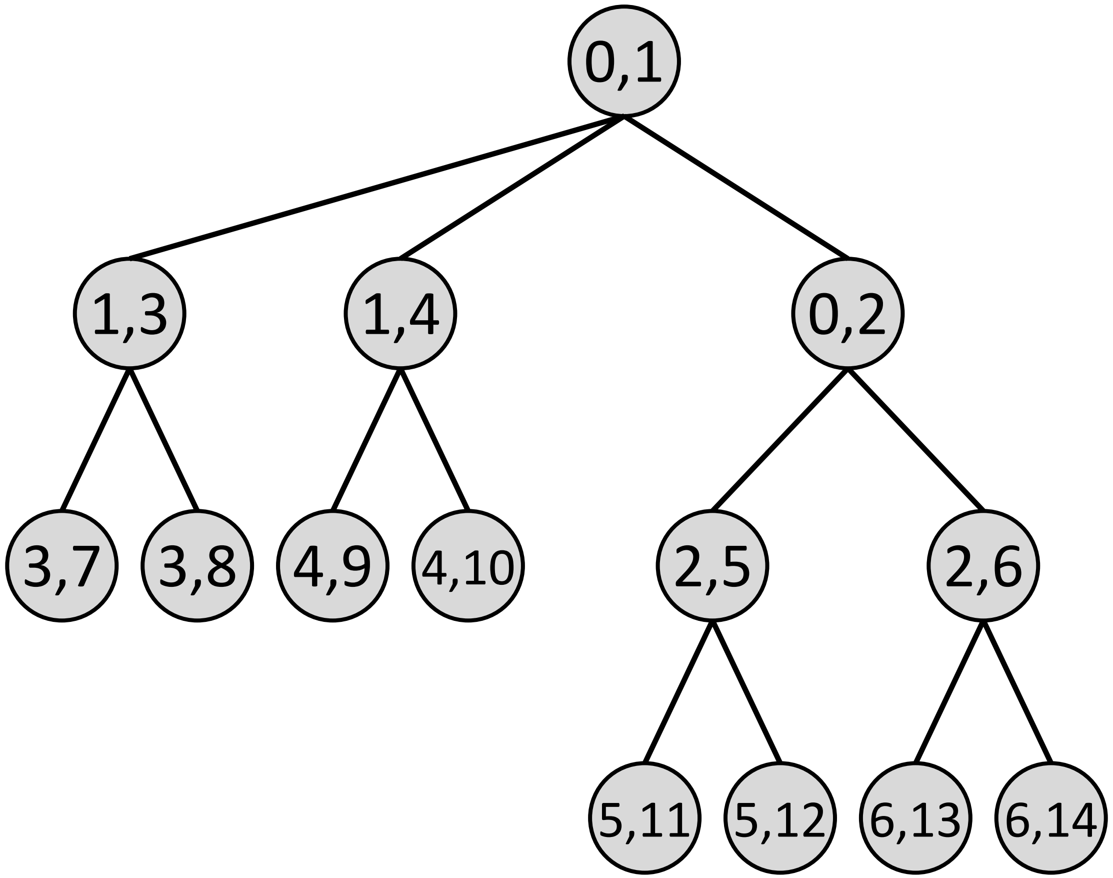


Robbers cannot hide on trees with 2 cops

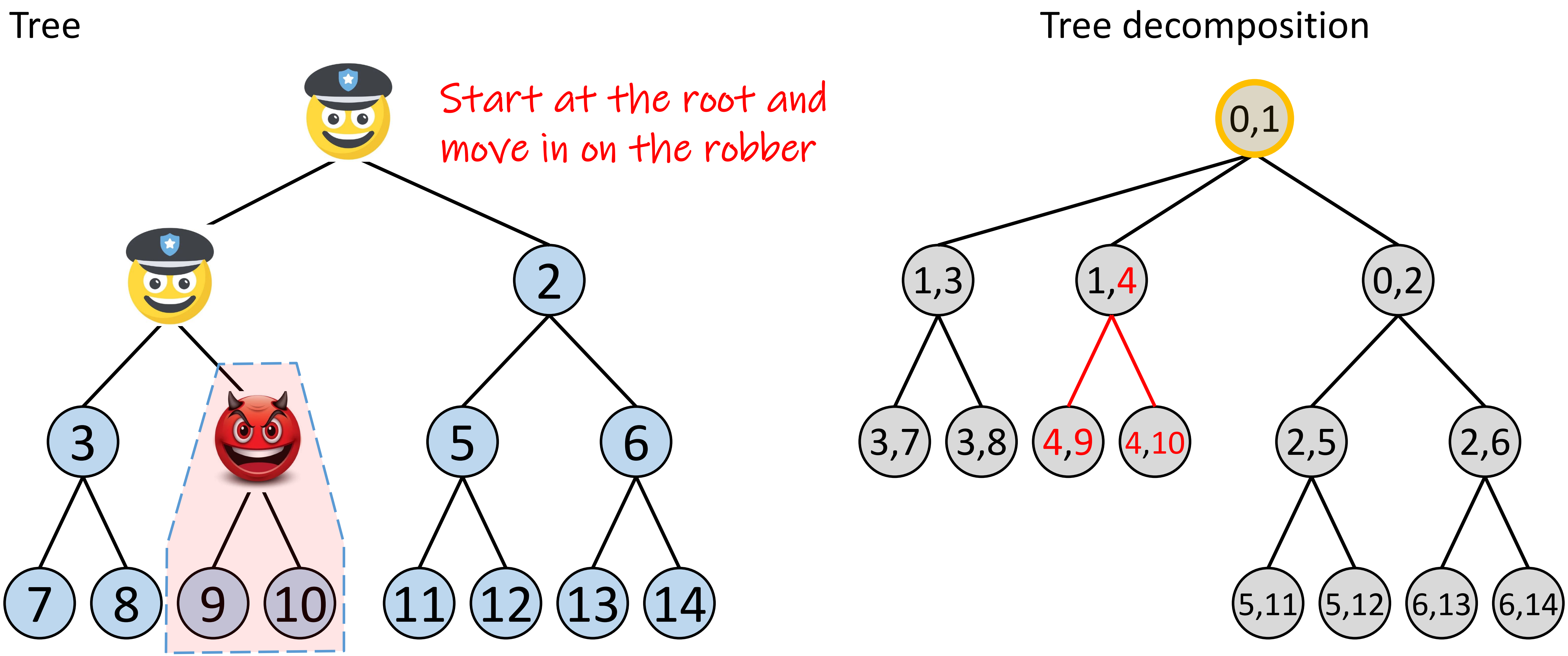
Tree



Tree decomposition

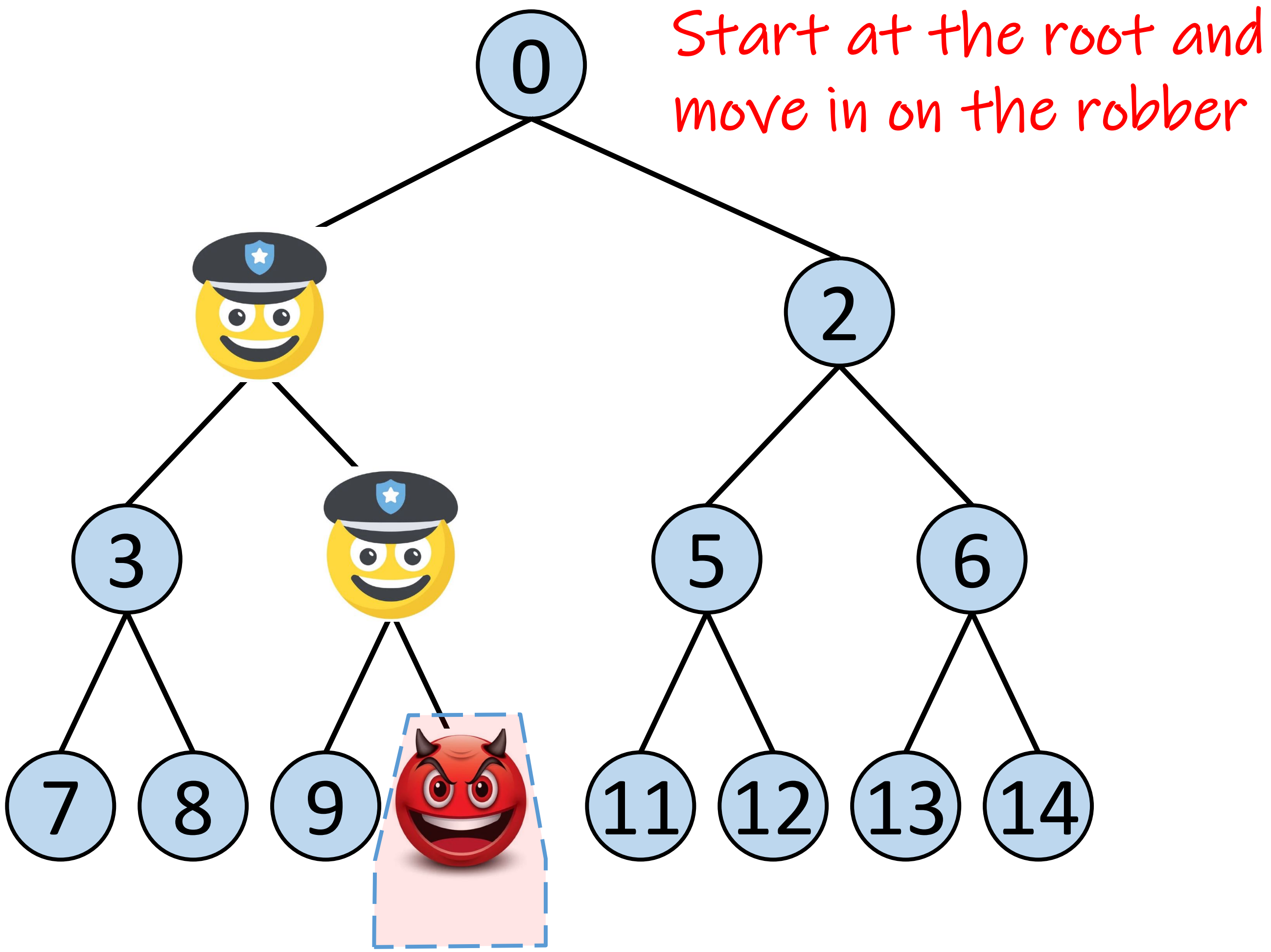


Robbers cannot hide on trees with 2 cops

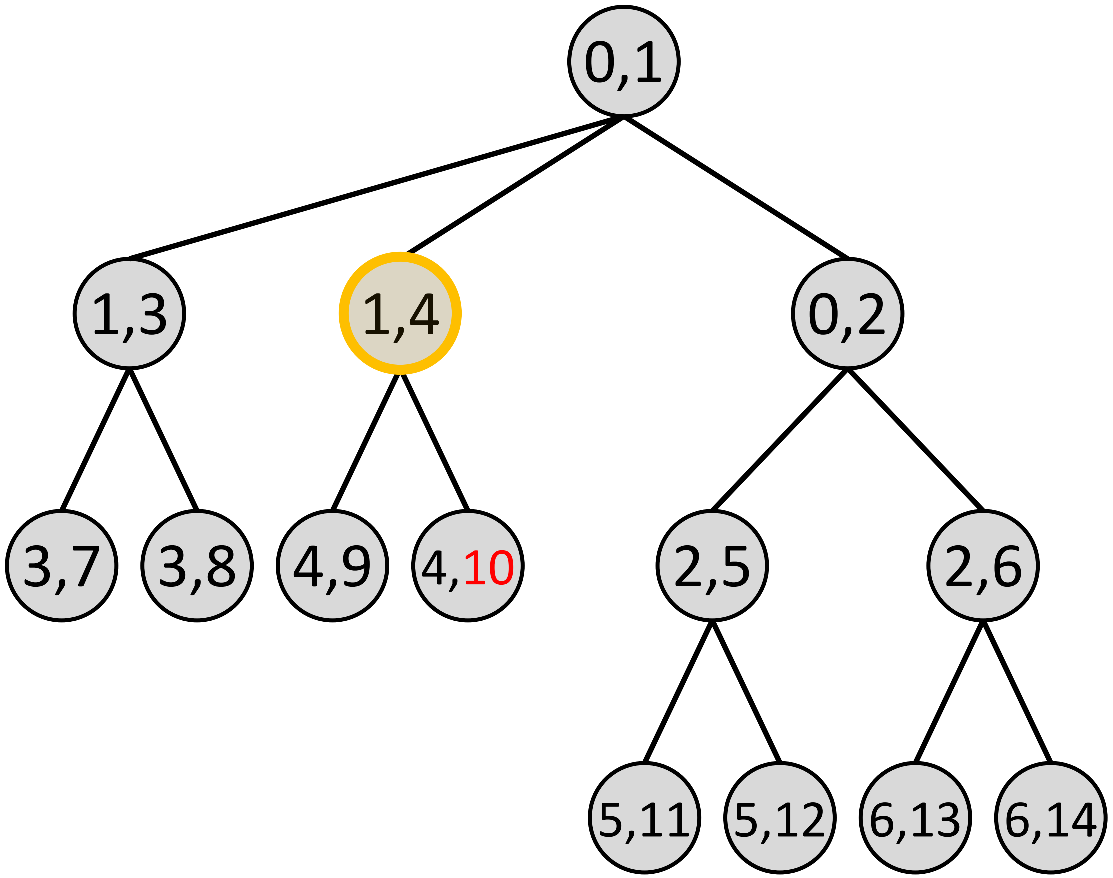


Robbers cannot hide on trees with 2 cops

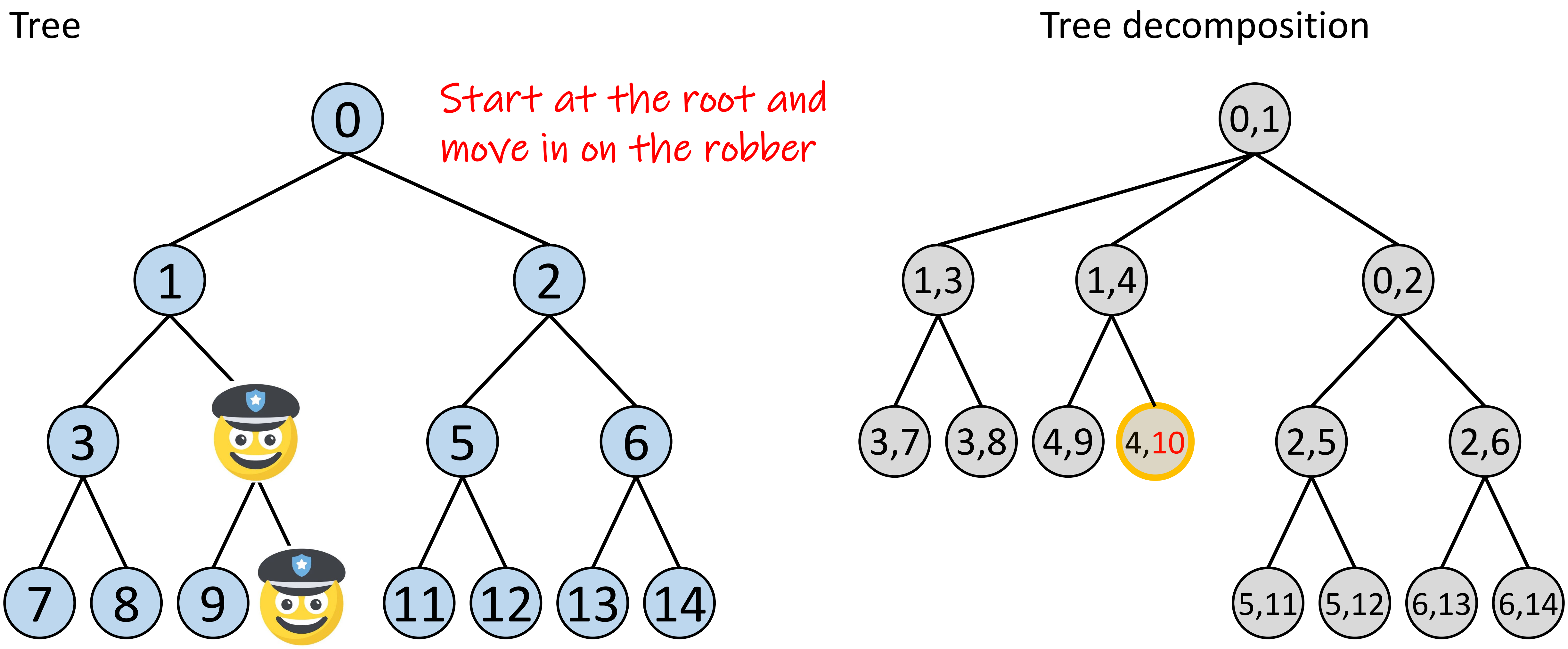
Tree



Tree decomposition

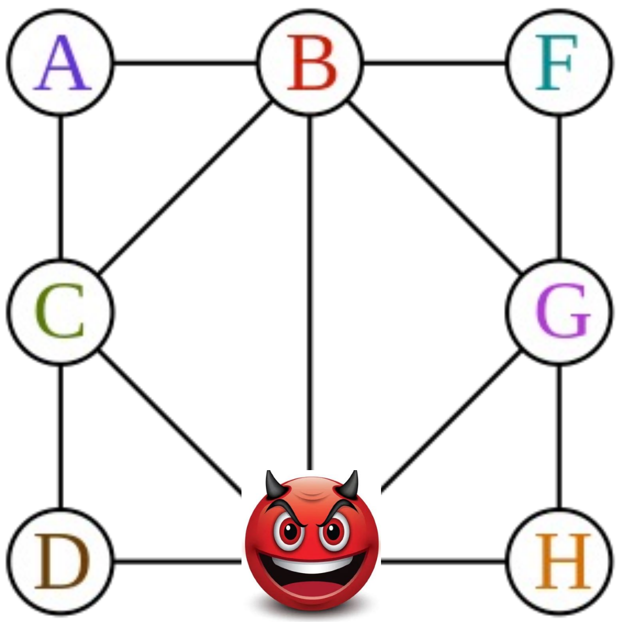


Robbers cannot hide on trees with 2 cops



Robbers cannot hide from k=3 cops on graph with treewidth=2

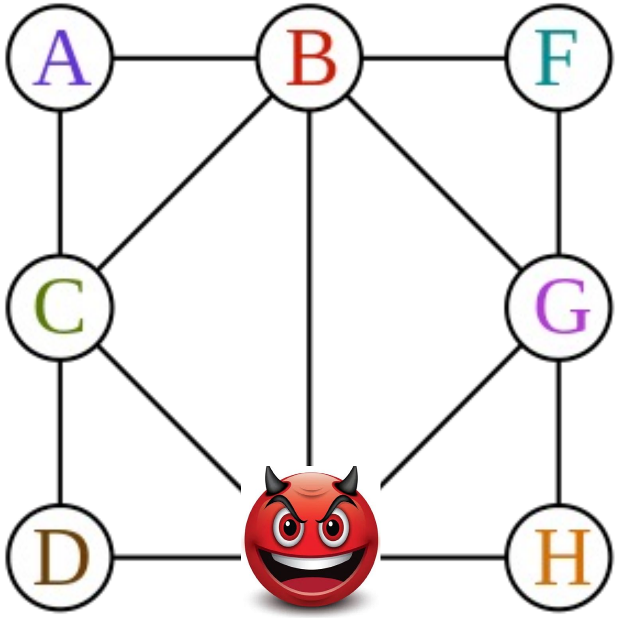
Graph with treewidth = 2



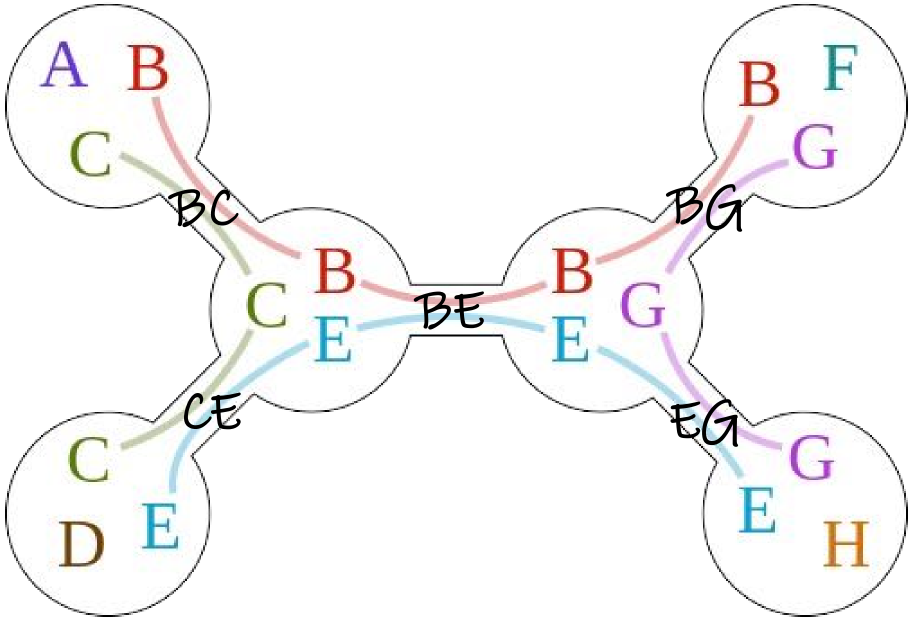
Robbers cannot hide from k=3 cops on graph with treewidth=2

Graph with treewidth = 2

You will need 3 cops



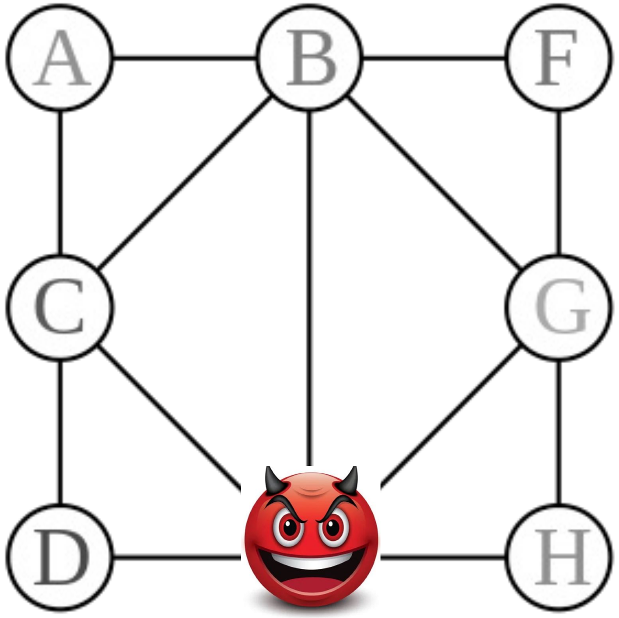
Tree decomposition



Robbers cannot hide from k=3 cops on graph with treewidth=2

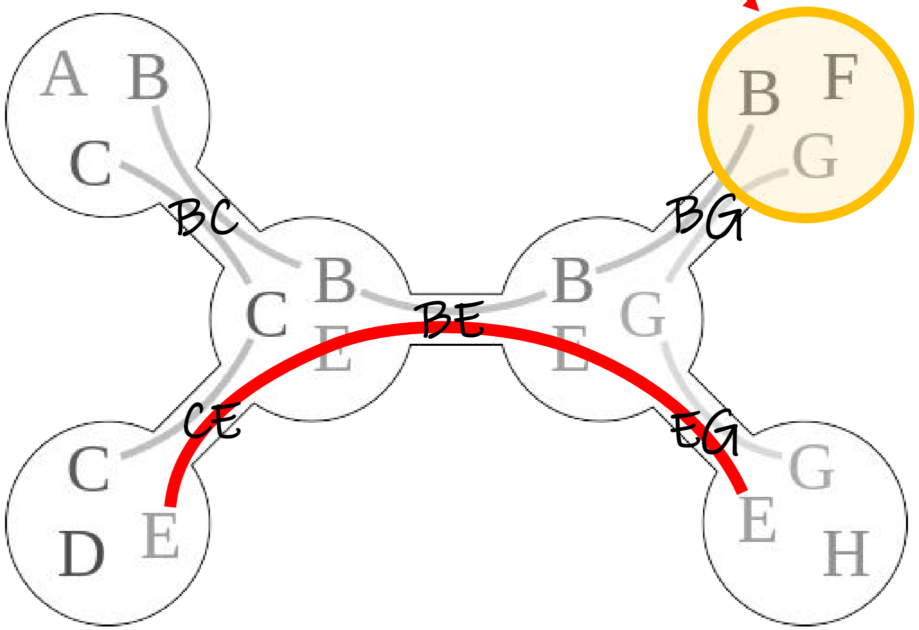
Graph with treewidth = 2

You will need 3 cops



Tree decomposition

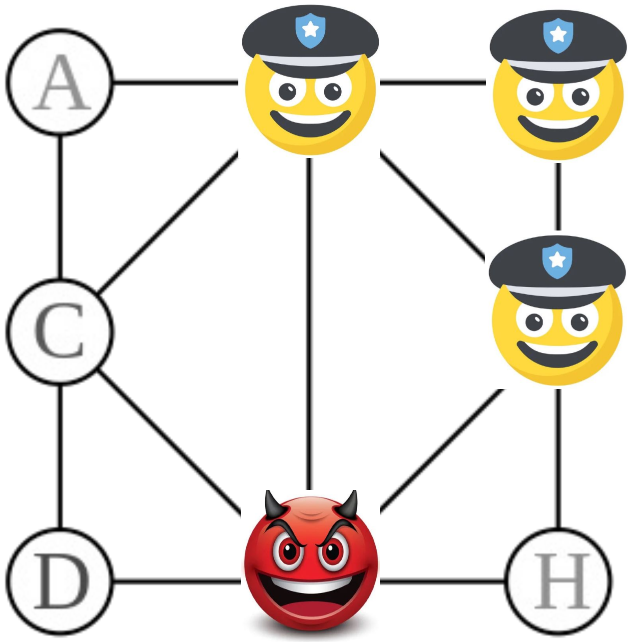
Pick some root



Robbers cannot hide from k=3 cops on graph with treewidth=2

Graph with treewidth = 2

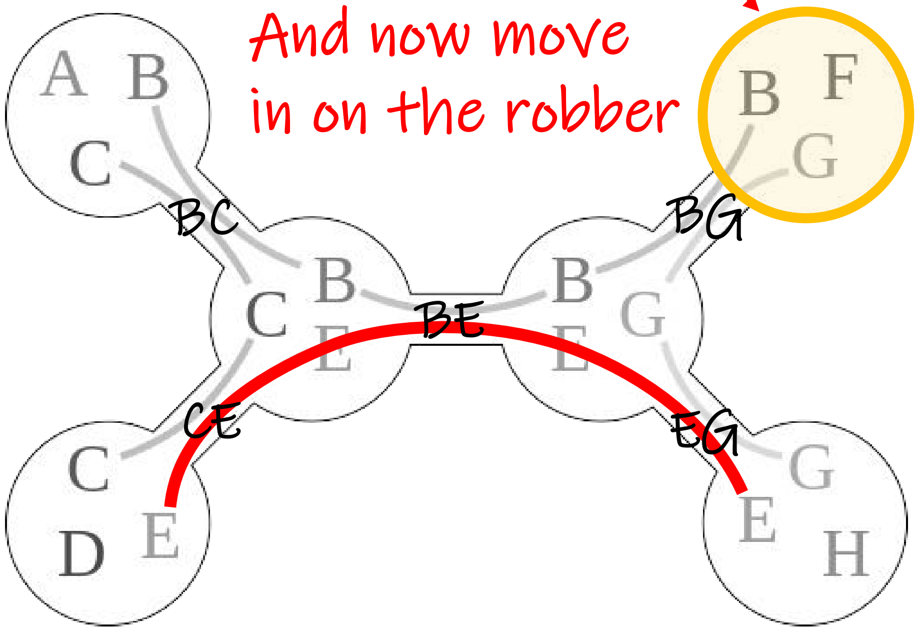
You will need 3 cops



Tree decomposition

Pick some root

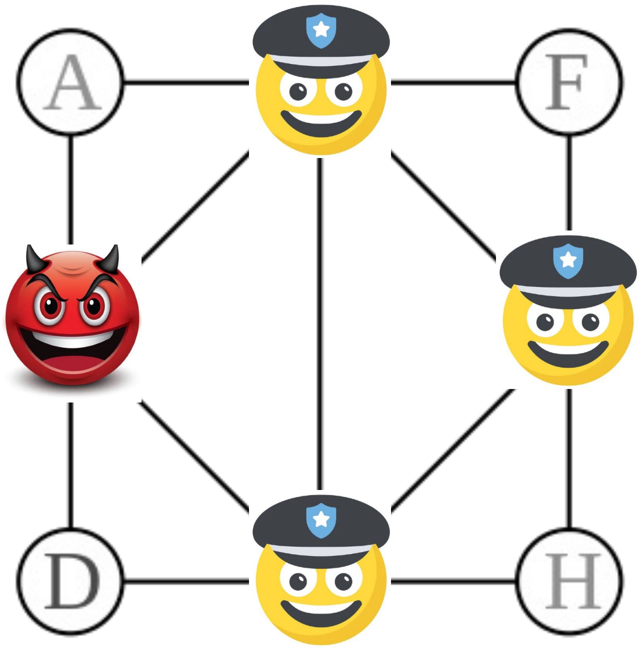
And now move
in on the robber



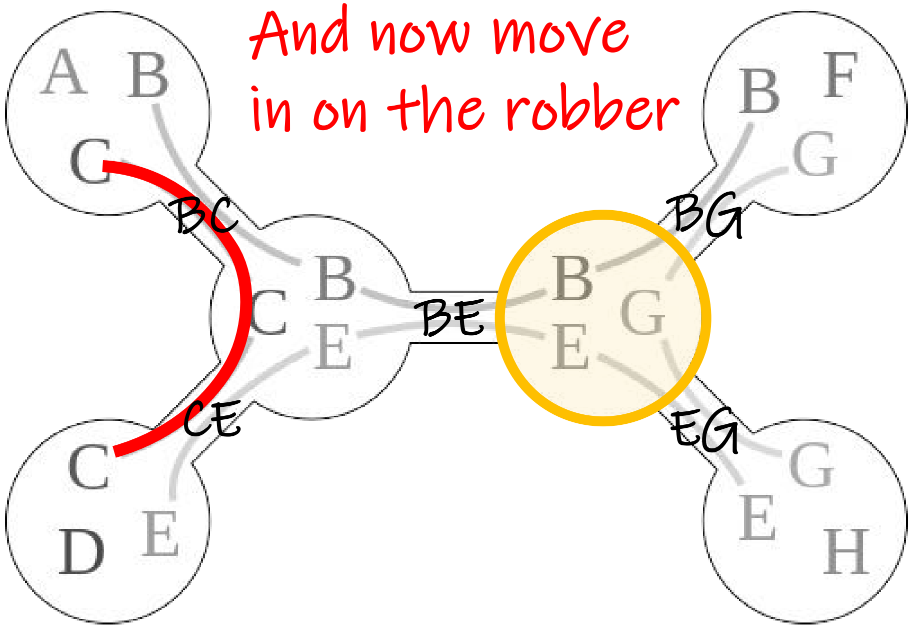
Robbers cannot hide from k=3 cops on graph with treewidth=2

Graph with treewidth = 2

You will need 3 cops



Tree decomposition

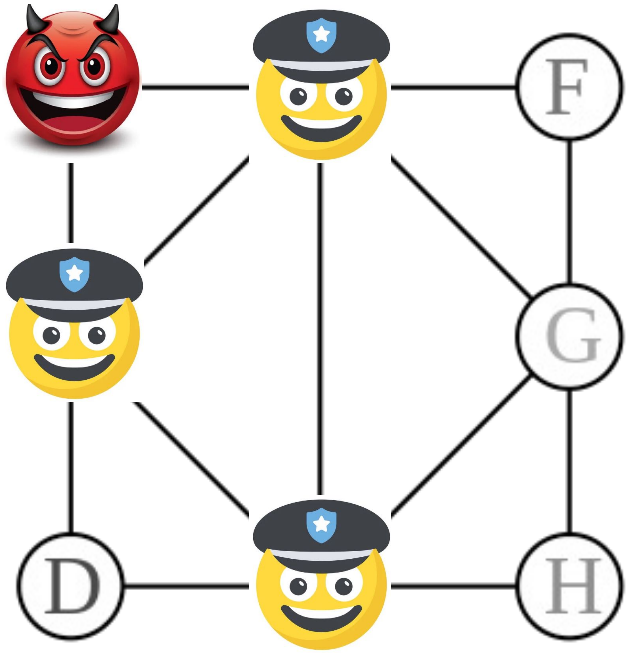


Graph and tree decomposition from: https://en.wikipedia.org/wiki/Tree_decomposition
Wolfgang Gatterbauer. Principles of scalable data management: https://northeastern-datalab.github.io/cs7240/

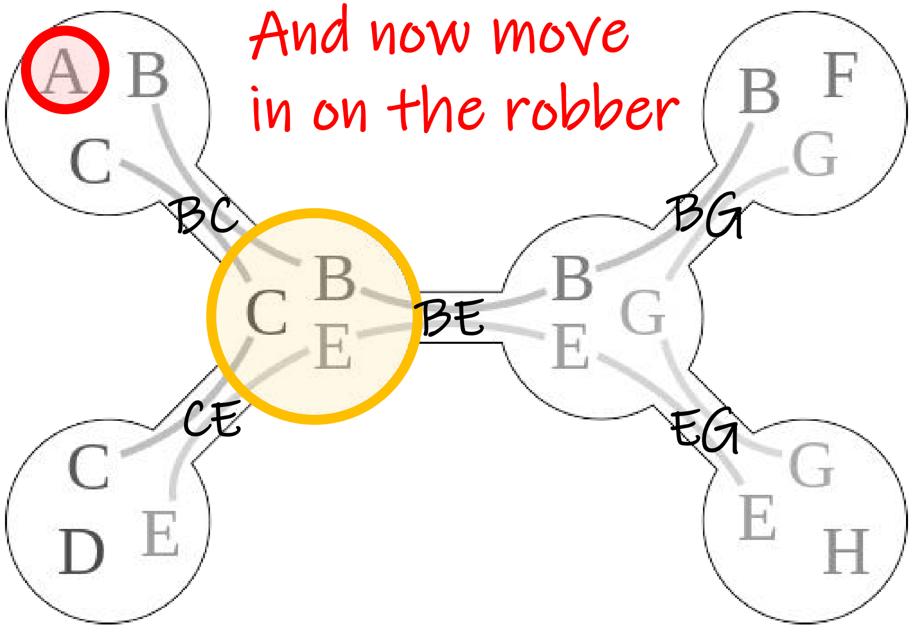
Robbers cannot hide from k=3 cops on graph with treewidth=2

Graph with treewidth = 2

You will need 3 cops



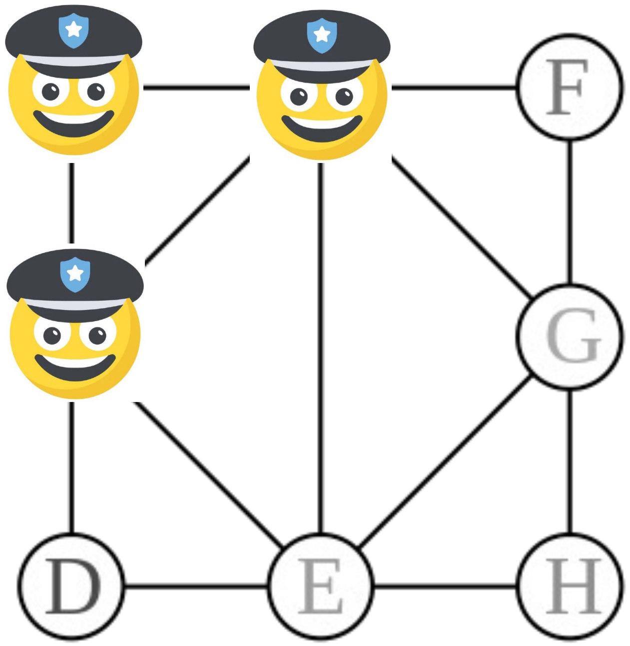
Tree decomposition



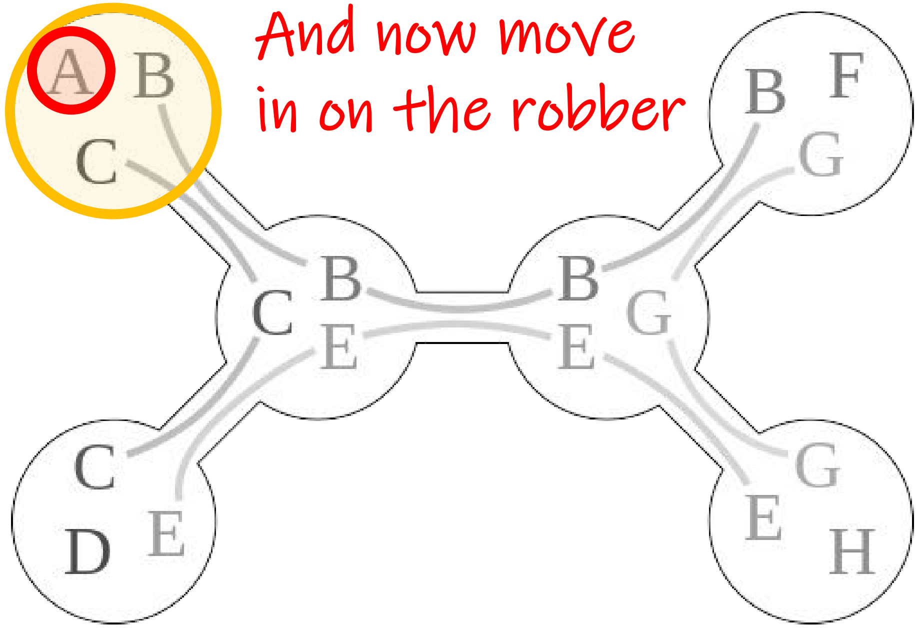
Robbers cannot hide from k=3 cops on graph with treewidth=2

Graph with treewidth = 2

You will need 3 cops
You caught the robber!



Tree decomposition



Updated 3/31/2023

Topic 3: Efficient query evaluation
Unit 2: Cyclic query evaluation
Lecture 22

Wolfgang Gatterbauer

CS7240 Principles of scalable data management (sp23)

https://northeastern-datalab.github.io/cs7240/sp23/

3/31/2023

Pre-class conversations

- Last class summary
- Scribes
  - Can you see my comments on your scribes and project notes?
  - also posting scribes on Piazza
- Project: (P3: today FRI, 3/31)
- Feedback on my slides
- Today:
  - Reducing cycles to trees (tree decompositions)
  - Reducing cycles in CQs to trees based on the domain or based on atoms (treewidth, query width hypertree decompositions)
  - Linear Programming Duality

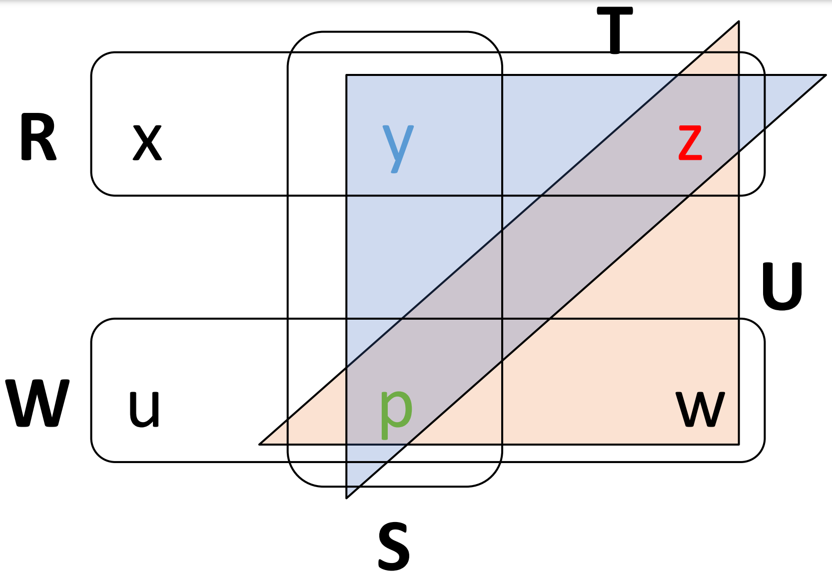
Outline: T3-2: Cyclic conjunctive queries

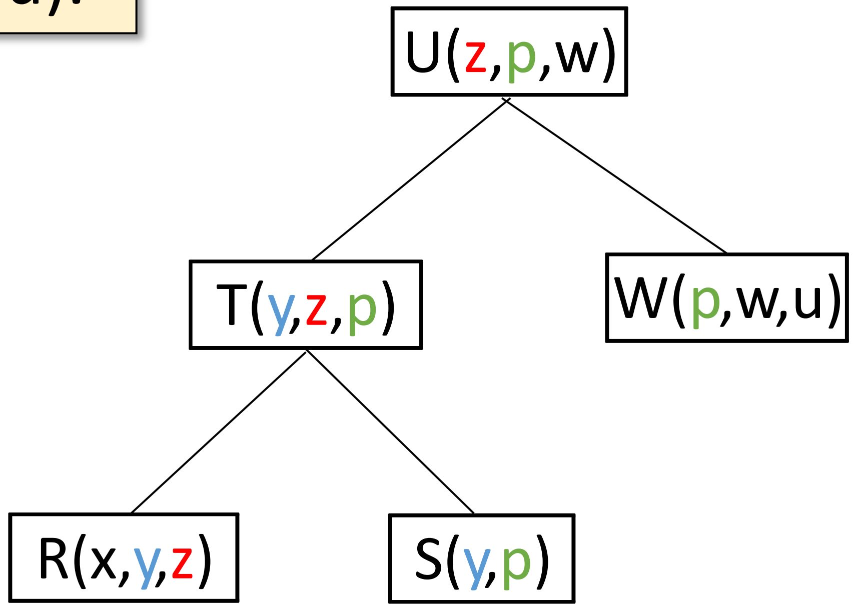
- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

Acyclic Conjunctive Queries

- A join tree for a hypergraph H=(V,E) is a labeled tree T =(N,F,λ) such that:
  - The nodes of T are formed by the hyperedges. In other words, λ: N→E s.t. for each hyperedge e ∈ E of H, there exists n ∈ N such that e = λ(n)
  - For each node u ∈ V of H, the set {n ∈ N | u ∈ λ(n)} induces a connected subtree of T. (also called: running intersection property)

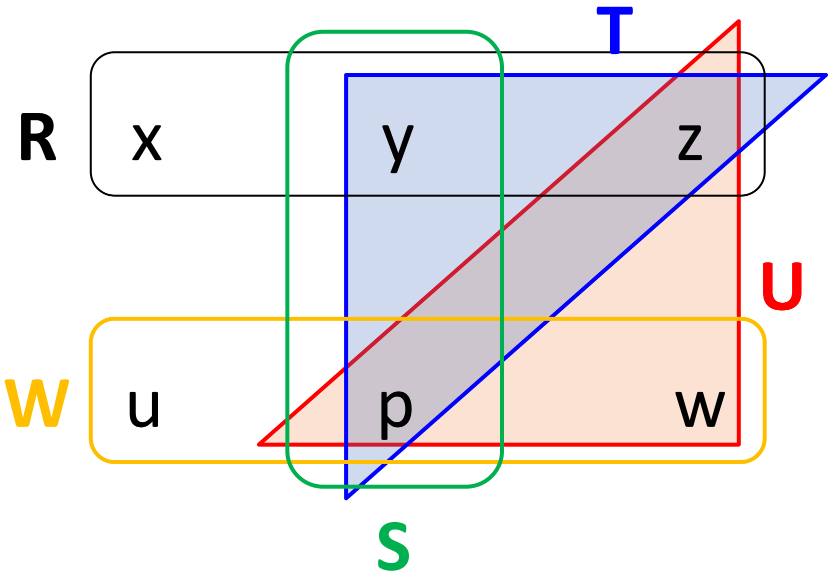
Q :- R(x,y,z), S(y,p), T(y,z,p), U(z,p,w), W(p,w,u).

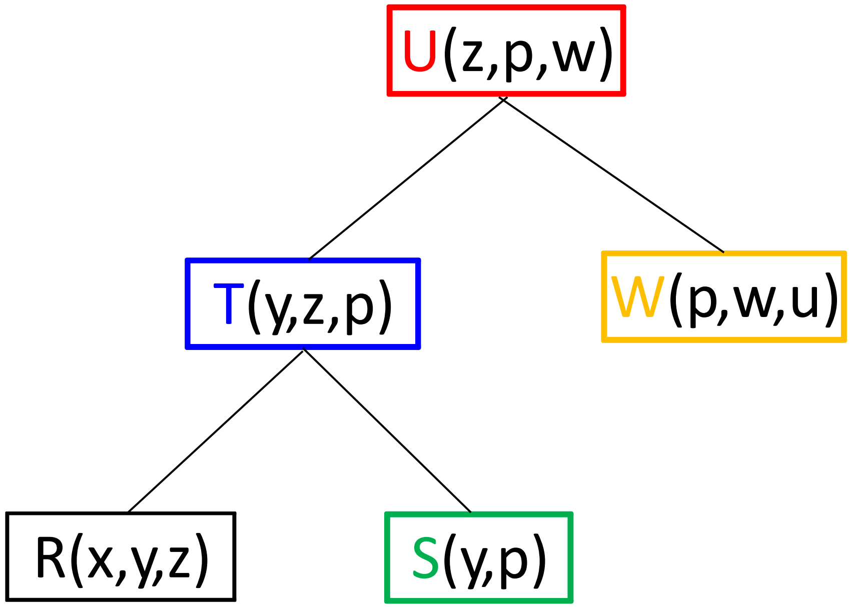




Acyclic Conjunctive Queries

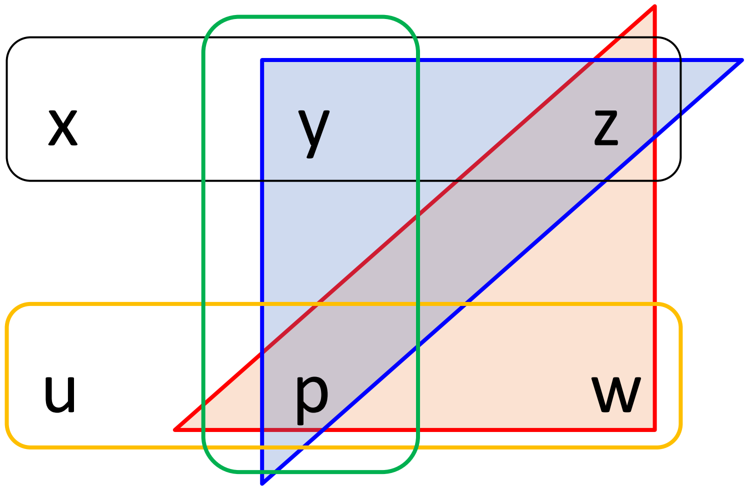
- A join tree for a hypergraph H=(V,E) is a labeled tree T =(N,F,λ) such that:
  - The nodes of T are formed by the hyperedges. In other words, λ: N→E s.t. for each hyperedge e ∈ E of H, there exists n ∈ N such that e = λ(n)
  - For each node u ∈ V of H, the set {n ∈ N | u ∈ λ(n)} induces a connected subtree of T. (also called: running intersection property)

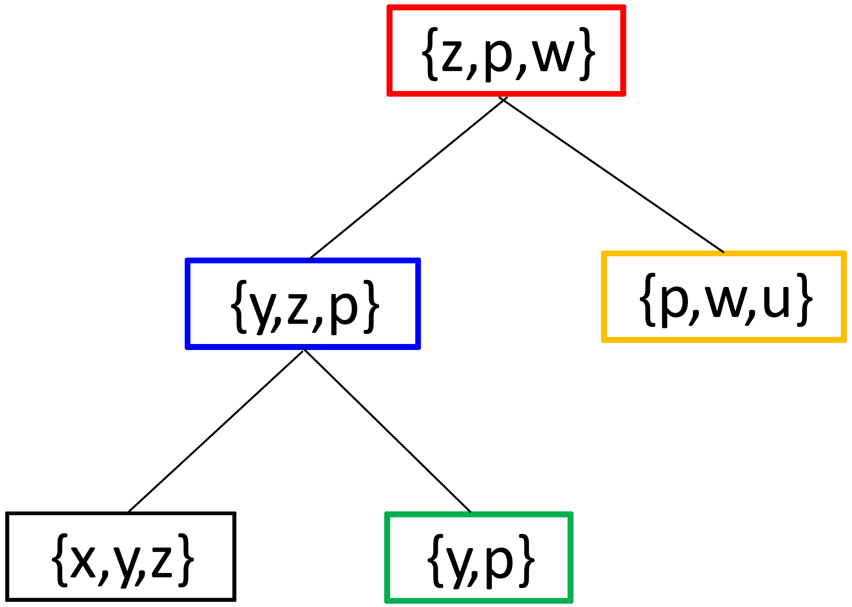




Acyclic Conjunctive Queries

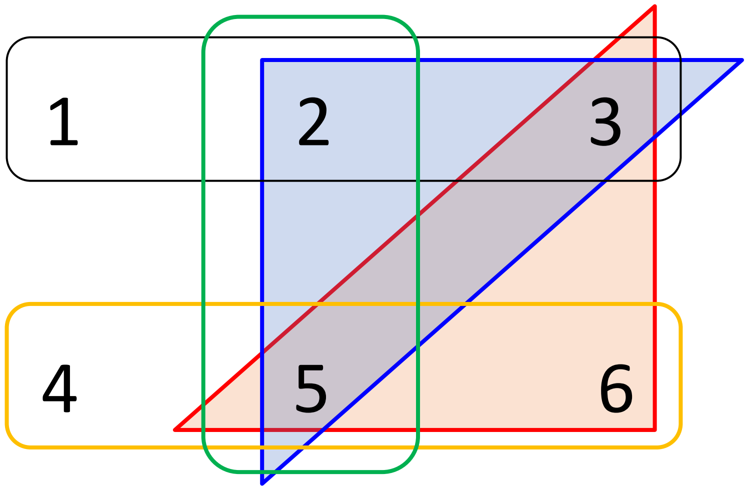
- A join tree for a hypergraph H=(V,E) is a labeled tree T =(N,F,λ) such that:
  - The nodes of T are formed by the hyperedges. In other words, λ: N→E s.t. for each hyperedge e ∈ E of H, there exists n ∈ N such that e = λ(n)
  - For each node u ∈ V of H, the set {n ∈ N | u ∈ λ(n)} induces a connected subtree of T. (also called: running intersection property)

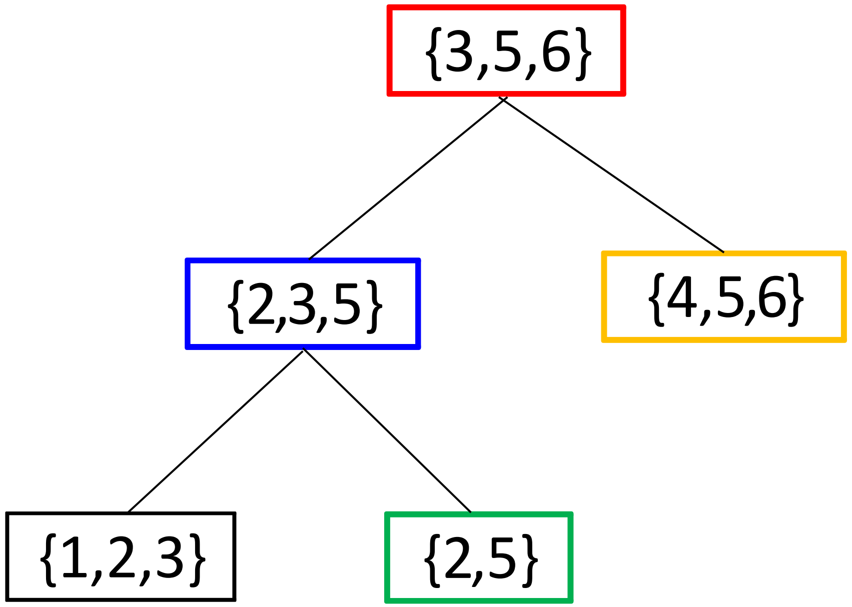




Acyclic Conjunctive Queries

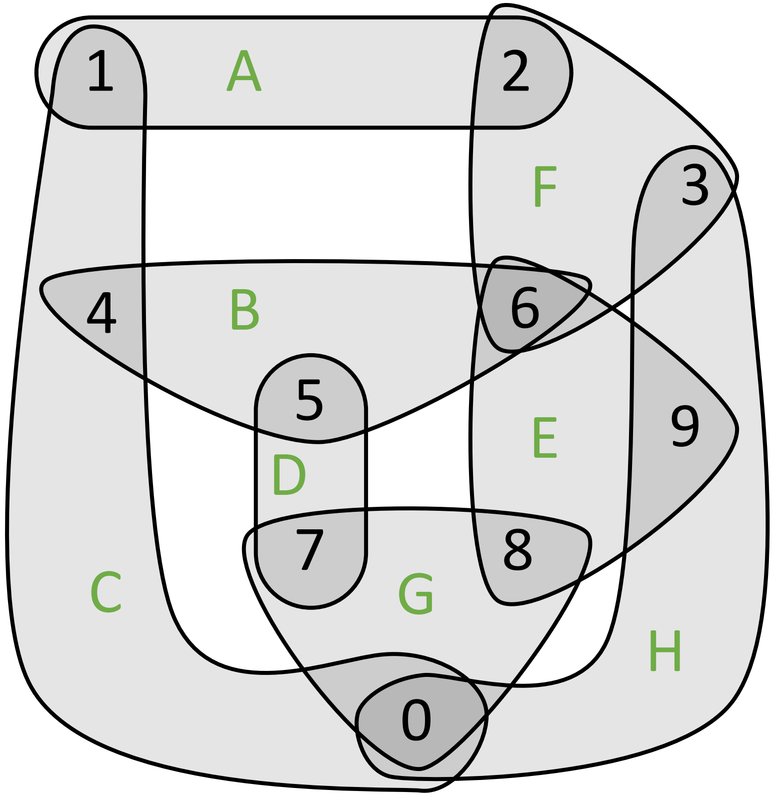
- A join tree for a hypergraph H=(V,E) is a labeled tree T =(N,F,λ) such that:
  - The nodes of T are formed by the hyperedges. In other words, λ: N→E s.t. for each hyperedge e ∈ E of H, there exists n ∈ N such that e = λ(n)
  - For each node u ∈ V of H, the set {n ∈ N | u ∈ λ(n)} induces a connected subtree of T. (also called: running intersection property)





Cyclic Conjunctive Queries

Hypergraph



For queries that are not acyclic, what bounds can we give on the data complexity of query evaluation, considering various structural properties of the query?

We will see:

- Coherence (as in TDs) is still a key structural criterion for efficiency!
- But treewidth does not generalize the notion of hypergraph acyclicity (because acyclic families of hypergraphs may have unbounded treewidth ☹)
- What will help is the number of atoms needed to cover sets of variables ☺.
- Reason: size of database is determined by number of tuples n not domain size m

Example adopted from: Markus Krötzsch. "Database theory: Lecture 6: Tree-like Conjunctive Queries." 2016. https://iccl.inf.tu-dresden.de/web/Database_Theory_(SS2016)/en

Issues with standard Treewidth (TW) for CQs

Treewidth based on graphs.
TW of CQ is TW of its clique graph (i.e. replace each hyperedge with a clique)

a clique is a graph where where every vertex is connected to every other vertex

Q(x,y,z,w) :- R(x,y,z,w).

Hypergraph

?

Clique graph

?

Treewidth: ?

Issues with standard Treewidth (TW) for CQs

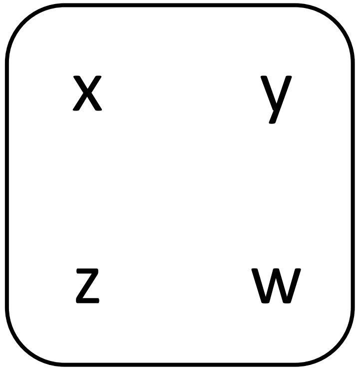
Treewidth based on graphs.

TW of CQ is TW of its clique graph (i.e. replace each hyperedge with a clique)

a clique is a graph where where every vertex is connected to every other vertex

Q(x,y,z,w) :- R(x,y,z,w).

Hypergraph



Clique graph

?

Treewidth: ?

Issues with standard Treewidth (TW) for CQs

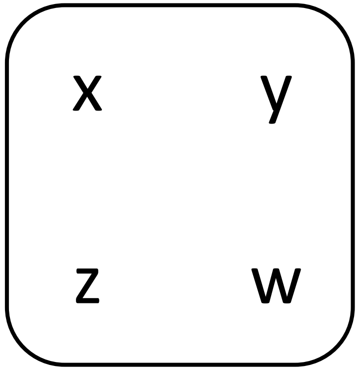
Treewidth based on graphs.

TW of CQ is TW of its clique graph (i.e. replace each hyperedge with a clique)

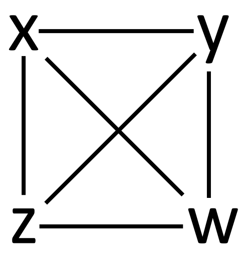
a clique is a graph where where every vertex is connected to every other vertex

Q(x,y,z,w) :- R(x,y,z,w).

Hypergraph



Clique graph



Treewidth: ?

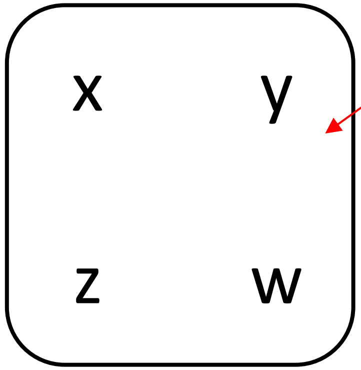
Issues with standard Treewidth (TW) for CQs

Treewidth based on graphs.
TW of CQ is TW of its clique graph (i.e. replace each hyperedge with a clique)

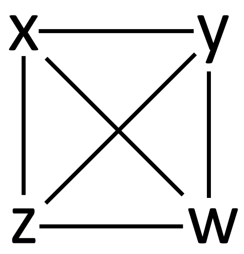
Q(x,y,z,w) :- R(x,y,z,w).

This is actually the best tree decomposition: Nodes of a clique need to appear in the same supernode

Hypertree



Clique graph



Treewidth: 3

Resulting complexity bound $O(m^4)$!

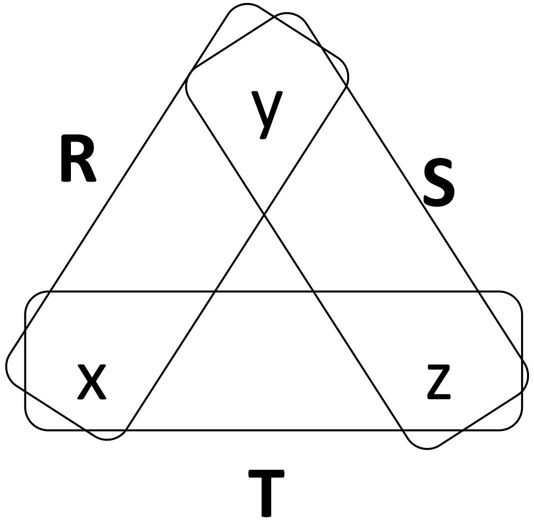
That's a pretty bad bound. We know we can evaluate this query in $O(n)$.

Issues with standard Treewidth (TW) for CQs

$Q_1(x,y,z)$:- $R(x,y)$, $S(y,z)$, $T(x,z)$.
$Q_2(x,y,z)$:- $R(x,y)$, $S(y,z)$, $T(x,z)$, $W(x,y,z)$.

We also know that these two queries have different maximal output sizes: $O(n^{1.5})$ vs. $O(n)$.
But TW cannot distinguish them ☹

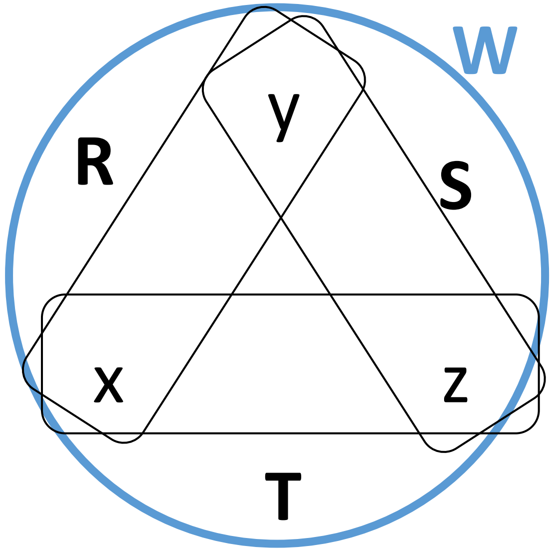
H_1



Clique graph

?

H_2

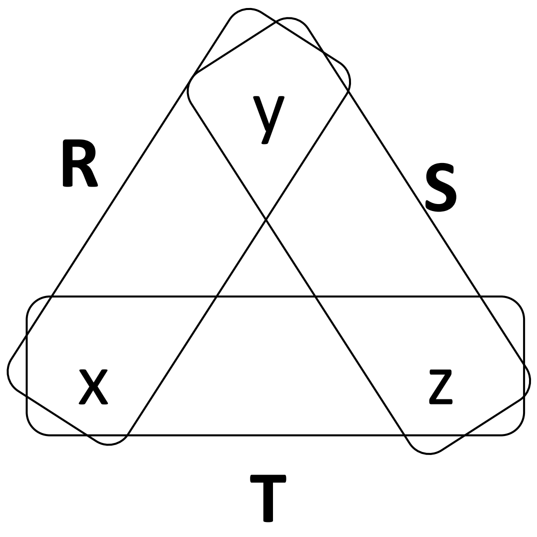


Issues with standard Treewidth (TW) for CQs

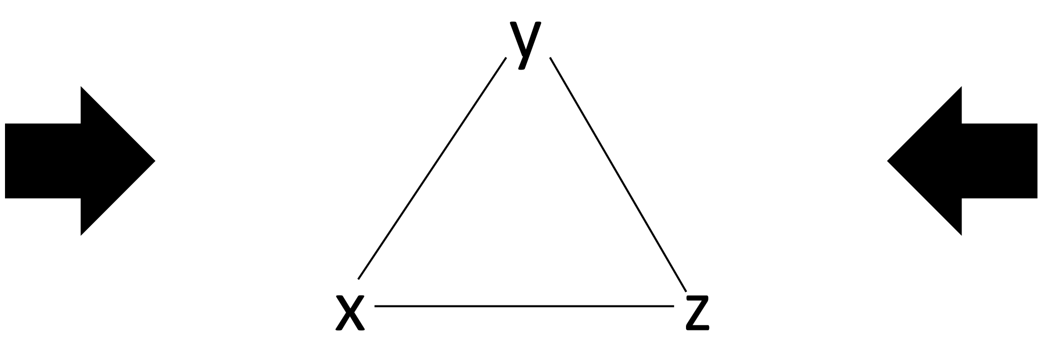
$Q_1(x,y,z)$:- $R(x,y)$, $S(y,z)$, $T(x,z)$.
$Q_2(x,y,z)$:- $R(x,y)$, $S(y,z)$, $T(x,z)$, $W(x,y,z)$.

We also know that these two queries have different maximal output sizes: $O(n^{1.5})$ vs. $O(n)$.
But TW cannot distinguish them ☹

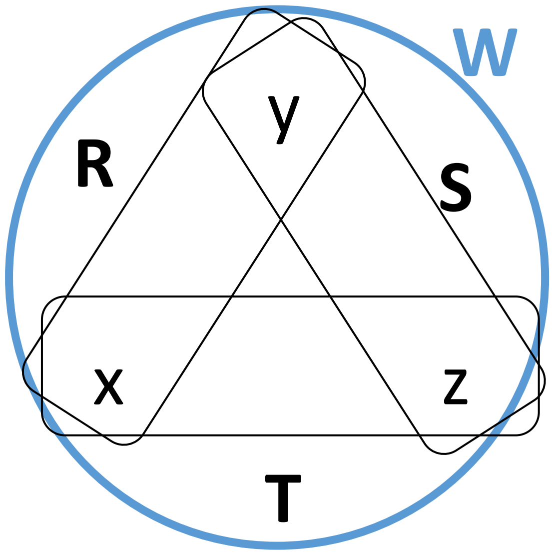
H$_1$



Clique graph



H$_2$



Same clique graph. Therefore:
→ same TW 2.
→ same complexity bound $O(m^3)$

"Query decomposition" [Chekuri, Rajaraman'97]

QUERY DECOMPOSITION

Tree decomposition with coherence conditions on both: 1) variables and 2) atoms.

Query width: max # of atoms in a supernode

A *query decomposition* of Q is a tree $T=(I,F)$, with a set $X(i)$ of subgoals and arguments associated with each vertex $i \in I$, such that the following conditions are satisfied:

- For each subgoal s of Q, there is an $i \in I$ such that $s \in X(i)$.
- For each subgoal s of Q, the set $\{i \in I \mid s \in X(i)\}$ induces a (connected) subtree of T.
- For each argument A of Q, the set

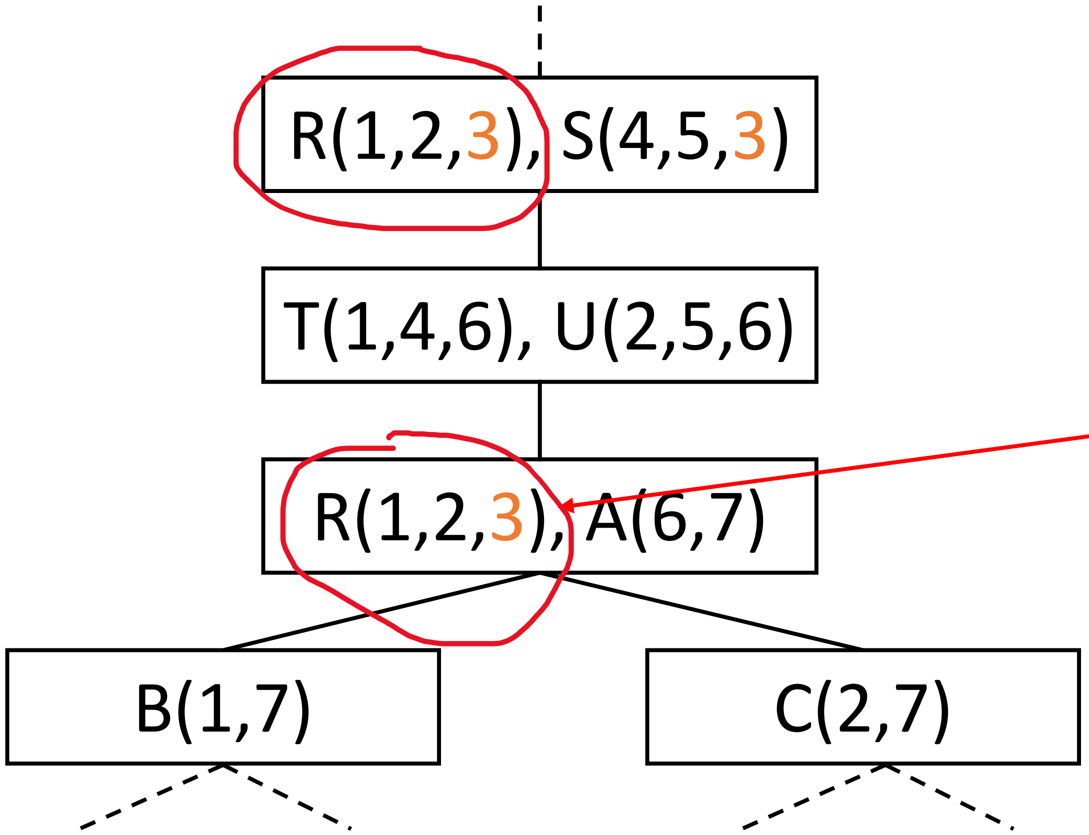
$$\{i \in I \mid A \in X(i)\} \cup \{i \in I \mid A \text{ appears in a subgoal } s \text{ such that } s \in X(i)\}$$

induces a (connected) subtree of T.

The *width* of the query decomposition is $\max_{i \in I} |X(i)|$. The *query width* of Q is the minimum width over all its query decompositions.

Important Observation 1

Some decomposition



"Query decomposition" as defined by [Chekuri, Rajaraman'97] is too strict about atoms needing to be connected and atoms not allowing projections

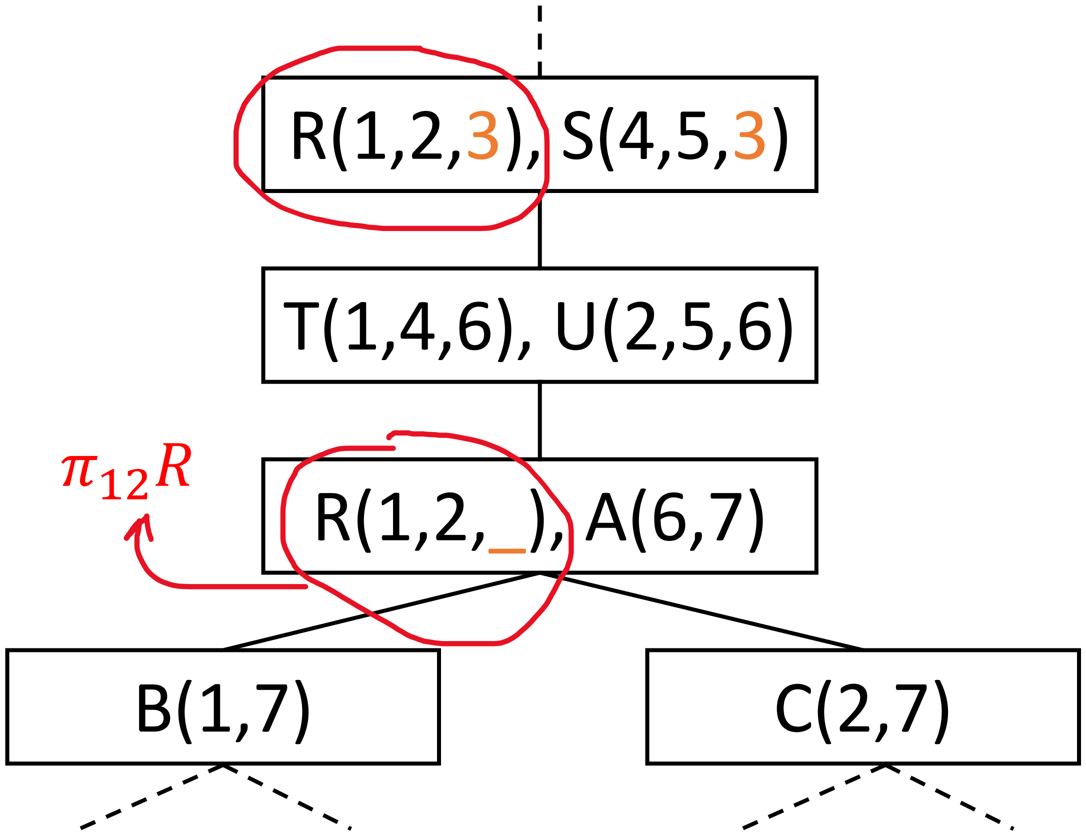
This decomposition would not be possible for original "query decomposition" because "3" is not connected.

But what if you project "3" away onto $R(1,2) = \pi_{12} R(1,2,3)$

Adopted from an example by Georg Gottlob

Important Observation 1

Some decomposition



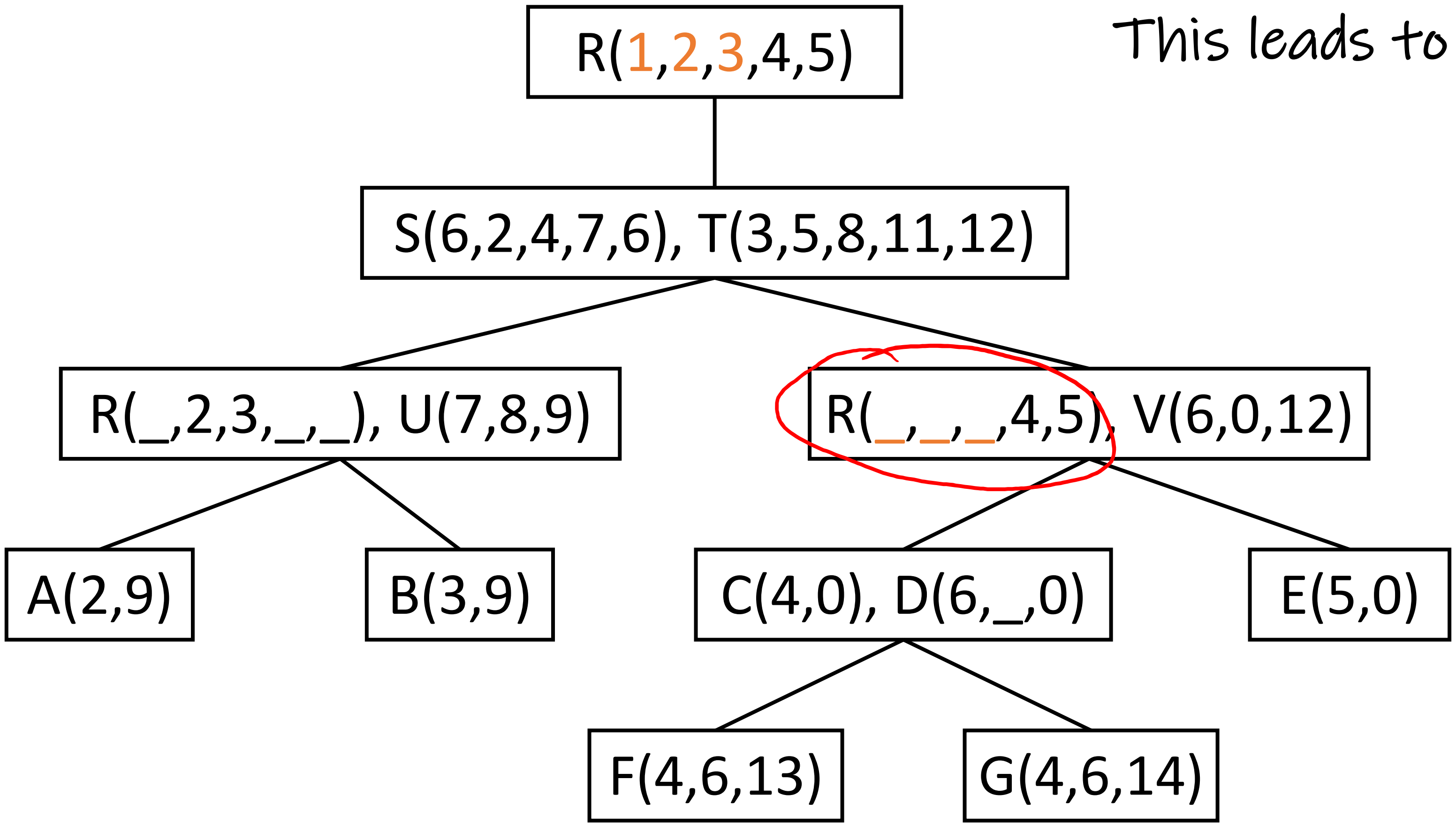
Here the reuse of R(1,2,3) is harmless: we could have added an atom R(1,2,_) here without changing the query.

Idea: allow query atoms to be reused partially (with projections) as long as the full atom appears somewhere else.

This leads to "generalized hypertree decompositions" which define coherence only based on variables, not atoms. More liberal than "query decomposition", and thus can give tighter bounds.

Important Observation 2

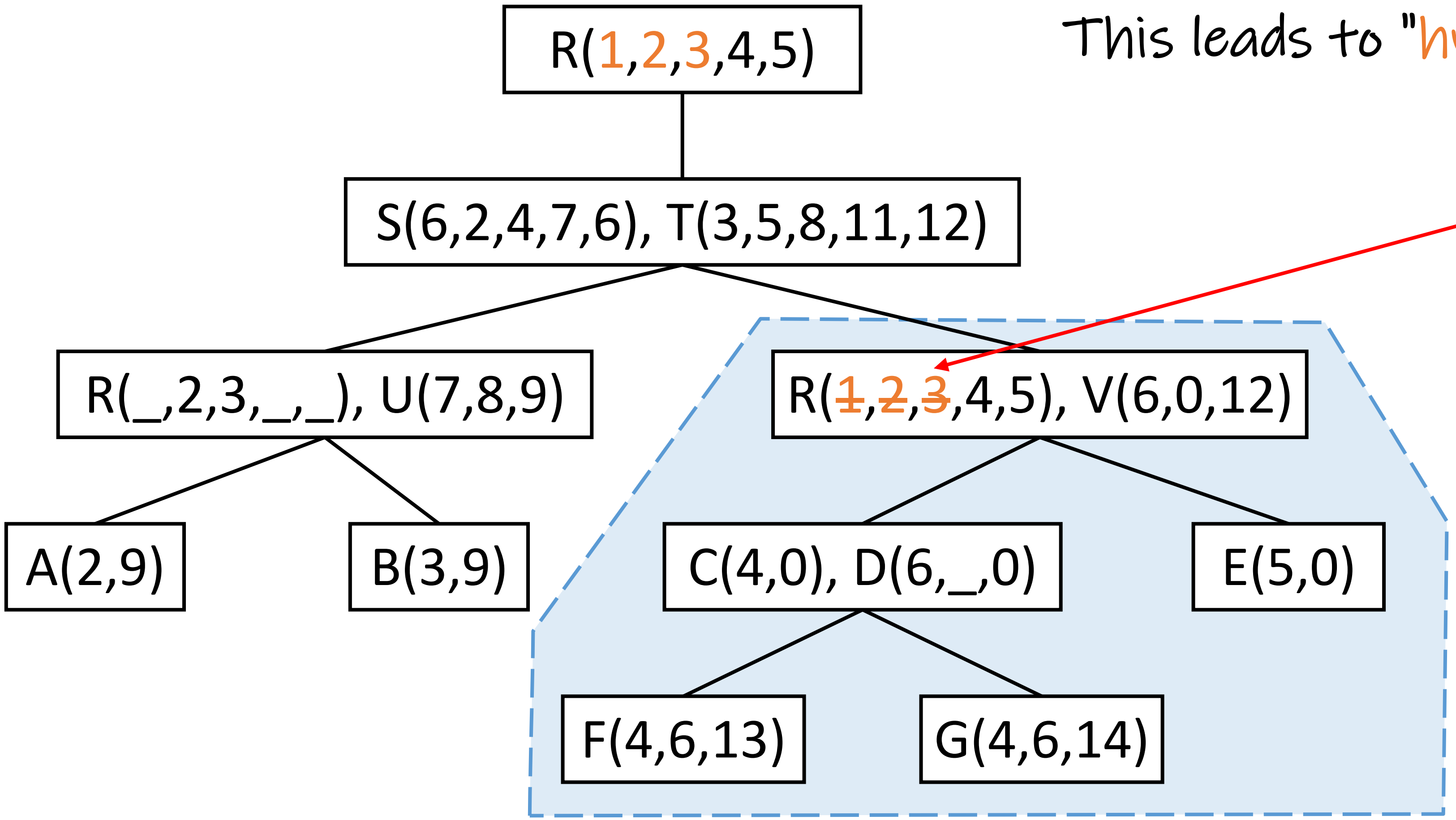
One can avoid NP-hardness of finding a minimal size decomposition by adding an additional syntactic "descendant condition".
This leads to "hypertree decompositions"



Adopted from an example by Georg Gottlob

Important Observation 2

One can avoid NP-hardness of finding a minimal size decomposition by adding an additional syntactic "descendant condition". This leads to "hypertree decompositions"



Each variable that disappears at some node, does not reappear in the subtree rooted at that node

HYPERTREE DECOMPOSITIONS AND TRACTABLE QUERIES *

Georg Gottlob
Inst. für Informationssysteme
Technische Universität Wien
A-1040 Vienna, Austria
gottlob@dbai.tuwien.ac.at

Nicola Leone
Inst. für Informationssysteme
Technische Universität Wien
A-1040 Vienna, Austria
leone@dbai.tuwien.ac.at

Francesco Scarcello
ISI-CNR
Via P. Bucci 41/C
I-87030 Rende, Italy
scarcello@si.deis.unical.it

Abstract

Several important decision problems on conjunctive queries (CQs) are NP-complete in general but become tractable, and actually highly parallelizable, if restricted to acyclic or nearly acyclic queries. Examples are the evaluation of Boolean CQs and query containment. These problems were shown tractable for conjunctive queries of bounded treewidth [9], and of bounded degree of cyclicity [24, 23]. The so far most general concept of nearly acyclic queries was the notion of queries of bounded query-width introduced by Chekuri and Rajaraman [9]. While CQs of bounded query-width are tractable, it remained unclear whether such queries are efficiently recognizable. Chekuri and Rajaraman [9] stated as an open problem whether for each constant k it can be determined in polynomial time if a query has query width $\leq k$. We give a negative answer by proving this problem NP-complete (specifically, for $k = 4$). In order to circumvent this difficulty, we introduce the new concept of hypertree decomposition of a query and the corresponding notion of hypertree width. We prove: (a) for each k, the class of queries with query width bounded by k is properly contained in the class of queries whose hypertree width is bounded by k; (b) unlike query width, constant hypertree-width is efficiently recognizable; (c) Boolean queries of constant hypertree-width can be efficiently evaluated.

descendent condition

Definition 3.1 A *hypertree decomposition* of a conjunctive query Q is a hypertree $\langle T, \chi, \lambda \rangle$ for Q which satisfies all the following conditions:

1. for each atom $A \in atoms(Q)$, there exists $p \in vertices(T)$ such that $var(A) \subseteq \chi(p)$;
2. for each variable $Y \in var(Q)$, the set $\{p \in vertices(T)$ s.t. $Y \in \chi(p)\}$ induces a (connected) subtree of T;
3. for each vertex $p \in vertices(T)$, $\chi(p) \subseteq var(\lambda(p))$;
4. for each vertex $p \in vertices(T)$, $var(\lambda(p)) \cap \chi(T_p) \subseteq \chi(p)$.

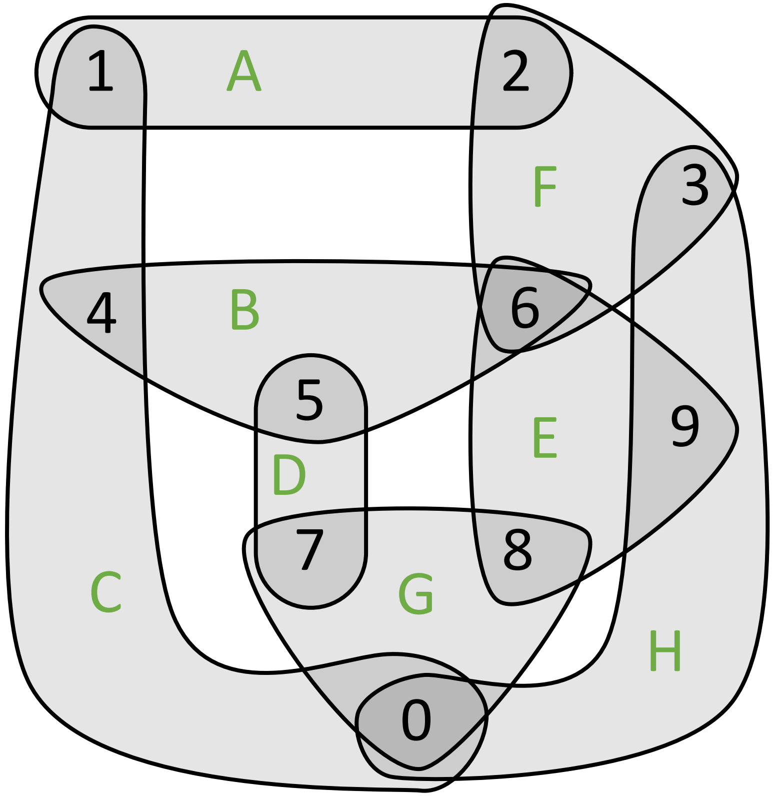
A hypertree decomposition $\langle T, \chi, \lambda \rangle$ of Q is a *complete decomposition* of Q if, for each atom $A \in atoms(Q)$, there exists $p \in vertices(T)$ such that $var(A) \subseteq \chi(p)$ and $A \in \lambda(p)$.

The *width* of the hypertree decomposition $\langle T, \chi, \lambda \rangle$ is $max_{p \in vertices(T)} |\lambda(p)|$. The *hypertree width* $hw(Q)$ of Q is the minimum width over all its hypertree decompositions.

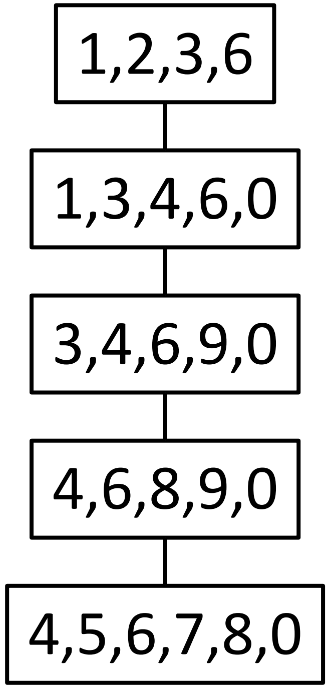
Source: Gottlob, Leone, Scarcello. "Hypertree decompositions and tractable queries." PODS 1999. https://doi.org/10.1145/303976.303979 (Gems of PODS 2016)
Wolfgang Gatterbauer. Principles of scalable data management: https://northeastern-datalab.github.io/cs7240/

Hypertree decomposition: full example

Hypergraph



Tree decomposition

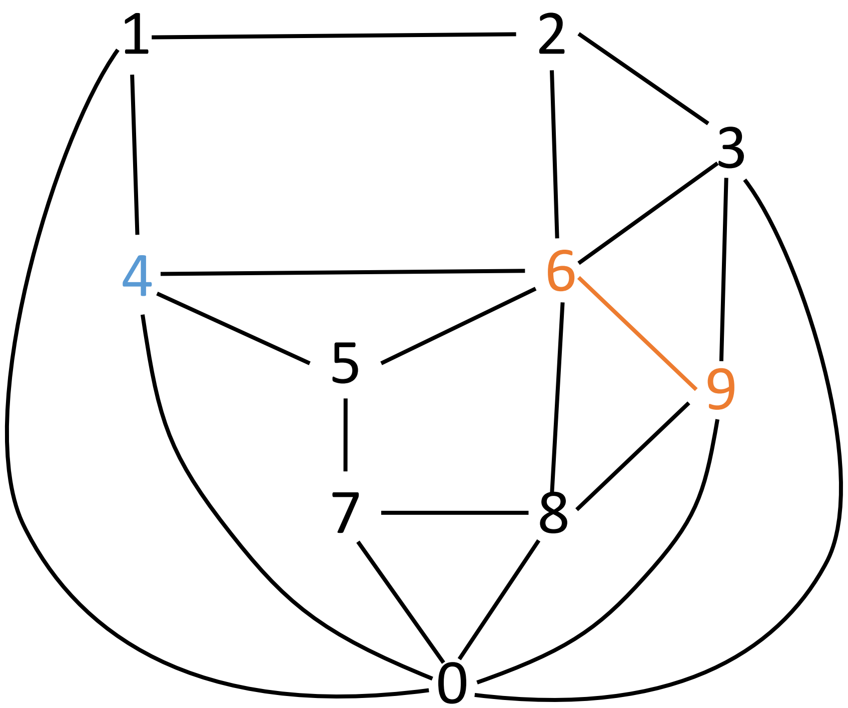


How to check that this is
a valid tree decomposition?

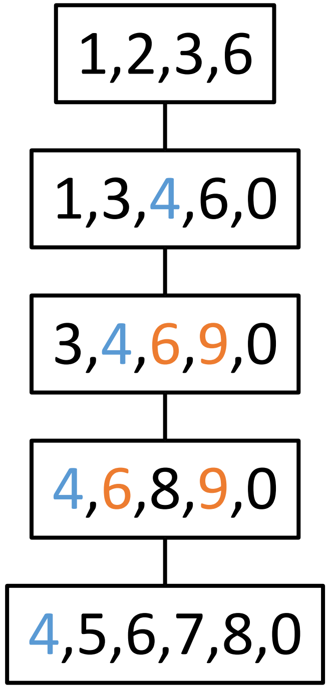
?

Hypertree decomposition: full example

Clique graph of Hypergraph
(also primal or Gaifman graph)



Tree decomposition



TREE DECOMPOSITION

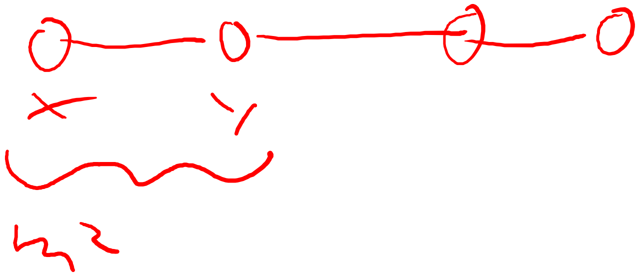
1. Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

2. Coherence

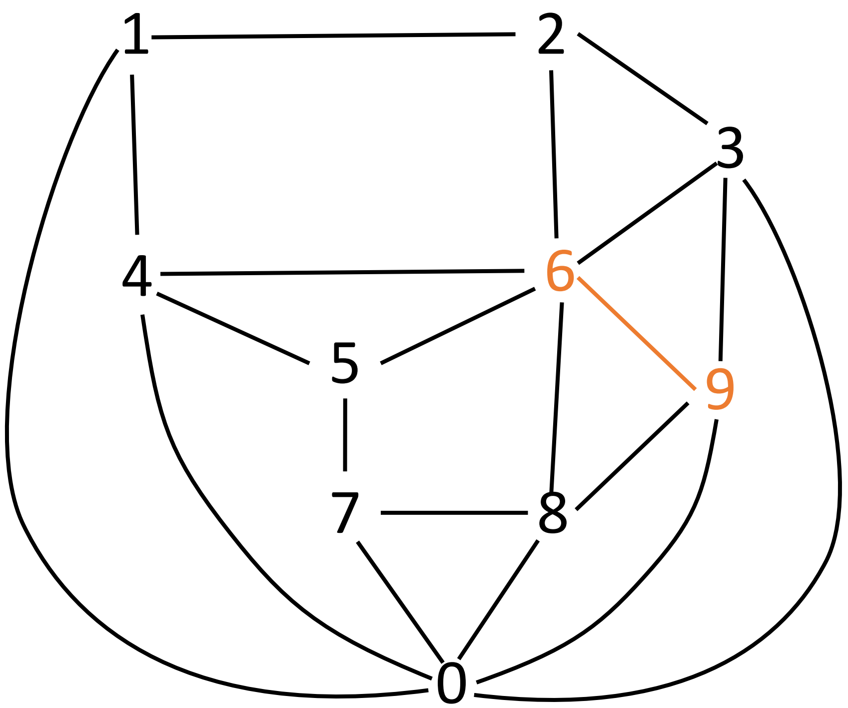
What is its width ?

Hypertree decomposition: full example

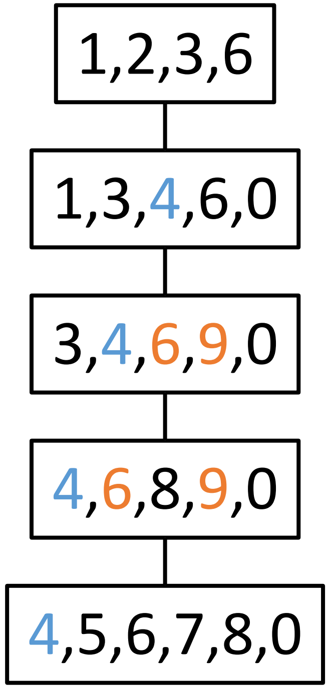
m=5



Clique graph of Hypergraph
(also primal or Gaifman graph)



Tree decomposition



TREE DECOMPOSITION

1. Edge coverage: For every edge e of G, there is a vertex in T that contains both ends of e

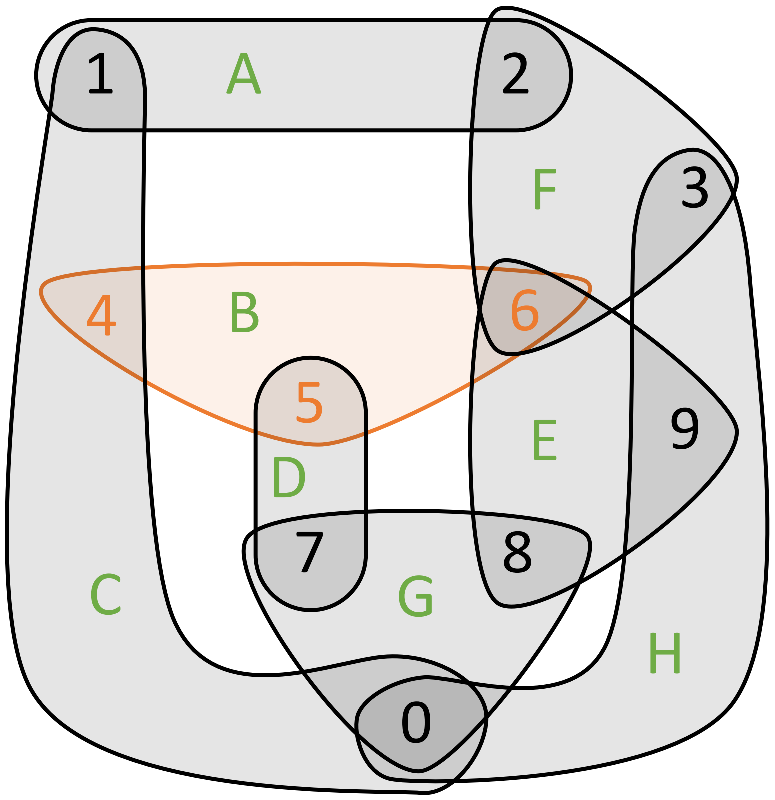
2. Coherence

guarantees evaluation in $O(m^6)$ where m is the domain size or $O(n^5)$ where n is size of largest relation

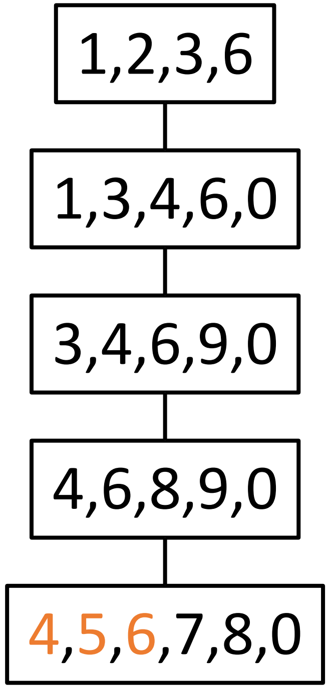
tree width = 5:
= size of largest supernode - 1

Hypertree decomposition: full example

Hypergraph



Tree decomposition
(width 5)



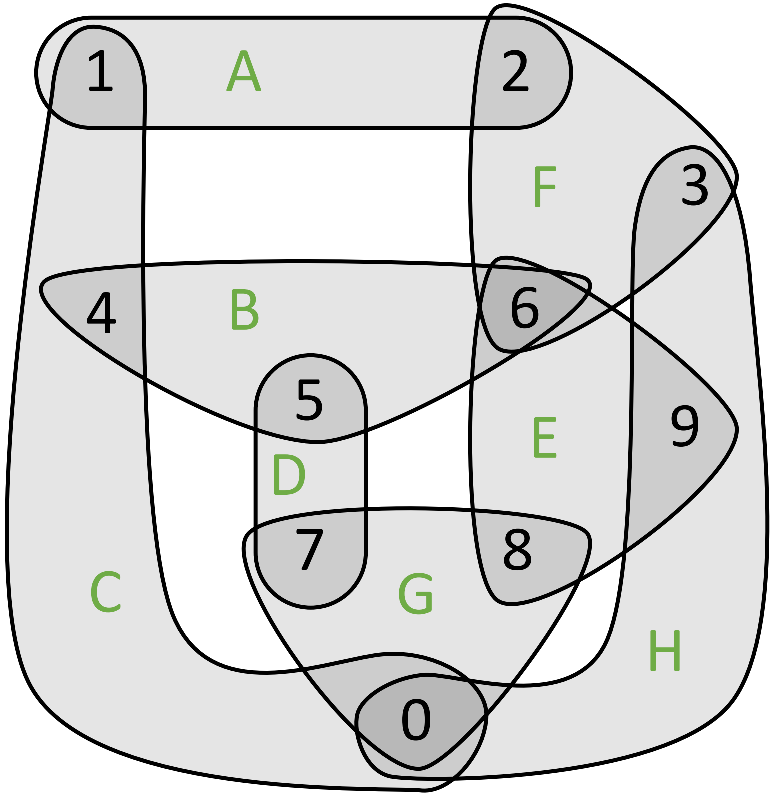
TREE DECOMPOSITION (ALTERNATIVE)

1. Hyperedge coverage: For every hyperedge h of H, there is a vertex in T that contains all its variables
2. Coherence

identical definition, because:
- hyperedge = clique in clique graph
- each clique needs to be contained in one supernode of the TD

Hypertree decomposition: full example

Hypergraph



Tree decomposition
(width 5)

1,2,3,6

1,3,4,6,0

3,4,6,9,0

4,6,8,9,0

4,5,6,7,8,0

Generalized hypertree decomp.
(width 2)

A{1,2}, F{2,3,6}

C{1,4,0}, F{~~2~~,3,6}

B{4,~~5~~,6}, H{3,9,0}

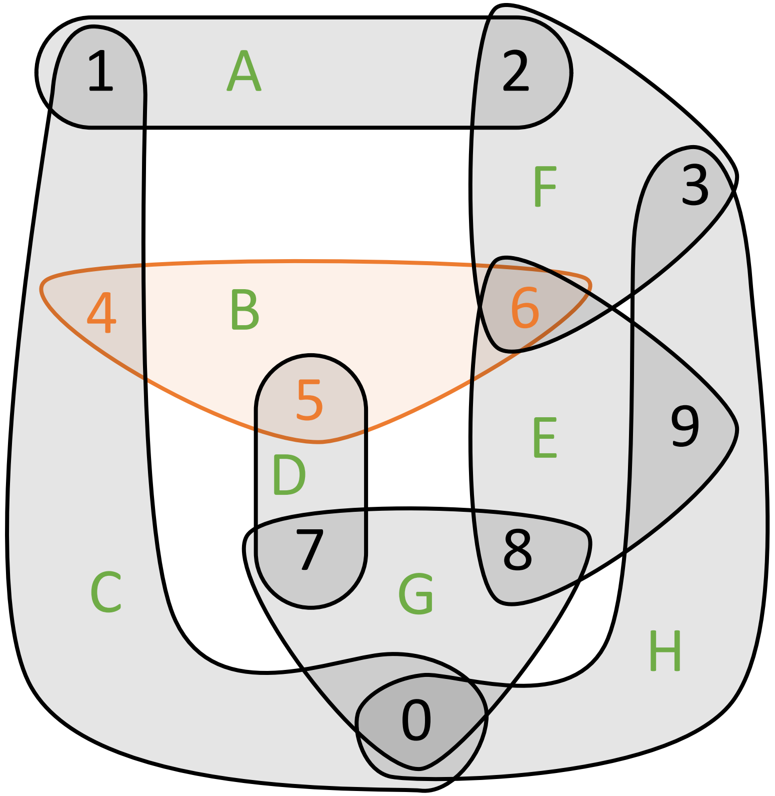
C{~~1~~,4,0}, E{6,8,9}

B{4,5,6}, G{7,8,0}

Why is this a valid "general. hypertree decomposition" ?

Hypertree decomposition: full example

Hypergraph



Tree decomposition (width 5)

GENERALIZED HT DECOMP.

1. Hyperedge coverage: For every hyperedge h of H, there is a vertex in T that contains all its variables
2. Coherence

Generalized hypertree decomp. (width 2)

A{1,2}, F{2,3,6}

C{1,4,0}, F{~~2~~,3,6}

B{4,~~5~~,6}, H{3,9,0}

C{~~1~~,4,0}, E{6,8,9}

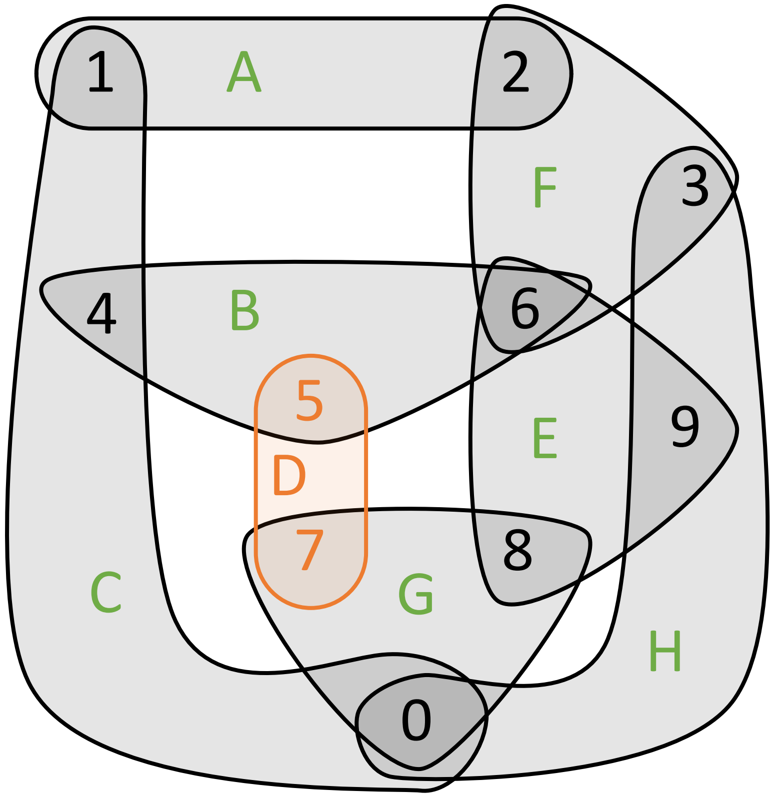
B{4,5,6}, G{7,8,0}

Basically identical to tree decomposition.
Just the width measure is different!

Example adopted from: Markus Krötzsch. "Database theory: Lecture 6: Tree-like Conjunctive Queries." 2016. https://iccl.inf.tu-dresden.de/web/Database_Theory_(SS2016)/en

Hypertree decomposition: full example

Hypergraph



Tree decomposition (width 5)

GENERALIZED HT DECOMP.

1. Hyperedge coverage: For every hyperedge h of H, there is a vertex in T that contains all its variables
2. Coherence

Basically identical to tree decomposition. Just the width measure is different!

Generalized hypertree decomp. (width 2)

A{1,2}, F{2,3,6}

C{1,4,0}, F{~~2~~,3,6}

B{4,~~5~~,6}, H{3,9,0}

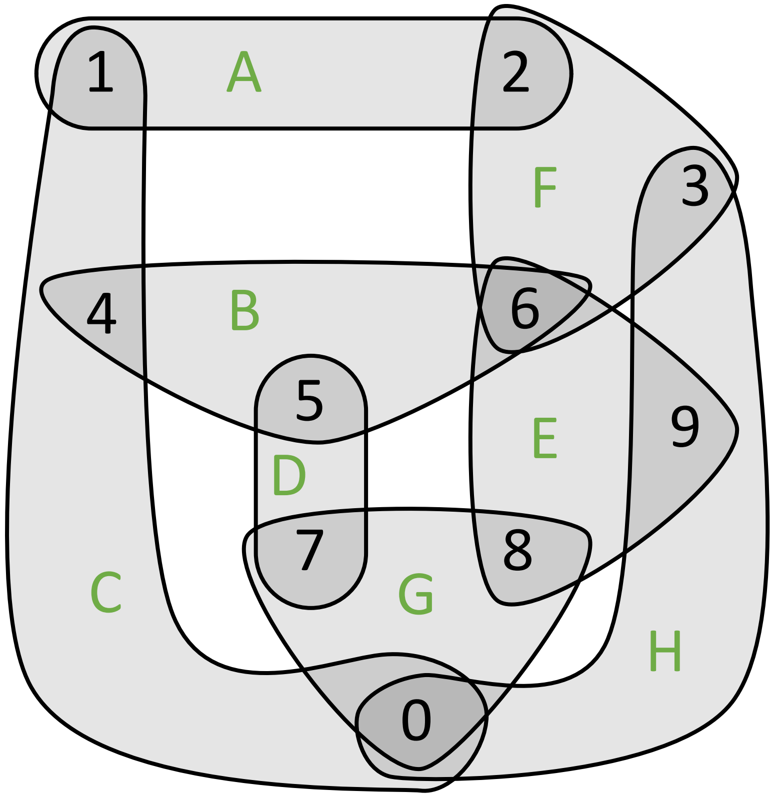
C{~~1~~,4,0}, E{6,8,9}

B{4,5,6}, G{7,8,0}

B and G together contain all variables from D

Hypertree decomposition: full example

Hypergraph



Generalized HT decomp.

1. Hyperedge coverage: For every hyperedge h of H, there is a vertex in T that contains all its variables
2. Coherence

Generalized hypertree decomp. (width 2)

A{1,2}, F{2,3,6}

C{1,4,0}, F{~~2~~,3,6}

B{4,~~5~~,6}, H{3,9,0}

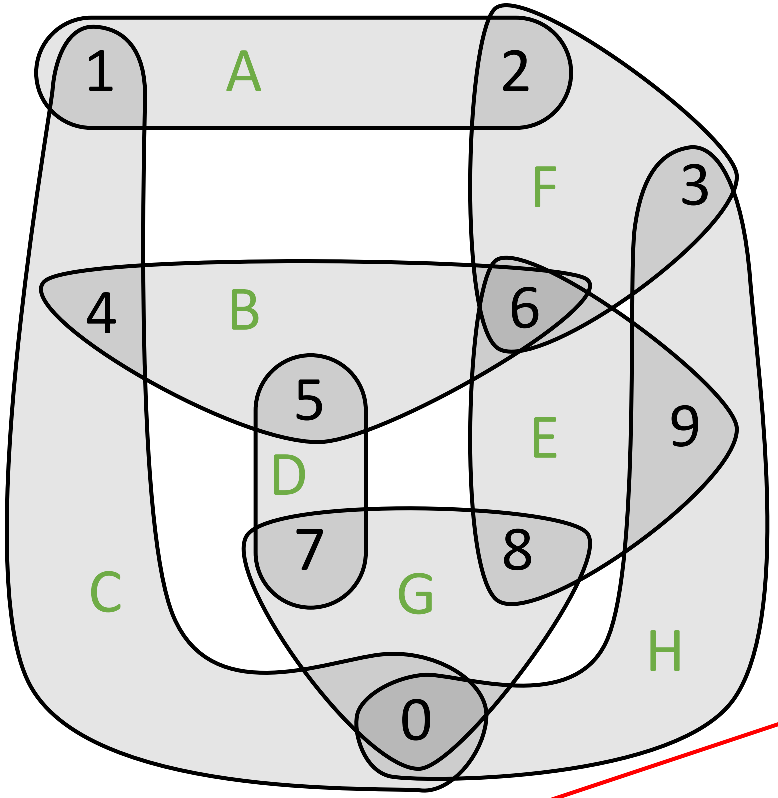
C{~~1~~,4,0}, E{6,8,9}

B{4,5,6}, G{7,8,0}

Is this also a valid "hypertree decomposition" ?

Hypertree decomposition: full example

Hypergraph



A condition to limit the search space of valid HD decompositions

HT DECOMP.

1. Hyperedge coverage: For every hyperedge h of H, there is a vertex in T that contains all its variables
2. Coherence
3. Descendant condition: Variables projected away from a hyperedge can not reappear in the subtree below

Generalized hypertree decomp. (width 2)

A{1,2}, F{2,3,6}

C{1,4,0}, F{~~2~~,3,6}

B{4,~~5~~,6}, H{3,9,0}

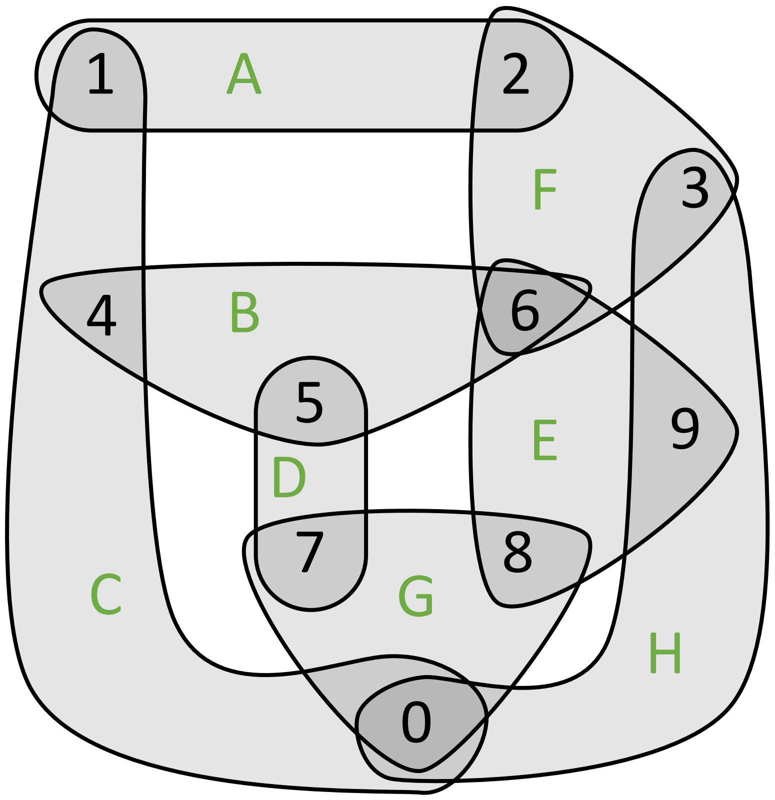
C{~~1~~,4,0}, E{6,8,9}

B{4,5,6}, G{7,8,0}

No: "5" got projected away, but reappears below. Also "1" in other direction

Hypertree decomposition: full example

Hypergraph



HT DECOMP.

1. Hyperedge coverage: For every hyperedge h of H, there is a vertex in T that contains all its variables
2. Coherence
3. Descendant condition: Variables projected away from a hyperedge can not reappear in the subtree below

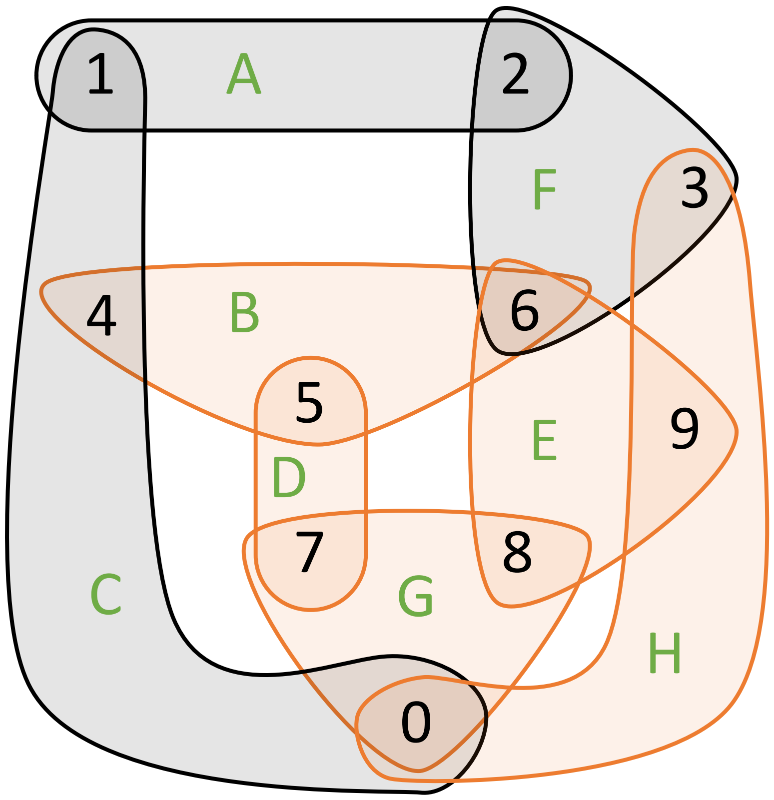
Hypertree decomposition

A{1,2}, C{1,4,0}, F{2,3,6}

B{4,5,6}, D{5,7}, E{6,8,9}, G{7,8,0}, H{3,9,0}

Hypertree decomposition: full example

Hypergraph



Hypertree decomposition

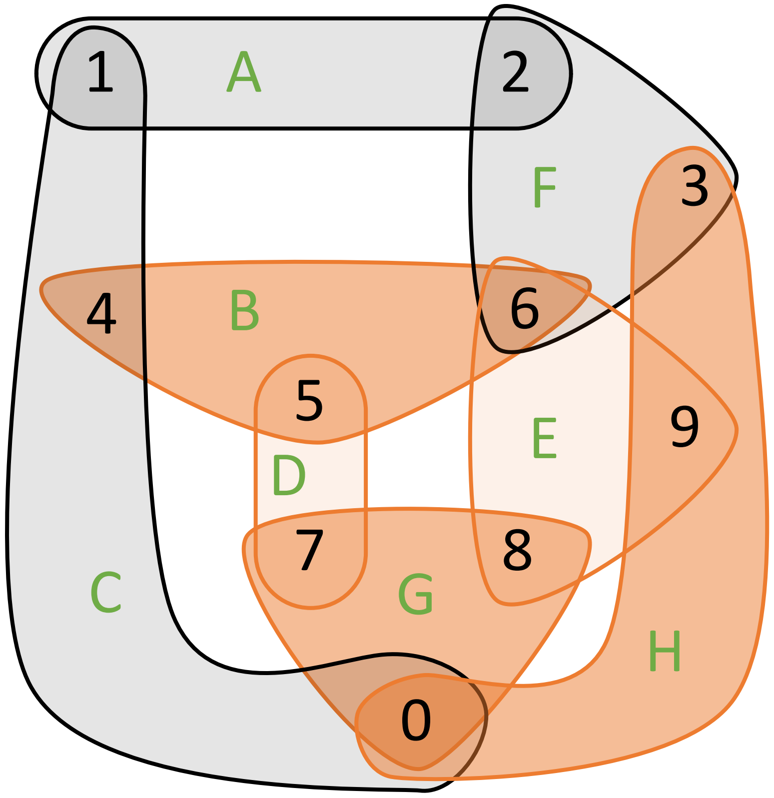
A{1,2}, C{1,4,0}, F{2,3,6}

B{4,5,6}, D{5,7}, E{6,8,9}, G{7,8,0}, H{3,9,0}

What should be the "width" of this HTD, i.e. what is the complexity of materializing this last supernode ?

Hypertree decomposition: full example

Hypergraph



Hypertree decomposition

A{1,2}, C{1,4,0}, F{2,3,6}

B{4,5,6}, D{5,7}, E{6,8,9},
G{7,8,0}, H{3,9,0}

B(4,5,6)⋈G(7,8,0)⋈H(3,9,0)

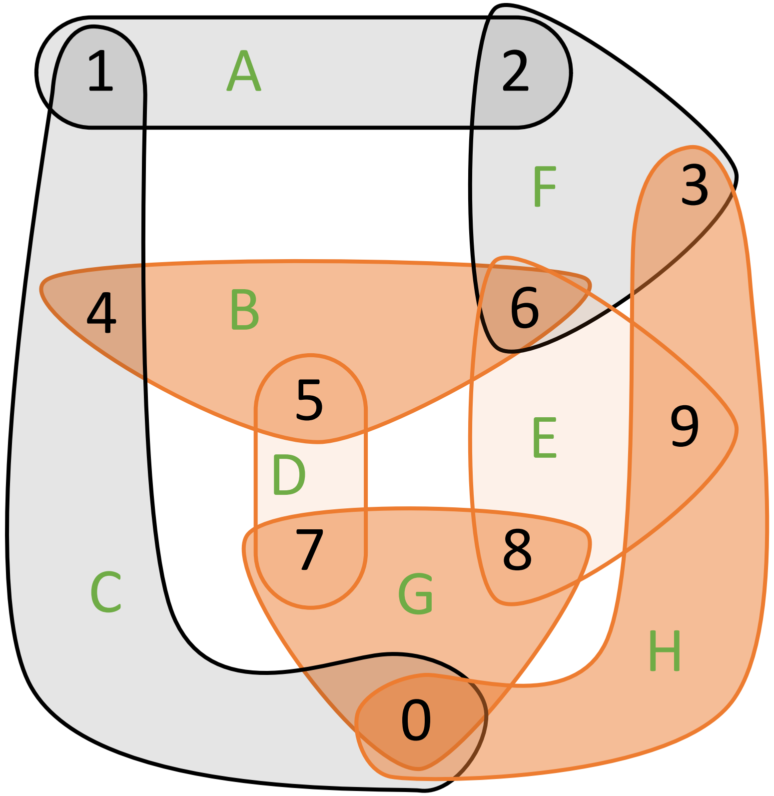
Notice that 3 relations alone "cover" all the variables.
The join can only be a subset of those tuples.

([(B(4,5,6) ⋈ G(7,8,0)) ⋈ H(3,9,0)] ← $O(n^3)$
⋉D(5,7)) ⋉E(6,8,9)

n… maximal size of relations

Hypertree decomposition: full example

Hypergraph



Hypertree decomposition
(width 3)

C,F: {1,2,3,4,6,0}
B,G,H:{3,4,5,6,7,8,9,0}

B⋈G⋈H

With of HTD = maximal width of any super node.
With of supernode = minimal number of relations to cover all variables. Here covered by B⋈G⋈H

Results in a modified database and modified acyclic query. Then perform Yannakakis: $O(n^3)$

Hypertree Decompositions: A Survey

Georg Gottlob[1], Nicola Leone[2], and Francesco Scarcello[3]

descendent condition

generalized. For instance, let us define the concept of *generalized hypertree decomposition* by just dropping condition 4 from the definition of hypertree decomposition (Def. 11). Correspondingly, we can introduce the concept of *generalized hypertree width* $ghw(\mathcal{H})$ of a hypergraph $\mathcal{H}$. We know that all classes of Boolean queries having bounded ghw can be answered in polynomial time. But we currently do not know whether these classes of queries are polynomially recognizable. This recognition problem is related to the mysterious *hypergraph*

Hypertree width and related hypergraph invariants

Isolde Adler[a], Georg Gottlob[b], Martin Grohe[c]

European Journal of Combinatorics 28 (2007) 2167–2181

$$\mathrm{ghw}(H) \leq \mathrm{hw}(H) \leq \mathrm{tw}(H) + 1.$$

$$\mathrm{hw}(H) \leq 3 \cdot \mathrm{ghw}(H) + 1$$

Generalized Hypertree Decompositions: NP-Hardness and Tractable Variants

Georg Gottlob
University of Oxford
Computing Laboratory
georg.gottlob@
comlab.ox.ac.uk

Zoltán Miklós
University of Oxford and
Technische Universität Wien
zoltan.miklos@
comlab.ox.ac.uk

Thomas Schwentick
Universität Dortmund
Lehrstuhl Informatik I
thomas.schwentick@
udo.edu

ABSTRACT

The generalized hypertree width $GHW(H)$ of a hypergraph H is a measure of its cyclicity. Classes of conjunctive queries or constraint satisfaction problems whose associated hypergraphs have bounded GHW are known to be solvable in polynomial time. However, it has been an open problem for several years if for a fixed constant k and input hypergraph H it can be determined in polynomial time whether $GHW(H) \leq k$. Here, this problem is settled by proving that even for $k = 3$ the problem is already NP-hard. On

Hypertree Decompositions and friends

Query decomposition
[Chekuri, Rajaraman 1997]

NP-complete to find the optimum

towards tighter bounds
(below is better)

Hypertree Decomposition (HD)
[Gottlob, Leone, Scarcello 1999]

PTIME to find the optimum

towards tighter bounds
(below is better)

Generalized Hypertree Decomposition (GHD)
[Gottlob, Leone, Scarcello 2001]

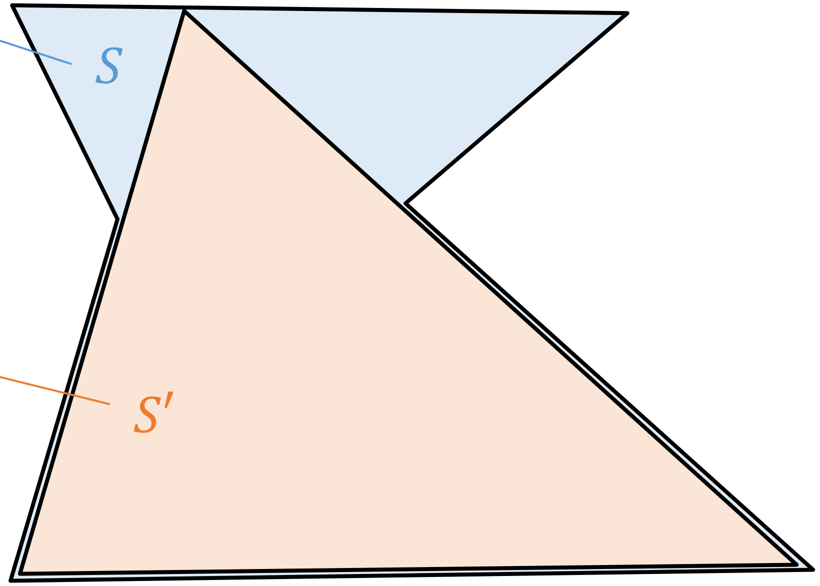
NP-complete to find the optimum

Chekuri, Rajaraman. "Conjunctive query containment revisited", TCS 2000. https://doi.org/10.1016/S0304-3975(99)00220-0 (ICDT'97 conference paper, ICDT'16 test-of-time award)
Gottlob, Leone, Scarcello. "Hypertree decompositions and tractable queries." PODS 1999. https://doi.org/10.1145/303976.303979 (Gems of PODS 2016)
Gottlob, Leone, Scarcello. "Hypertree decompositions: a survey." MFCS 2001. https://dl.acm.org/doi/10.5555/645730.668191

Hypertree Decomposition: an unfortunate naming

1. Generalized Hypertree Decomposition (GHD):
explores the whole search space of valid decompositions
(illustrated here with a non-convex search space S in blue)

2. Hypertree Decomposition (HD):
limits the search space in a way that makes it tractable
to find the optimal solution within that limited subspace
(illustrated here with a convex search space $S' \subseteq S$)



Better names would be:

1. Hypertree Decomposition (HD) instead of GHD

2. Restricted Hypertree Decomposition (RHD) instead of HD

Updated 4/3/2023

Topic 3: Efficient query evaluation
Unit 2: Cyclic query evaluation
Lecture 23

Wolfgang Gatterbauer

CS7240 Principles of scalable data management (sp23)

https://northeastern-datalab.github.io/cs7240/sp23/

4/3/2023

Pre-class conversations

- Last class summary
- Project: comments finished on about 1/3 (4)
- Scribes

- Today:
  - Linear Programming Duality, min-cut-max-flow

Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

Topic Duality in Linear Programming (LP)

- Subtopics
  - Connections between (max) set packing and (min) set covers in graphs
  - Linear Programming (LP) and duality gaps
  - LP relaxations of ILP problems (Integer Linear Programming)
  - Duality b/w independent vertex sets and edge covers

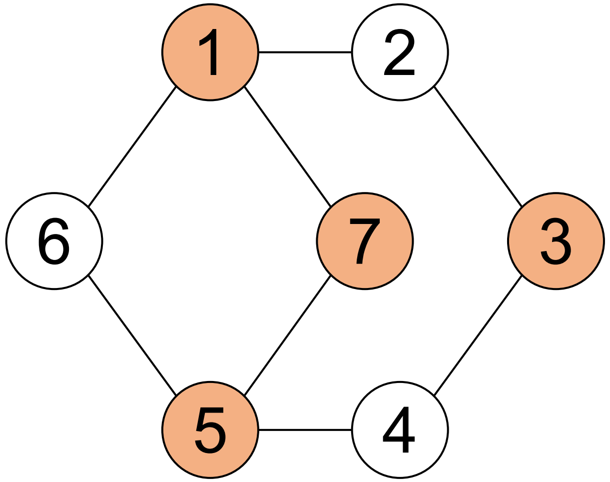
Duality in linear programming: Intuitively, every Linear Program has a dual problem with the same optimal solution, but the variables in the dual problem correspond to constraints in the primal problem and vice versa.
But the notion of duality is more general:

- "Over and over again, it turns out that one can associate with a given mathematical object a related, 'dual' object that helps one ... understand the properties of the object one started with." [The Princeton Companion to Mathematics, 2008]
- "Fundamentally, duality gives two different points of view of looking at the same object." [Michael Atiyah, 2007]

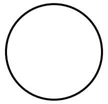
Let's use graphs to explain duality in LP (Linear Programming)

- **(max) Packing** problems: max number of disjoint subsets
  - max set packing: max number of subsets that are pairwise disjoint
  - max independent (vertex) set: max number of vertices not sharing edges
  - max independent edge set = matching: maximum number of edges that don't share any nodes (every vertex can be in max one matching)
- **(min) Coverings** problems: min number of subsets to cover all elements
  - min set cover: min number of subsets to cover the entire domain
  - min vertex cover: min number of vertices to cover all edges
  - min edge cover: min number of edges to cover all vertices
- Some packing problem is the dual problems of some covering problem
  - Min Vertex Cover (VC) is the dual of Max matching
  - Max Independent Set (IS) is the dual of Min edge cover

Independent set

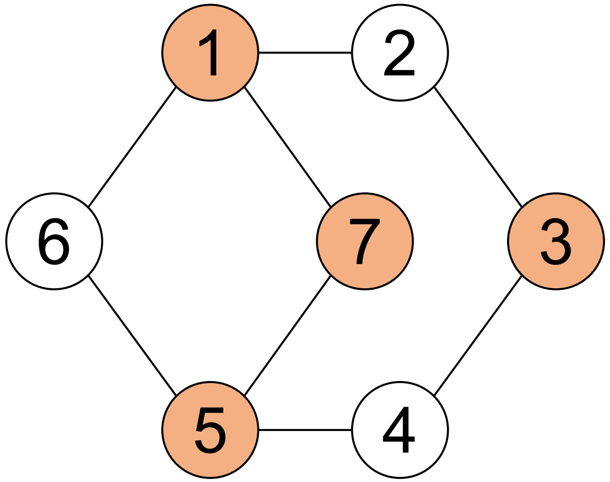


Independent set (IS): set of vertices that are not connected (white)



max

VC vs. Ind set ?



Independent set (IS): set of vertices that are not connected (white)

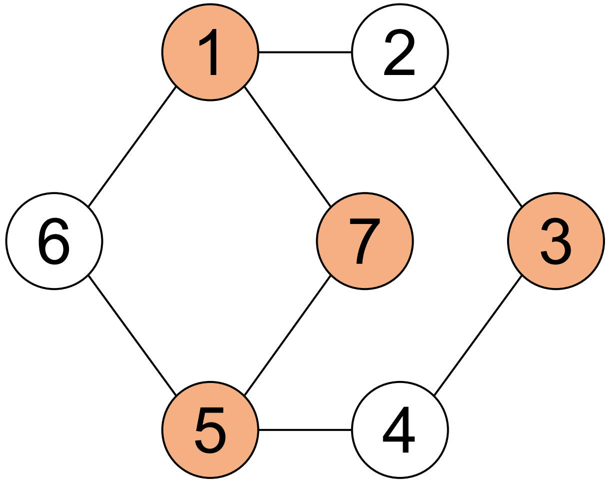
max

Vertex cover (VC): set of vertices that covers all edges

Assume you are given an independent set.
How do you find a vertex cover?

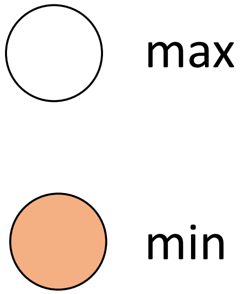
?

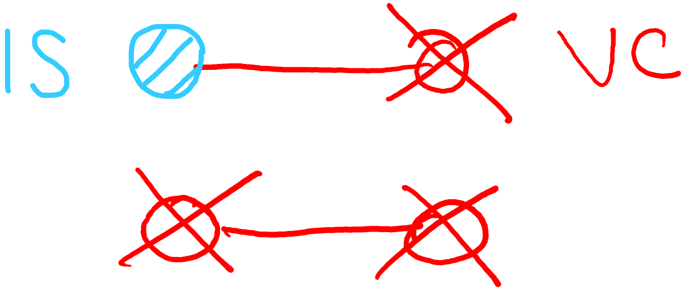
VC $=^c$ Ind set



Independent set (IS): set of vertices that are not connected (white)

Vertex cover (VC): set of vertices that covers all edges (orange)

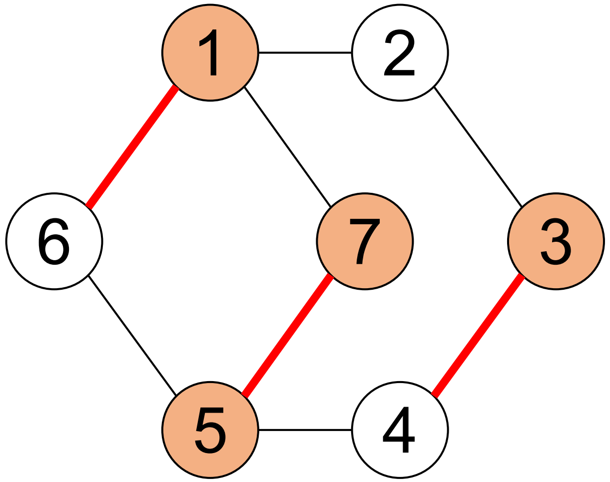




Set S is a **VC** iff the **complement** $V^c = V - S$ is an **IS**

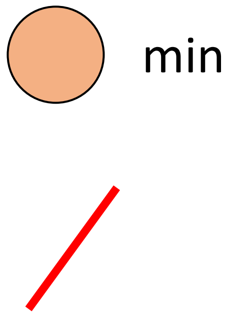
Proof: for each edge at most one vertex is in V^c.
Thus at least one vertex is in Set S.

Matching vs. VC?



Vertex cover (VC): set of vertices that covers all edges (orange)

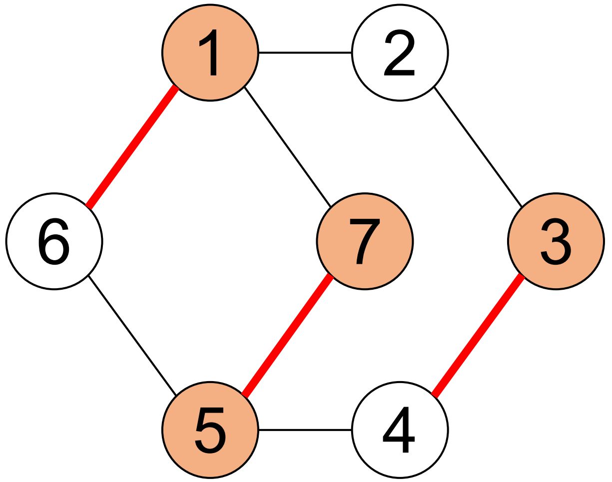
Matching (Ind edge set): set of edges w/o common vertices (red)



What is a possible connection between VC and matchings

?

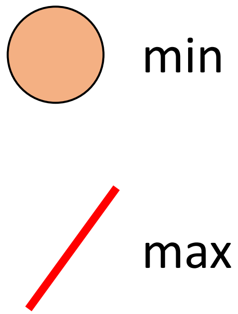
Matching ≤ VC



≤

Vertex cover (VC): set of vertices that covers all edges (orange)

Matching (Ind edge set): set of edges w/o common vertices (red)



A **VC** needs to cover at least each edge from any matching

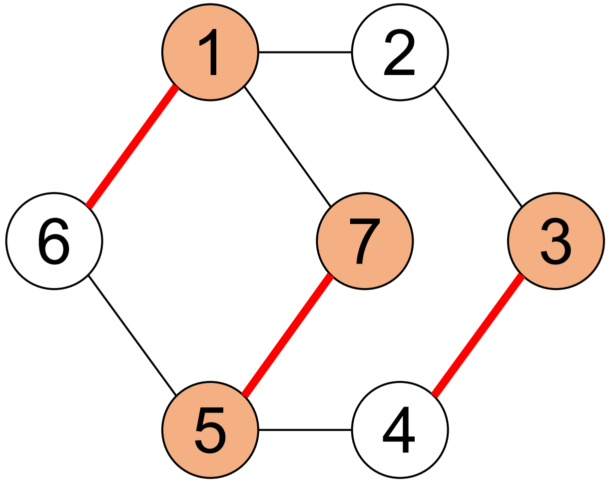
Thus, any VC has at least the size of any matching

⇒ **Size of any matching** ≤ any VC

That turns out to be the dual:

Max Matching ≤ Min VC

Matching ≤ VC $=^c$ Ind set (summary so far)



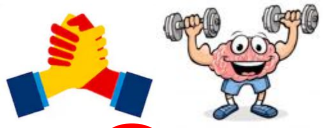
What intuitive problem is missing ?

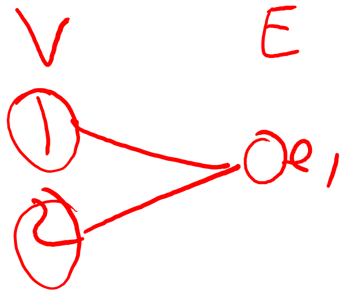
$\subset$

$\leq$

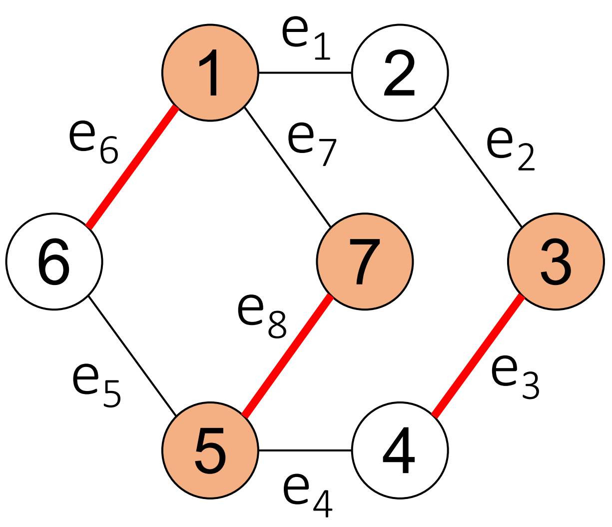
Independent set (IS): set of vertices that are not connected (white)	○	max
Vertex cover (VC): set of vertices that covers all edges (orange)	●	min
Matching (Ind edge set): set of edges w/o common vertices (red)	/	max

Matching ≤ VC =^{c} Ind set (summary so far)





What intuitive problem is missing ?



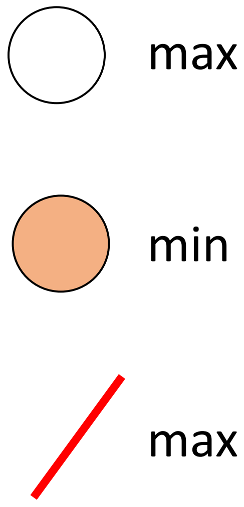
Independent set (IS): set of vertices that are not connected (white)

c

Vertex cover (VC): set of vertices that covers all edges (orange)

≤

Matching (Ind edge set): set of edges w/o common vertices (red)



Edges = Sets

Vertices = elements

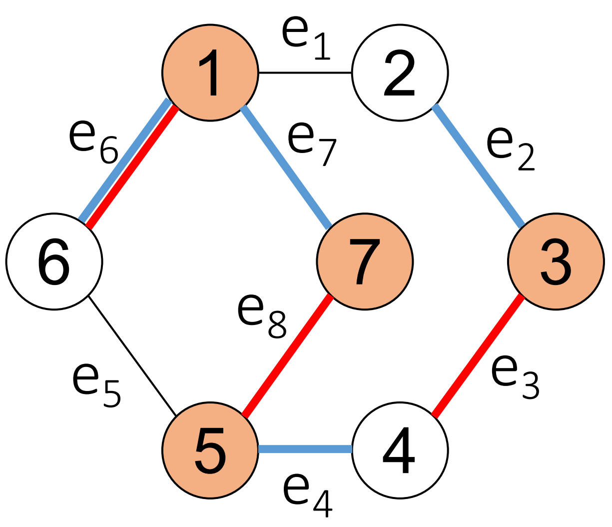
	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
1	o					o	o	
2	o	o						
3		o	o					
4			o	o				
5				o	o			o
6					o	o		
7							o	o

Cover problems: set of subsets that cover all elements

Packing problems: set of disjoint subsets

Matching ≤ VC =^{c} Ind set vs. Edge cover

What is its connection to IS ?



Edge cover: set of edges that cover all vertices (blue)

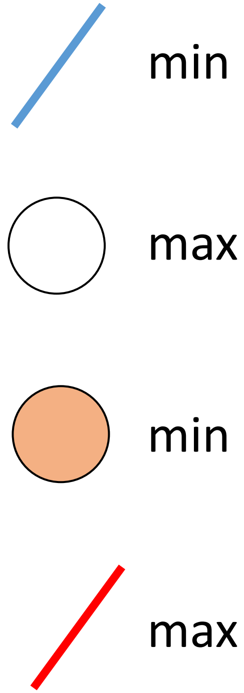
Independent set (IS): set of vertices that are not connected (white)

c

Vertex cover (VC): set of vertices that covers all edges (orange)

≤

Matching (Ind edge set): set of edges w/o common vertices (red)



Edges = Sets

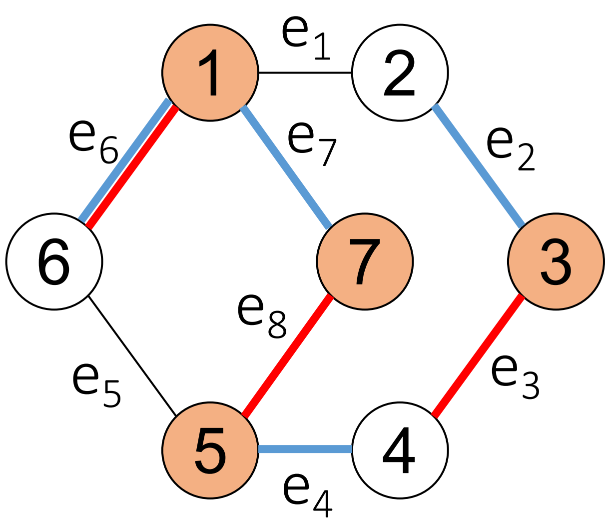
Vertices = elements

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
1	o					o	o	
2	o	o						
3		o	o					
4			o	o				
5				o	o			o
6					o	o		
7							o	o

Cover problems: set of subsets that cover all elements (min set cover: min vertex cover, min edge cover)

Packing problems: set of disjoint subsets (max set packing: max ind set, max matching)

Matching ≤ VC =^{c} Ind set ≤ Edge cover



Edges = Sets

Vertices = elements

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
1	o					o	o	
2	o	o						
3		o	o					
4			o	o				
5				o	o			o
6					o	o		
7							o	o

≤

Edge cover: set of edges that cover all vertices (blue) — min

Independent set (IS): set of vertices that are not connected (white) — max

⊂

≤

Vertex cover (VC): set of vertices that covers all edges (orange) — min

Matching (Ind edge set): set of edges w/o common vertices (red) — max

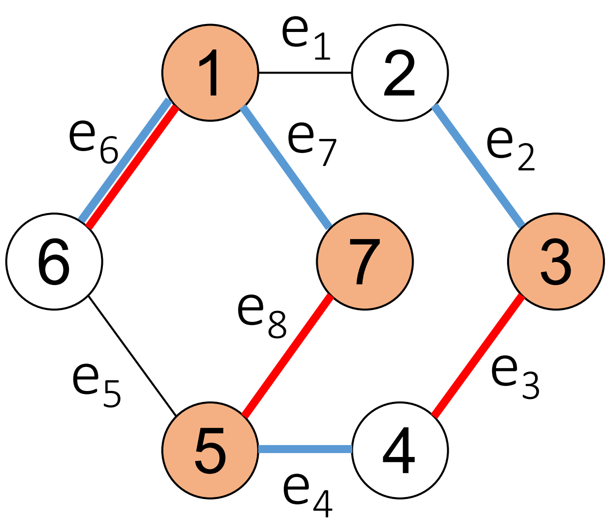
An **edge cover** needs to cover at least each vertex from any IS

Thus, any IS is lower bound to the size of any edge cover

⇒ **Size of min edge cover** ≥ max IS

Duality: Max IS ≤ Min edge cover

4 graph problems in the incidence matrix



Edges = Sets

Vertices = elements

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
1	o					o	o	
2	o	o						
3		o	o					
4			o	o				
5				o	o			o
6					o	o		
7							o	o

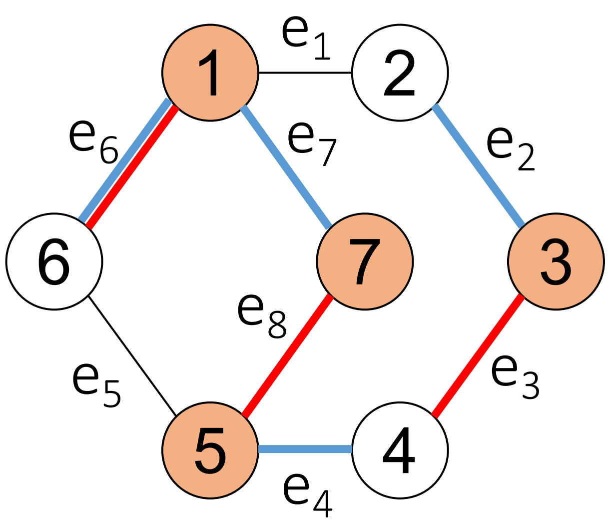
	Choose Vertices	Choose Edges
Set **Cover**	min=4 Vertex Cover	min=4 Edge Cover
Set **Packing**	max=3 Independent Set	max=3 Matching = Ind. edge set

complement

$\leq$ dual

$\geq$ dual

4 graph problems in the incidence matrix



Edges = Sets

Vertices = elements

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
1	o					o	o	
2	o	o						
3		o	o					
4			o	o				
5				o	o			o
6					o	o		
7							o	o

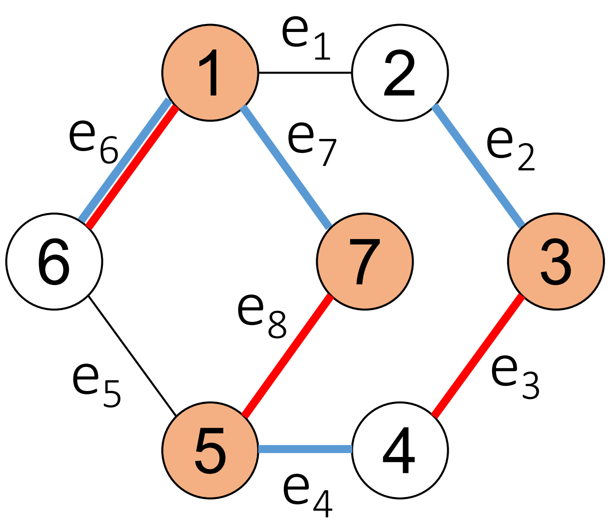
	NP-complete	PTIME
	Choose Vertices	Choose Edges
Set **Cover**	min=4 Vertex Cover	min=4 Edge Cover
Set **Packing**	max=3 Independent Set	max=3 Matching = Ind. edge set

complement

≤ dual

≥ dual

4 graph problems in the incidence matrix



Edges = Sets

Vertices = elements

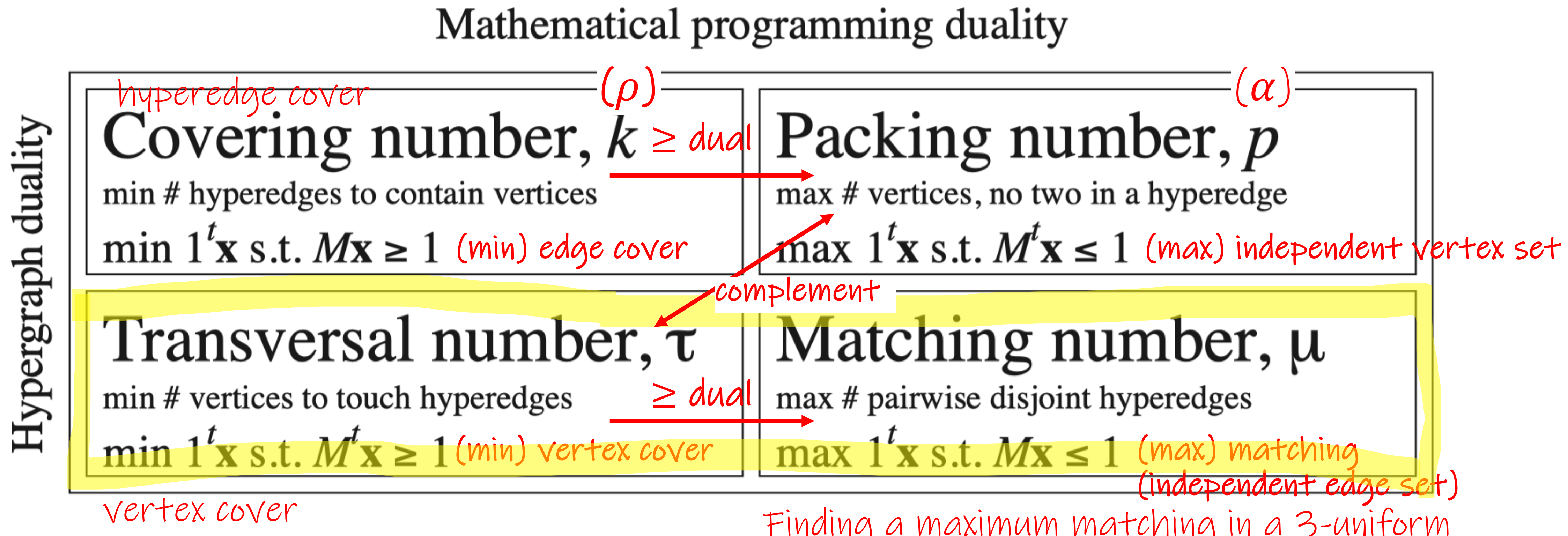
	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
1	o					o	o	
2	o	o						
3		o	o					
4			o	o				
5				o	o			o
6					o	o		
7							o	o

	NP-complete Choose Vertices	PTIME Choose Edges
Set **Cover**	τ=4 (min) Vertex Cover	ρ=4 (min) Edge Cover
Set **Packing**	α=3 (max) Independent Set	μ=3 (max) Matching = Ind. edge set

complement

$\leq$ dual

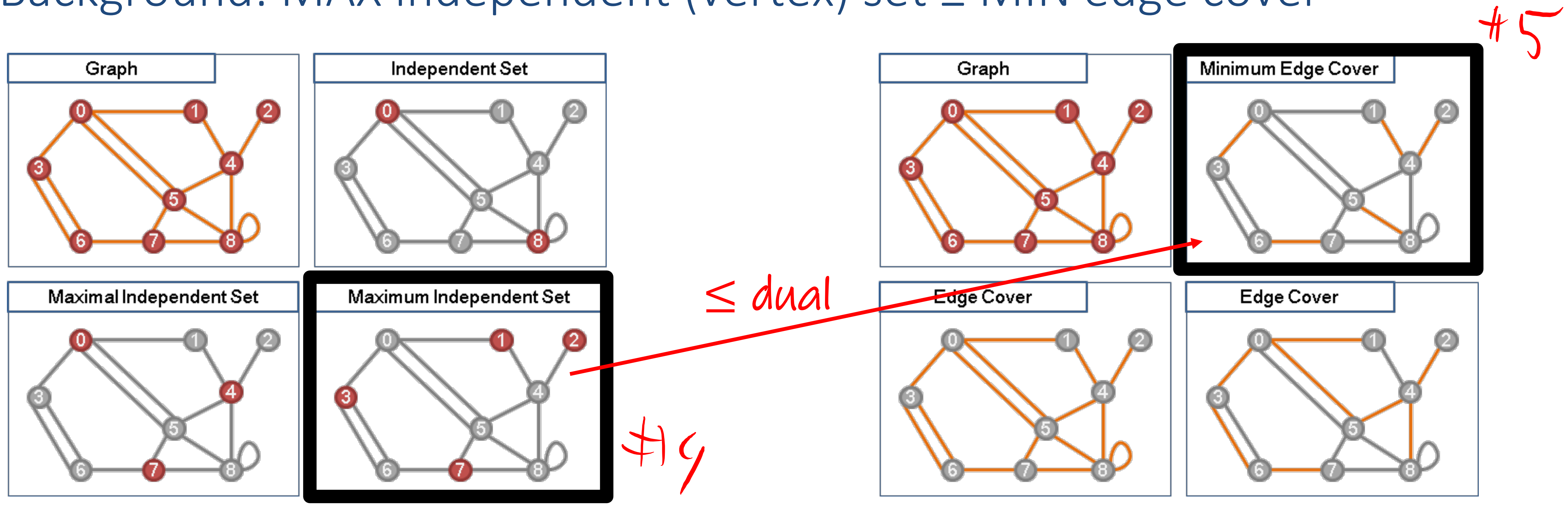
$\geq$ dual



Finding a maximum matching in a 3-uniform hypergraph is NP-hard (3-dimensional matching), but is in PTIME for simple (2-uniform) graphs.

Figure 1.1. The dualities between the covering, packing, transversal, and matching numbers of a hypergraph.

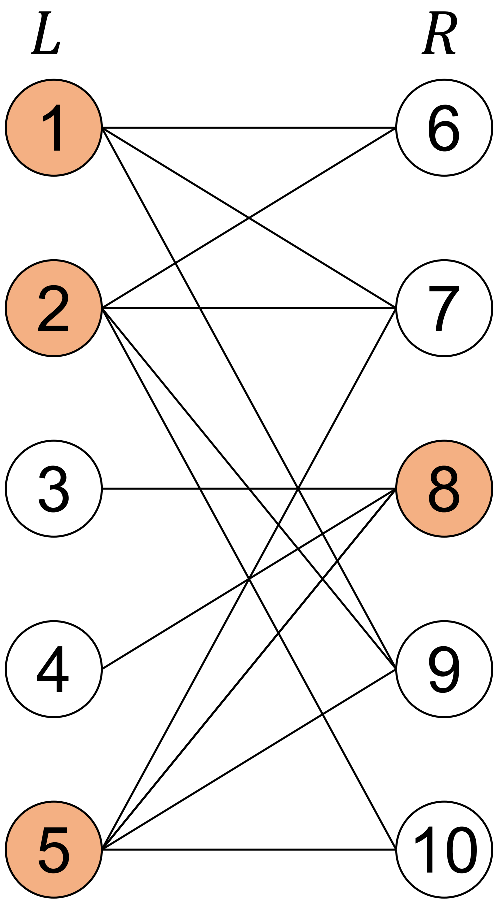
Background: MAX independent (vertex) set ≤ MIN edge cover



- Assume graph G is connected. Thus, every vertex has at least one edge (unless just one vertex)
- Suppose S is an independent set and E is an edge cover.
- Then for each vertex $v \in S$ there exists at least one edge $e \in E$ incident with v.
- By definition of independent set no two $u, v \in S$, have a common edge in E.
- Therefore $|S| \leq |E|$

Matching ≤ VC: what changes in bipartite graphs?

Nodes are partitioned into Left and Right



≤

Vertex cover (VC): set of vertices that covers all edges (orange)

Matching (Ind edge set): set of edges w/o common vertices (red)

min

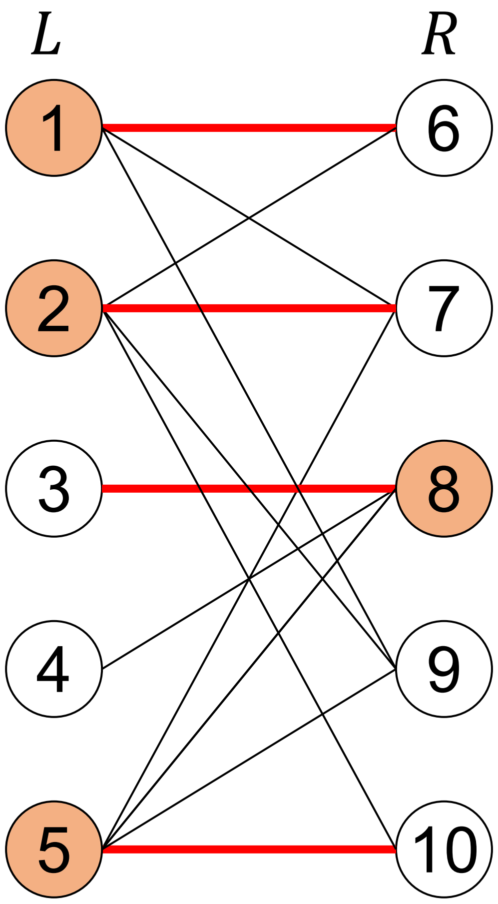
max

A **VC** needs to cover at least each edge from any matching

Thus, min VC at least the size of any matching

⇒ **Size of any matching** ≤ any VC

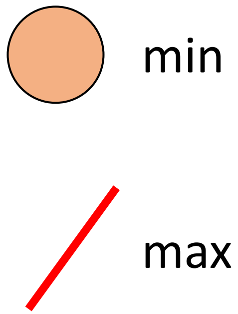
matching = VC ... in bipartite graphs!



=

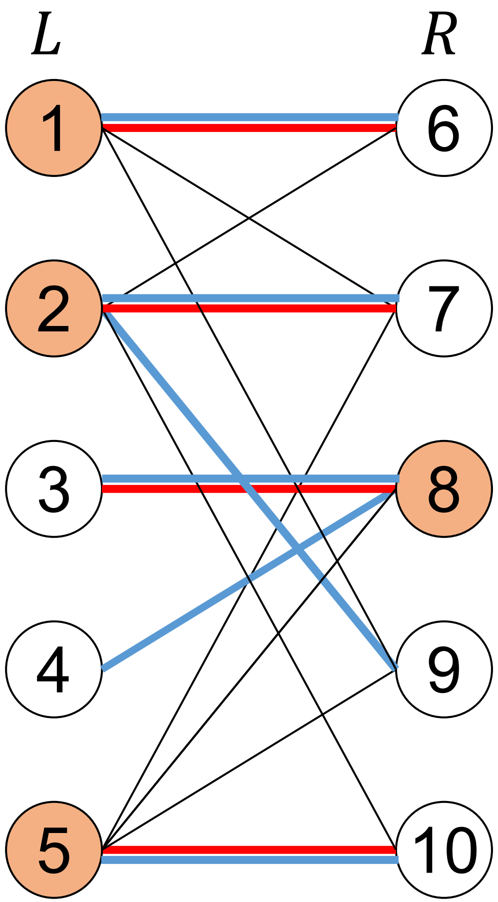
Vertex cover (VC): set of vertices that covers all edges (orange)

Matching (Ind edge set): set of edges w/o common vertices (red)



Kőnig-Egeváry theorem for bipartite graphs:
Max matching equivalent **to Min VC**

All for 4 problems become easy in bipartite graphs



PTIME

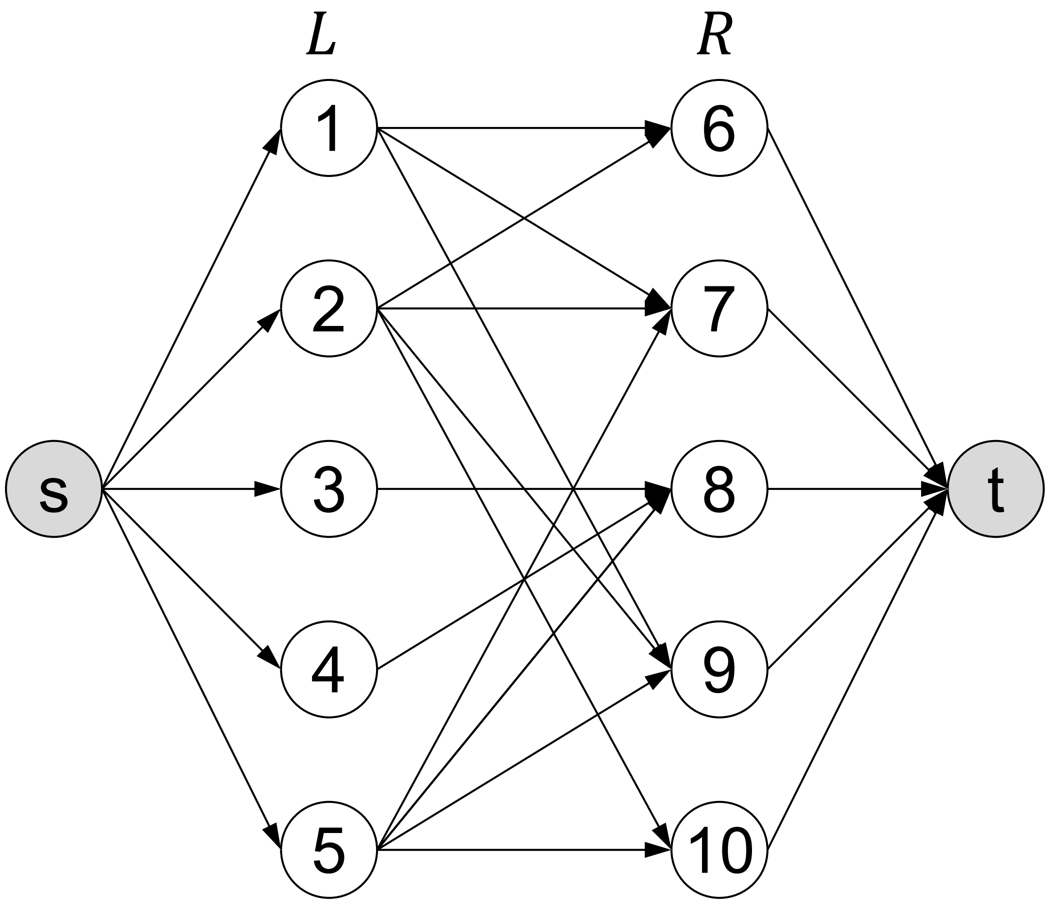
	Choose Vertices	Choose Edges
Set Cover	Vertex Cover	Edge Cover
Set Packing	Independent Set	Matching = Ind. edge set

complement

= dual

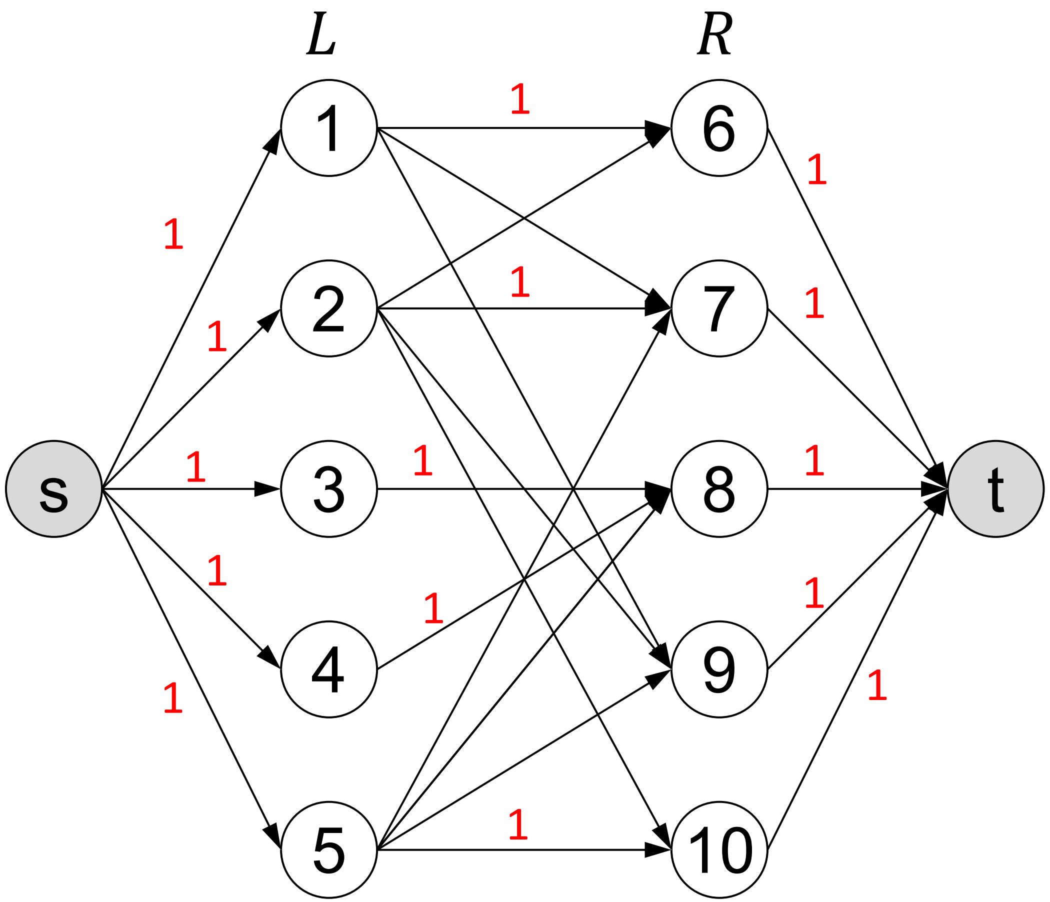
= dual

Cuts and Flows in directed graphs $G = (V, E)$



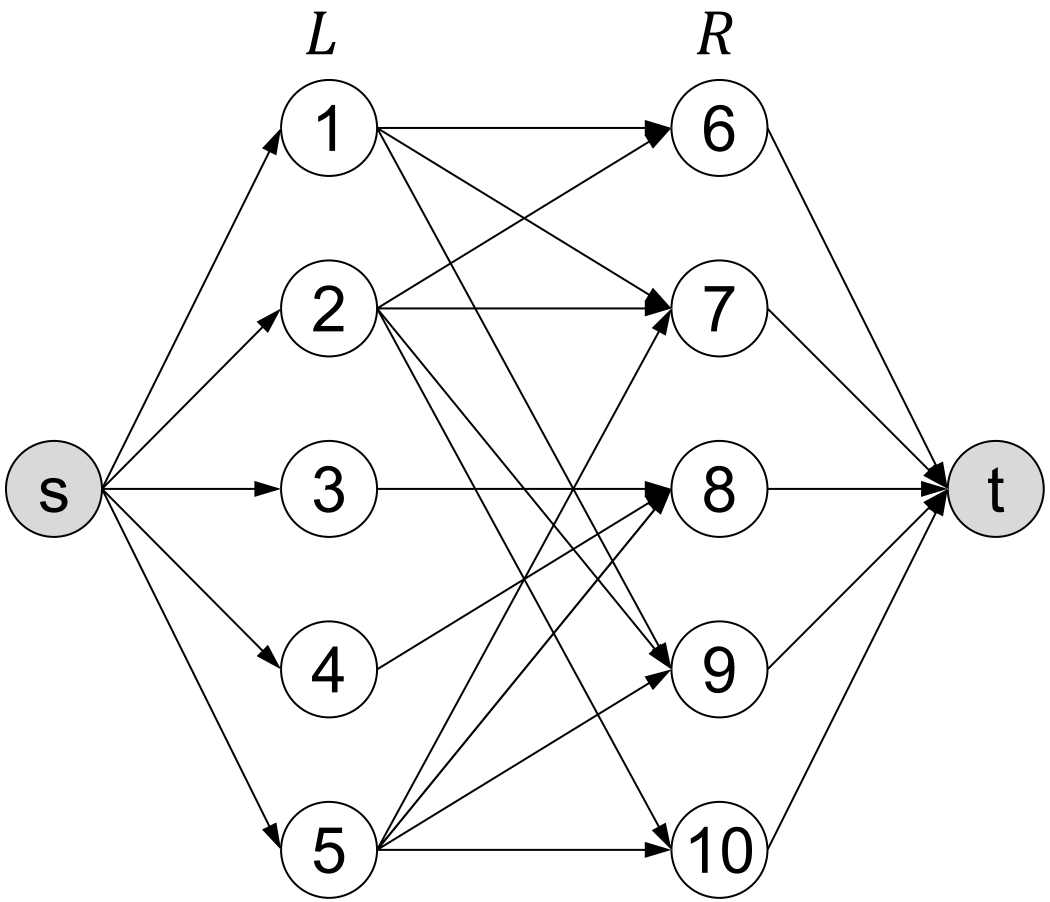
Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity c_{uv} which is the max amount of flow that can pass through it.



Cuts and Flows in directed graphs $G = (V, E)$

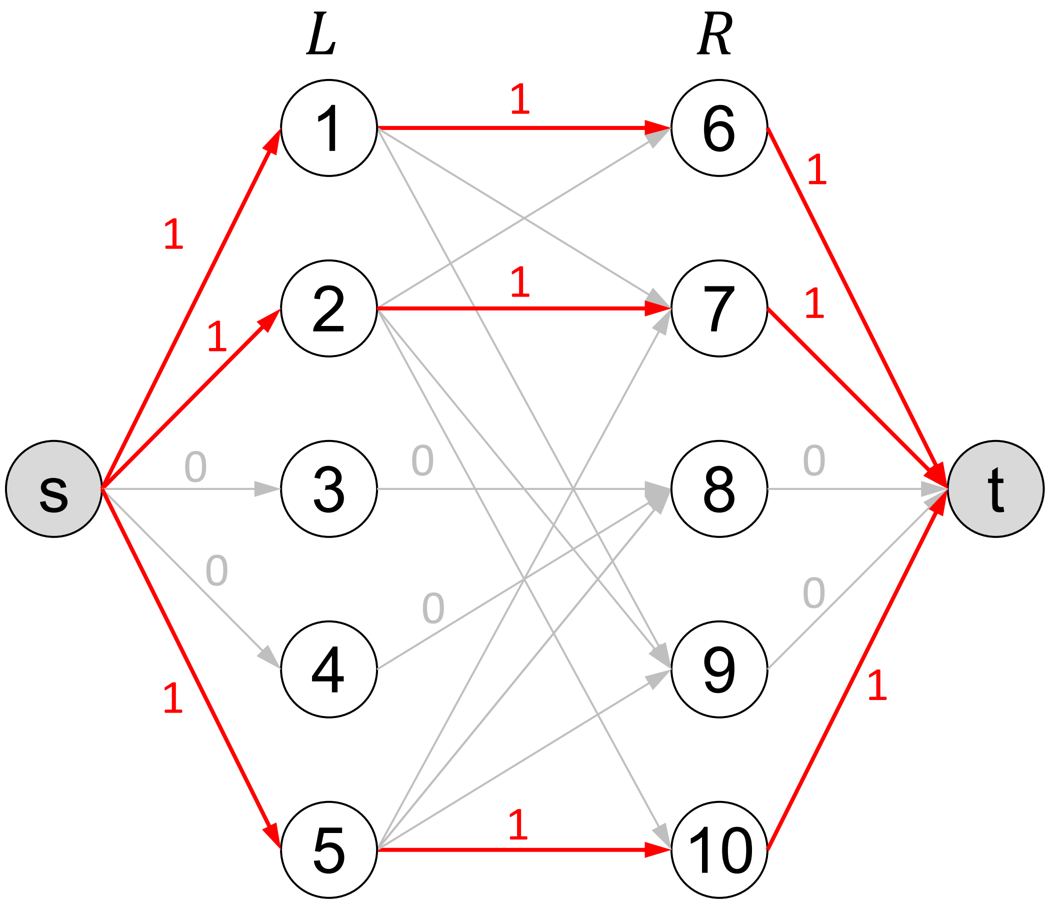
Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.



A **flow** is a mapping of edges to flows $f: E \to \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.

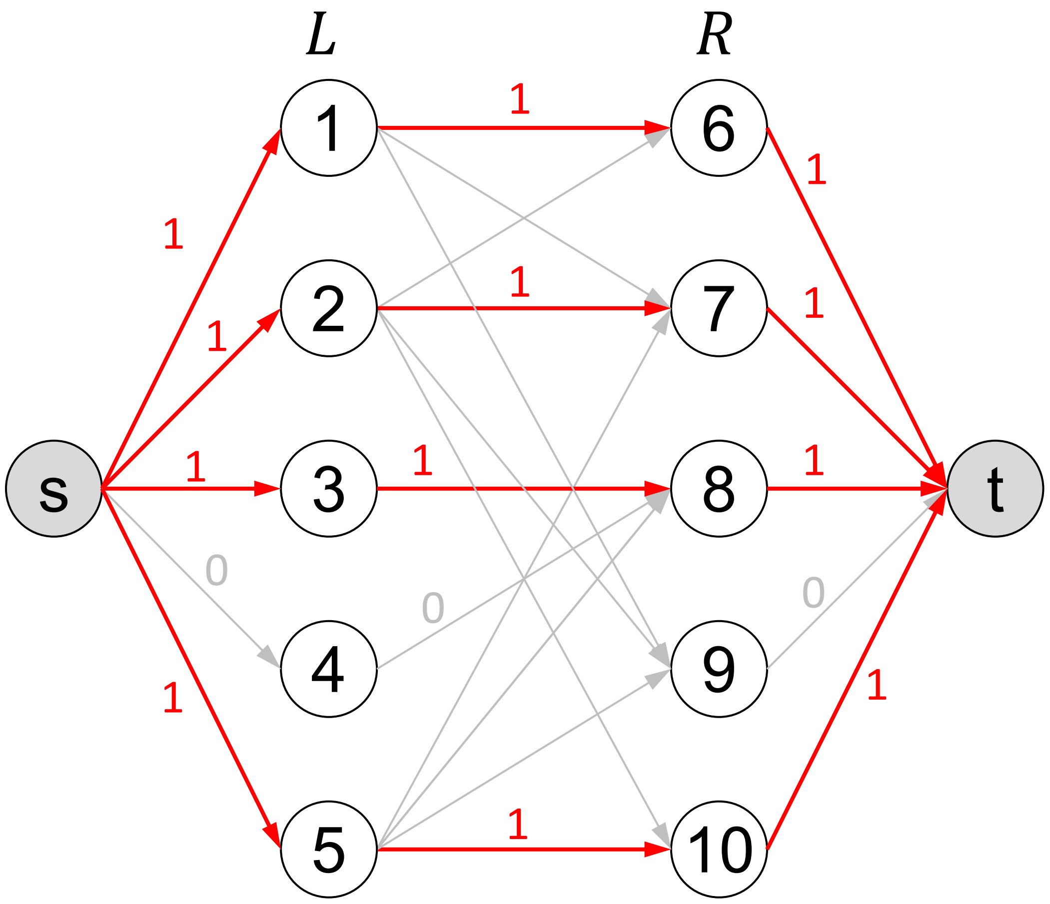


A **flow** is a mapping of edges to flows $f: E \to \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

$|f| = 3$

Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.



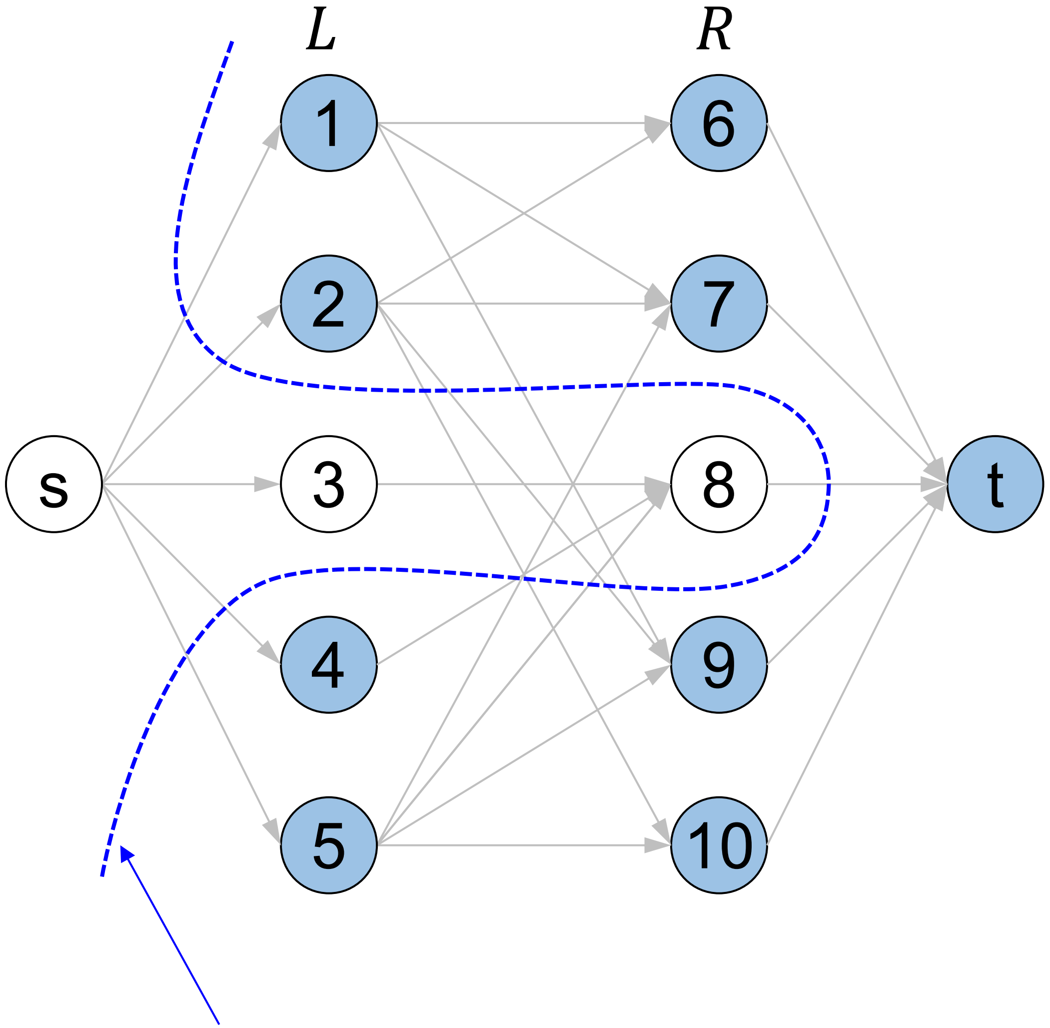
A **flow** is a mapping of edges to flows $f: E \rightarrow \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

$|f| = 4$

Definitions adapted from: https://en.wikipedia.org/wiki/Max-flow_min-cut_theorem

Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.



A **flow** is a mapping of edges to flows $f: E \to \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

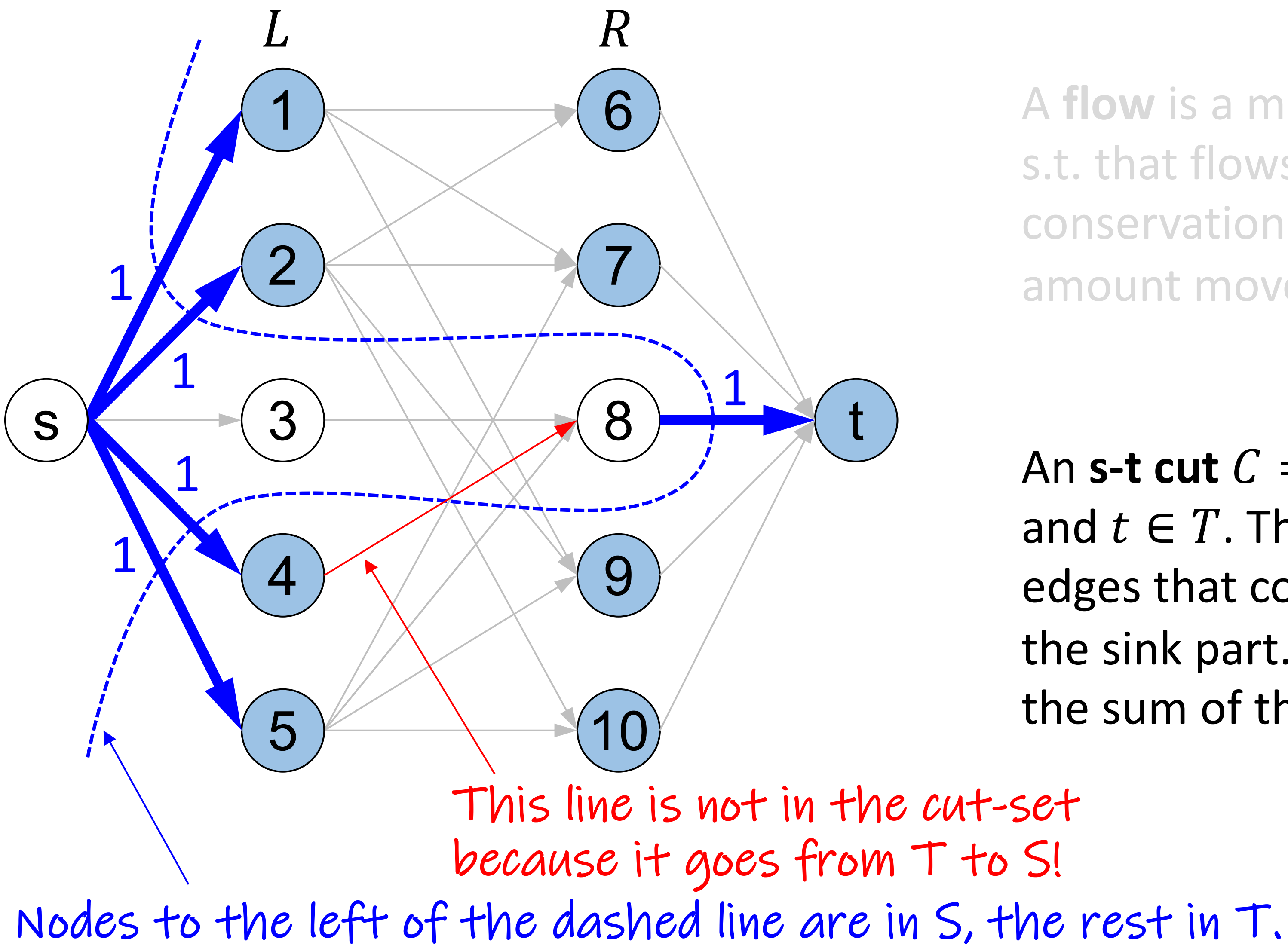
An **s-t cut** $C = (S, T)$ is a partition of V s.t. $s \in S$ and $t \in T$. The **cut-set** X_C of a cut C is the set of edges that connect the source part of the cut to the sink part. The **capacity** $c(S, T)$ of an s-t cut is the sum of the capacities of the edges in its cut-set.

Nodes to the left of the dashed line are in S, the rest in T.

Definitions adapted from: https://en.wikipedia.org/wiki/Max-flow_min-cut_theorem

Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.



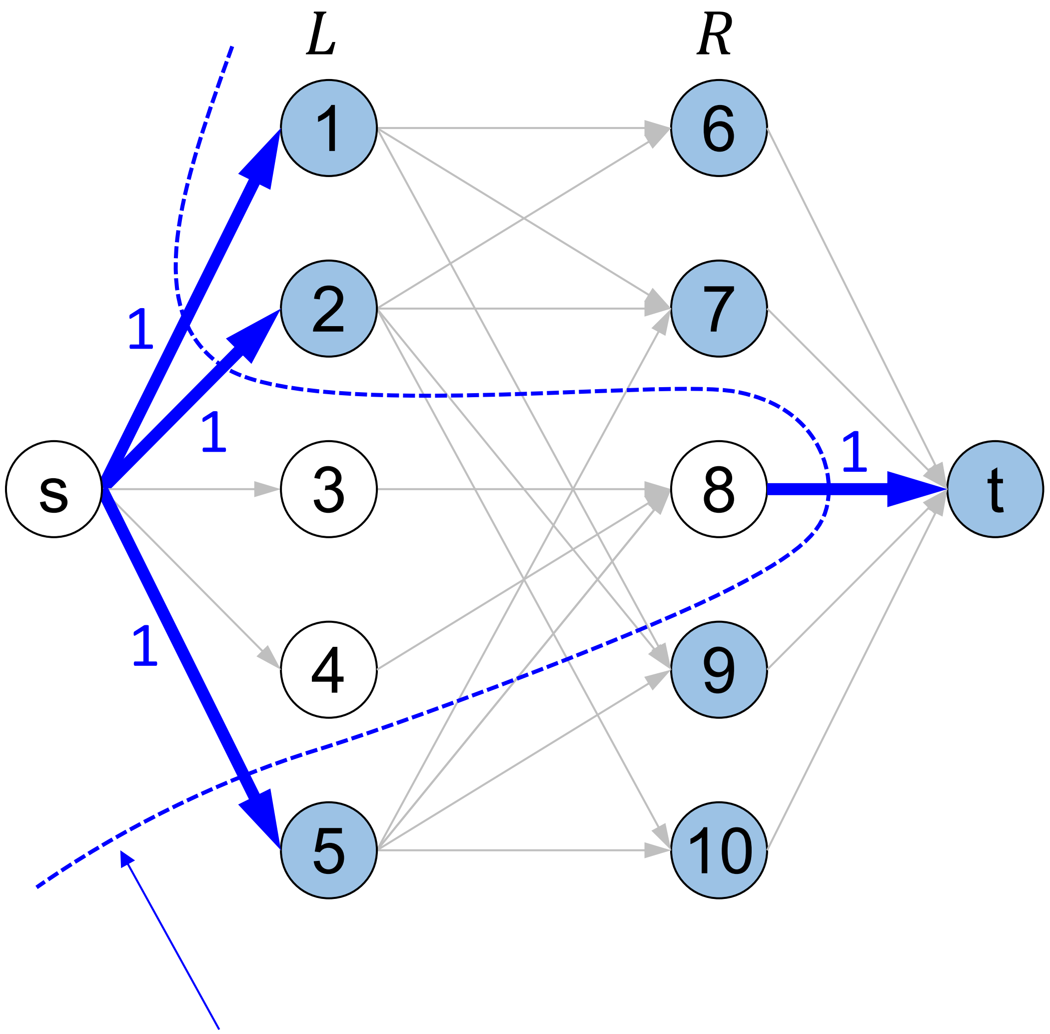
A **flow** is a mapping of edges to flows $f: E \to \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

An **s-t cut** $C = (S, T)$ is a partition of V s.t. $s \in S$ and $t \in T$. The **cut-set** X_C of a cut C is the set of edges that connect the source part of the cut to the sink part. The **capacity** $c(S, T)$ of an s-t cut is the sum of the capacities of the edges in its cut-set.

$c(S, T) = 5$

Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.



A **flow** is a mapping of edges to flows $f: E \to \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

An **s-t cut** $C = (S, T)$ is a partition of V s.t. $s \in S$ and $t \in T$. The **cut-set** X_C of a cut C is the set of edges that connect the source part of the cut to the sink part. The **capacity** $c(S, T)$ of an s-t cut is the sum of the capacities of the edges in its cut-set.

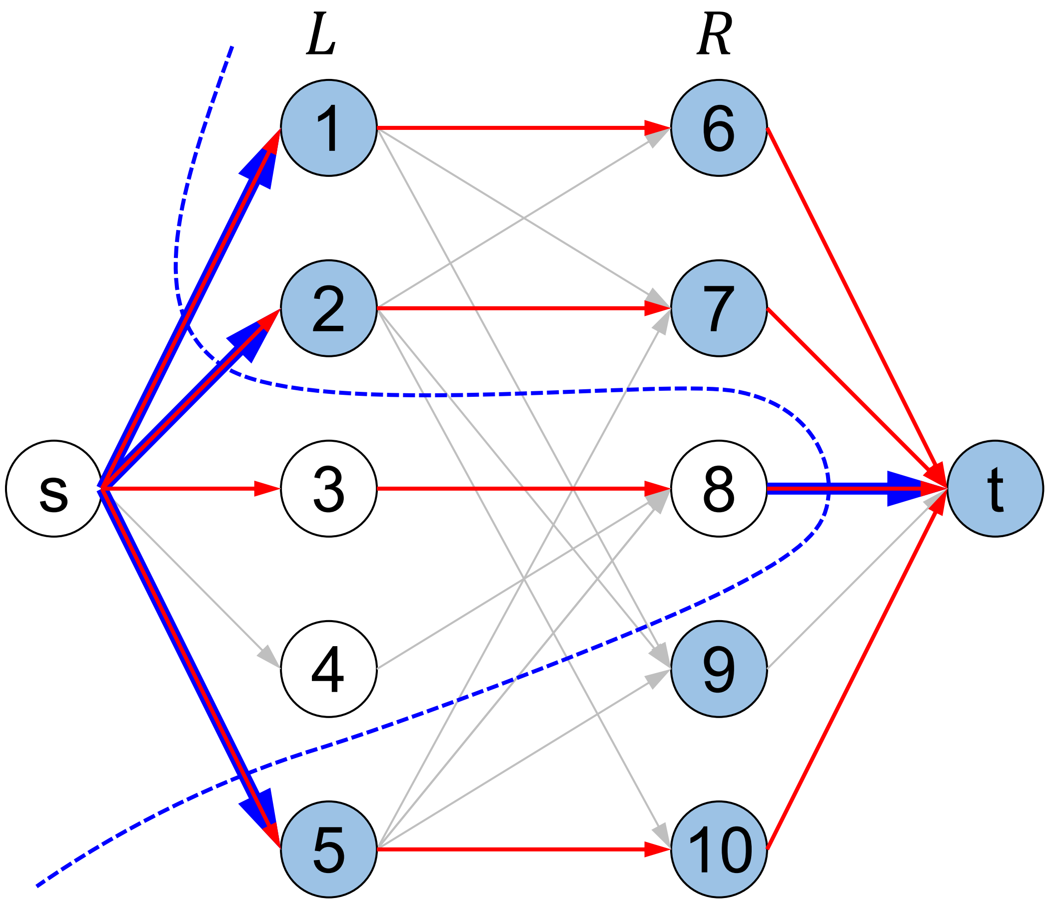
$c(S, T) = 4$

Nodes to the left of the dashed line are in S, the rest in T.

Definitions adapted from: https://en.wikipedia.org/wiki/Max-flow_min-cut_theorem

Cuts and Flows in directed graphs $G = (V, E)$

Each edge (u, v) has a capacity $c_{uv} = 1$ which is the max amount of flow that can pass through it.



A **flow** is a mapping of edges to flows $f: E \rightarrow \mathbb{R}^+$ s.t. that flows obey their capacities $f_{uv} \leq c_{uv}$ and conservation laws. The **value** $|f|$ of a flow is the amount moved from S to T through the network.

$|f| = 4$

An **s-t cut** $C = (S, T)$ is a partition of V s.t. $s \in S$ and $t \in T$. The **cut-set** X_C of a cut C is the set of edges that connect the source part of the cut to the sink part. The **capacity** $c(S, T)$ of an s-t cut is the sum of the capacities of the edges in its cut-set.

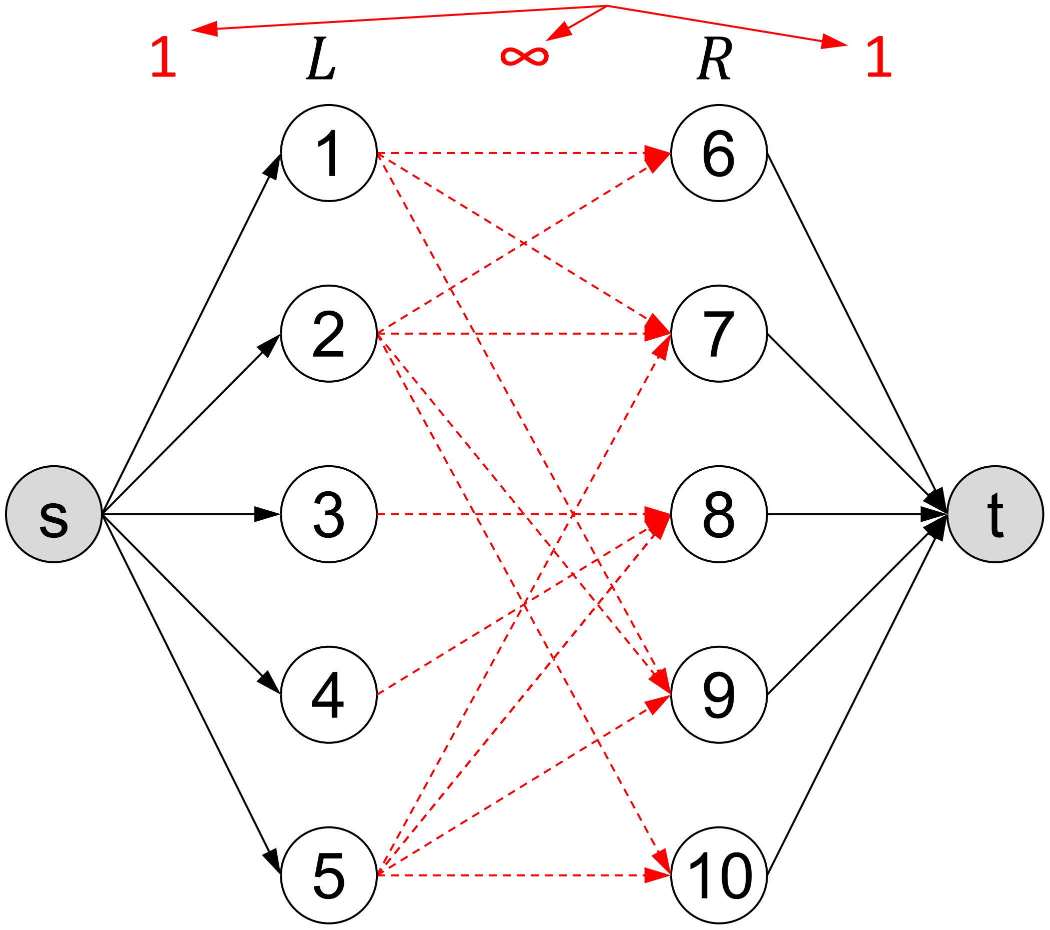
$c(S, T) = 4$

MAX-FLOW MIN-CUT THEOREM.
The maximum value of an s-t flow is equal to the minimum capacity over all s-t cuts.

Definitions adapted from: https://en.wikipedia.org/wiki/Max-flow_min-cut_theorem

Proof Kőnig-Egeváry: outline

Notice the now infinite capacities in the middle:

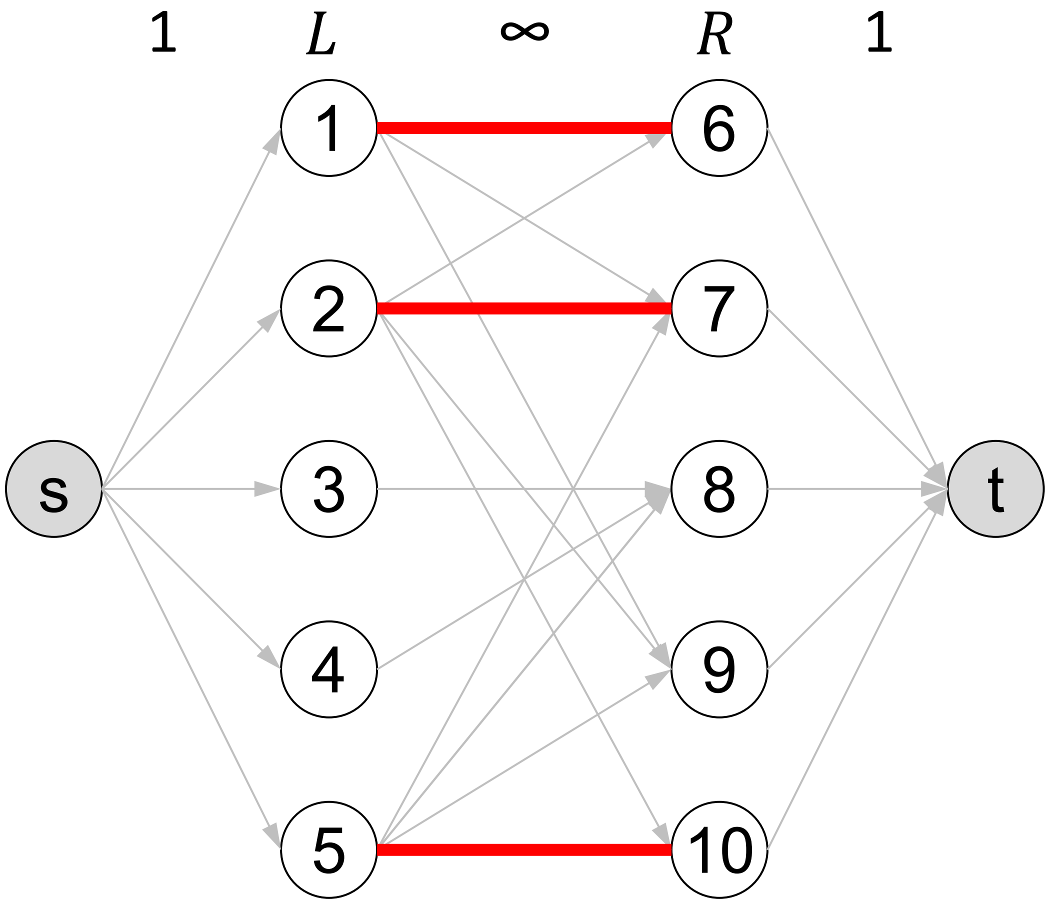


Proof outline:

Consider the flow graph to the left with capacities chosen to avoid a cut between L and R. We will show:

1. every integral flow ⇔ some matching
2. every (finite capacity) cut ⇔ some VC
3. Then we know that max matching = min VC, from the max-flow min-cut theorem

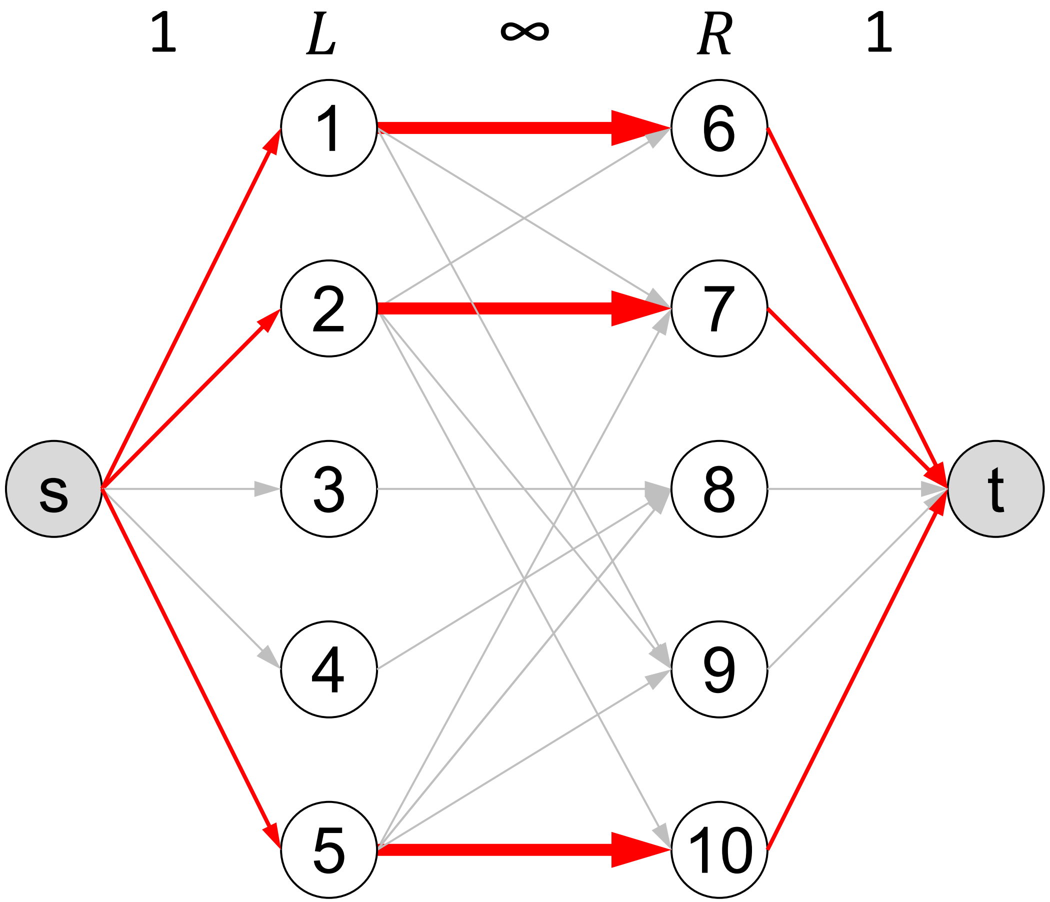
Proof Kőnig-Egeváry 1: matching = flow



1. A matching of size x corresponds to an integral flow of same value.

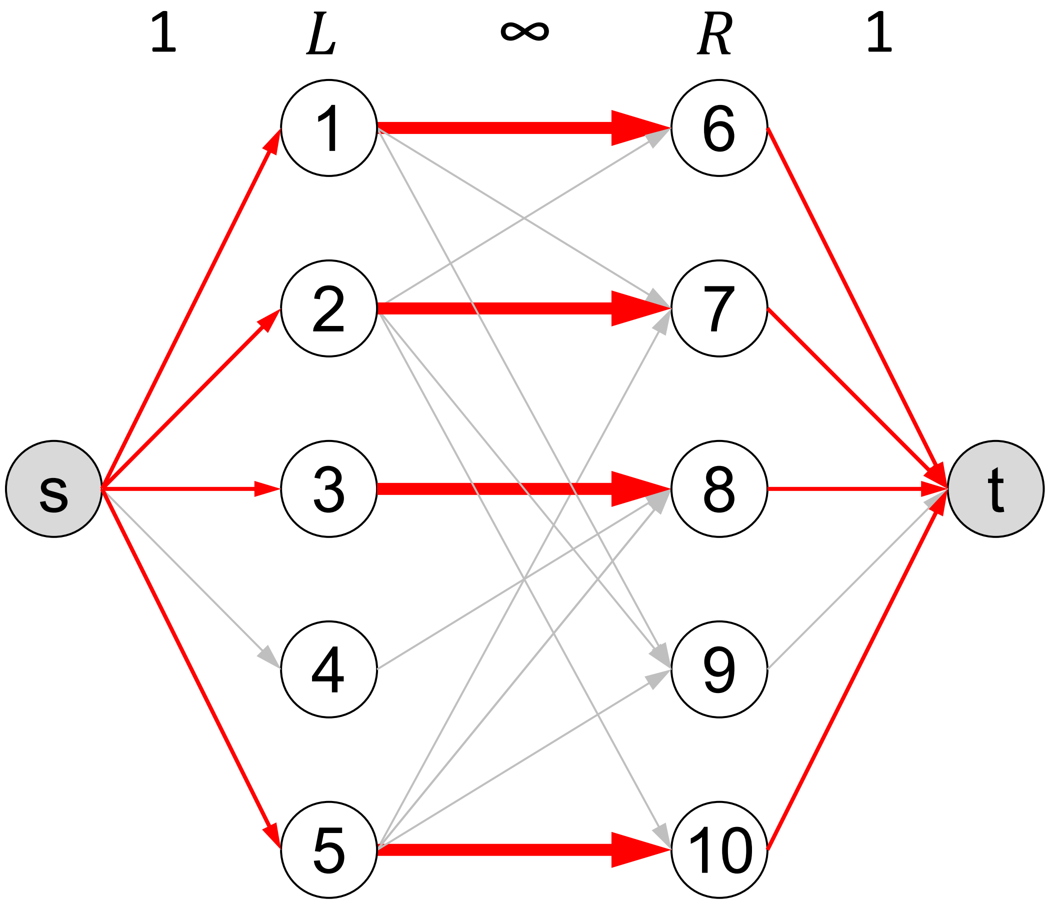
#VC = 5

Proof Kőnig-Egeváry 1: matching = flow



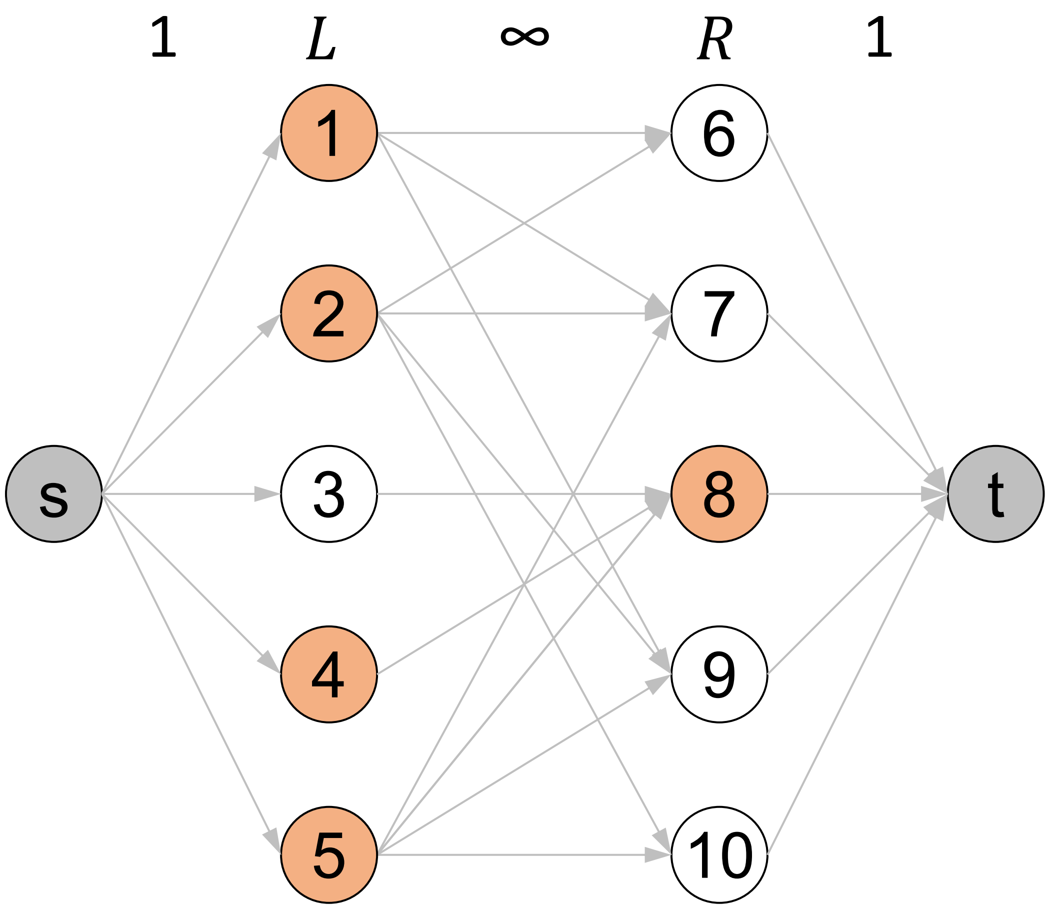
1. A matching of size x corresponds to an integral flow of same value.

Proof Kőnig-Egeváry 1: matching = flow



1. A matching of size x corresponds to an integral flow of same value.

Proof Kőnig-Egeváry 2: VC = cut



1. A matching of size x corresponds to an integral flow of same value.

2. Any VC of size x defines a cut of same capacity.

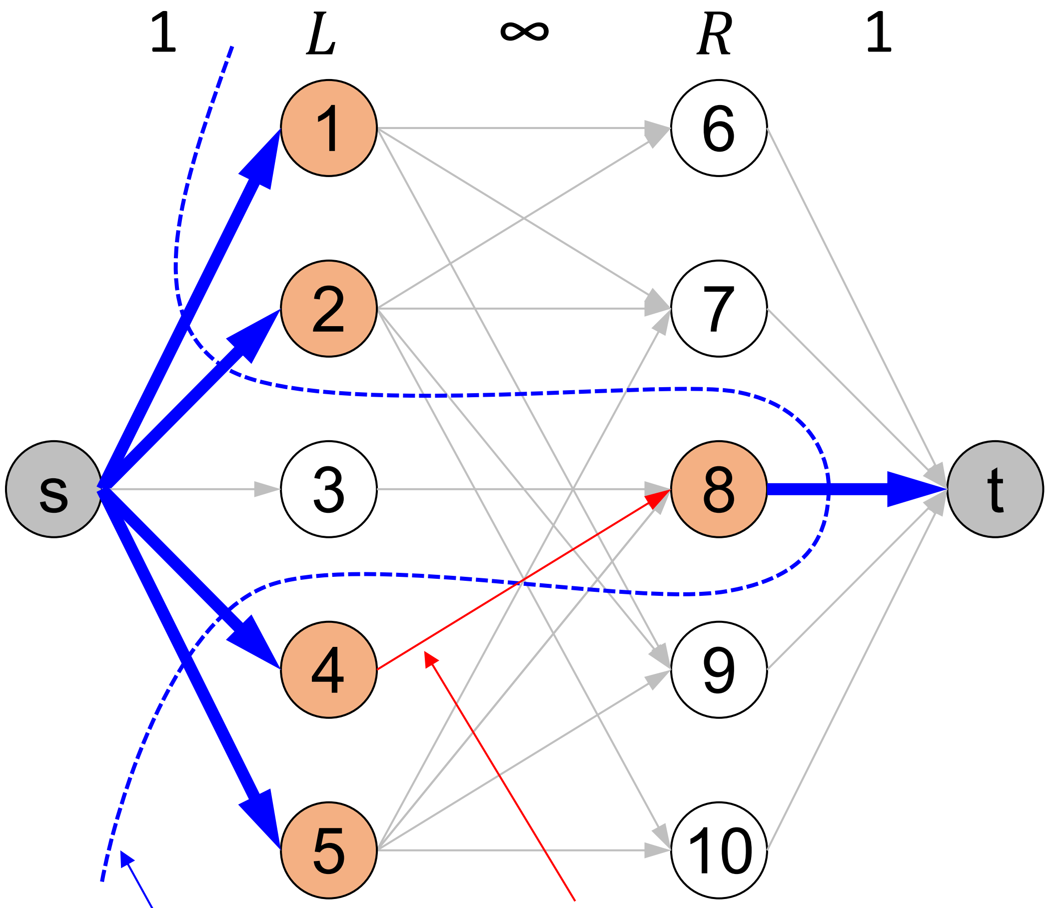
Let C be the VC, $C(L) = C \cap L$, $C(R) = C \cap R$.

Then define: $S \coloneqq \{s\} \cup \big(L - C(L)\big) \cup C(R)$

$T \coloneqq \{t\} \cup \big(R - C(R)\big) \cup C(L)$

#VC = 5

Proof Kőnig-Egeváry 2: VC = cut



This line is not in the cut-set because it goes from T to S!

Nodes to the left of the dashed line are in S, the rest in T

1. A matching of size x corresponds to an integral flow of same value.

2. Any VC of size x defines a cut of same capacity.

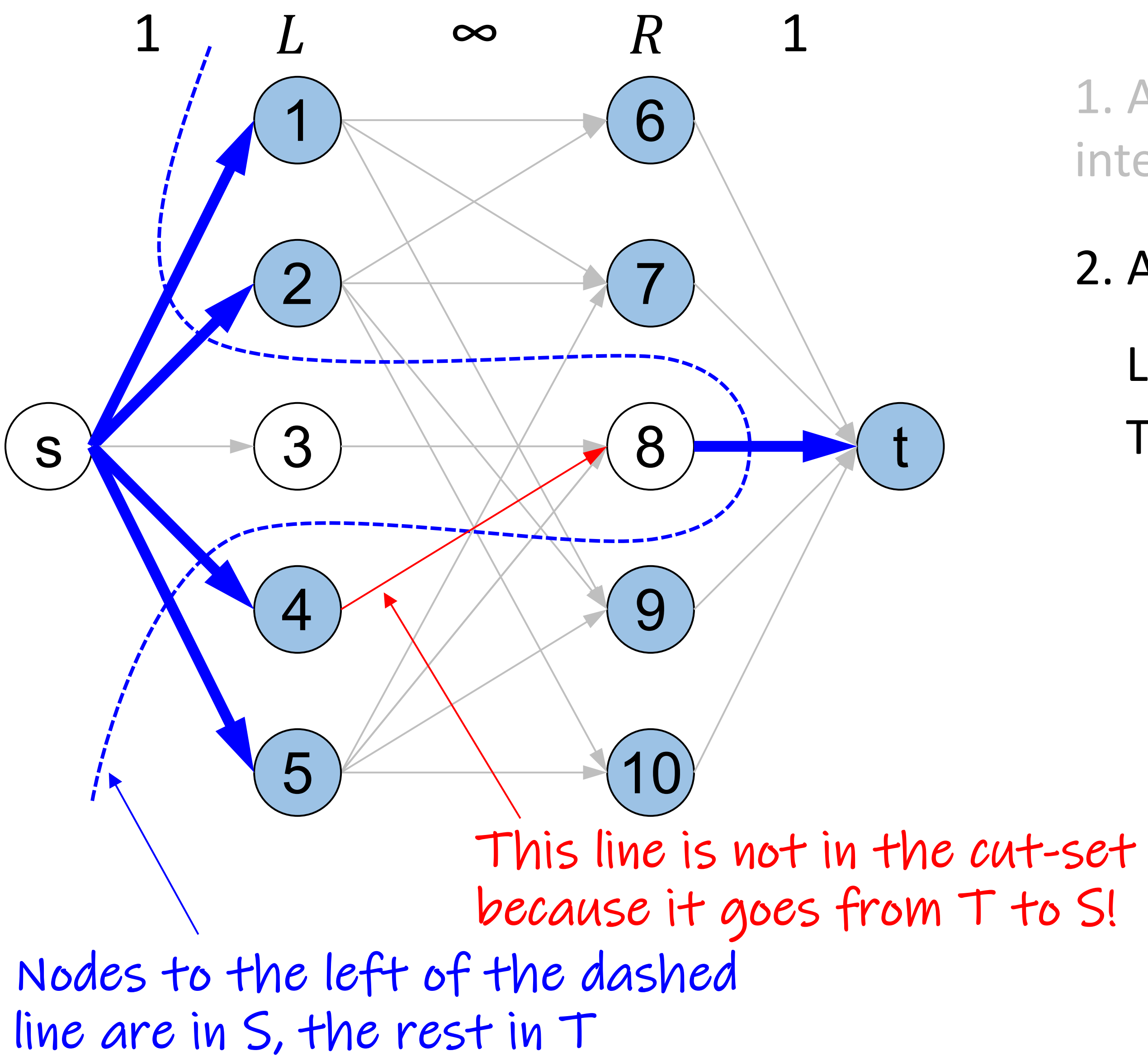
Let C be the VC, $C(L) = C \cap L$, $C(R) = C \cap R$.

Then define: $S \coloneqq \{s\} \cup \big(L - C(L)\big) \cup C(R)$

$T \coloneqq \{t\} \cup \big(R - C(R)\big) \cup C(L)$

#VC = $c(S,T) = 5$

Proof Kőnig-Egeváry 2: VC = cut



1. A matching of size x corresponds to an integral flow of same value.

2. Any VC of size x defines a cut of same capacity.

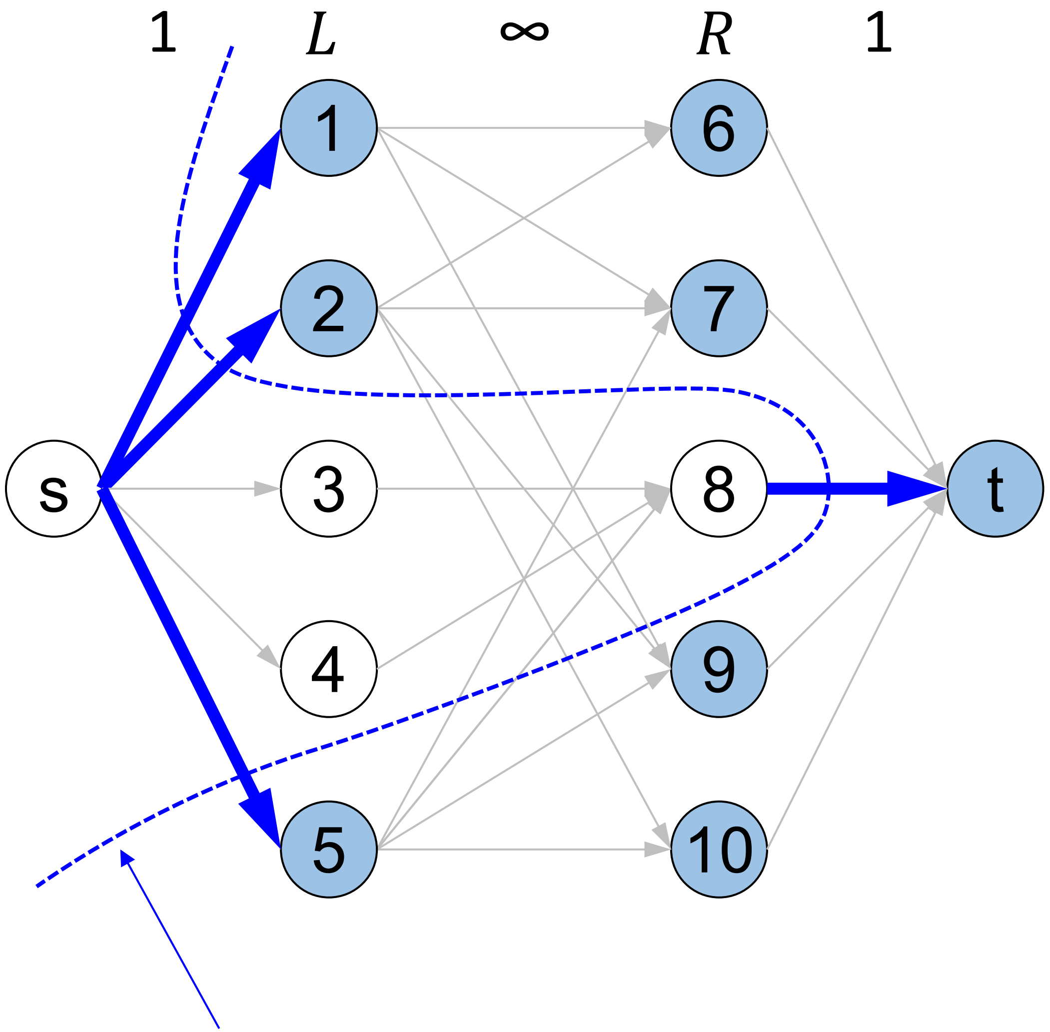
Let C be the VC, $C(L) = C \cap L$, $C(R) = C \cap R$.

Then define: $S \coloneqq \{s\} \cup \big(L - C(L)\big) \cup C(R)$

$T \coloneqq \{t\} \cup \big(R - C(R)\big) \cup C(L)$

#VC = $c(S, T) = 5$

Proof Kőnig-Egeváry 2: VC = cut



Nodes to the left of the dashed line are in S, the rest in T

1. A matching of size x corresponds to an integral flow of same value.

2. Any VC of size x defines a cut of same capacity.

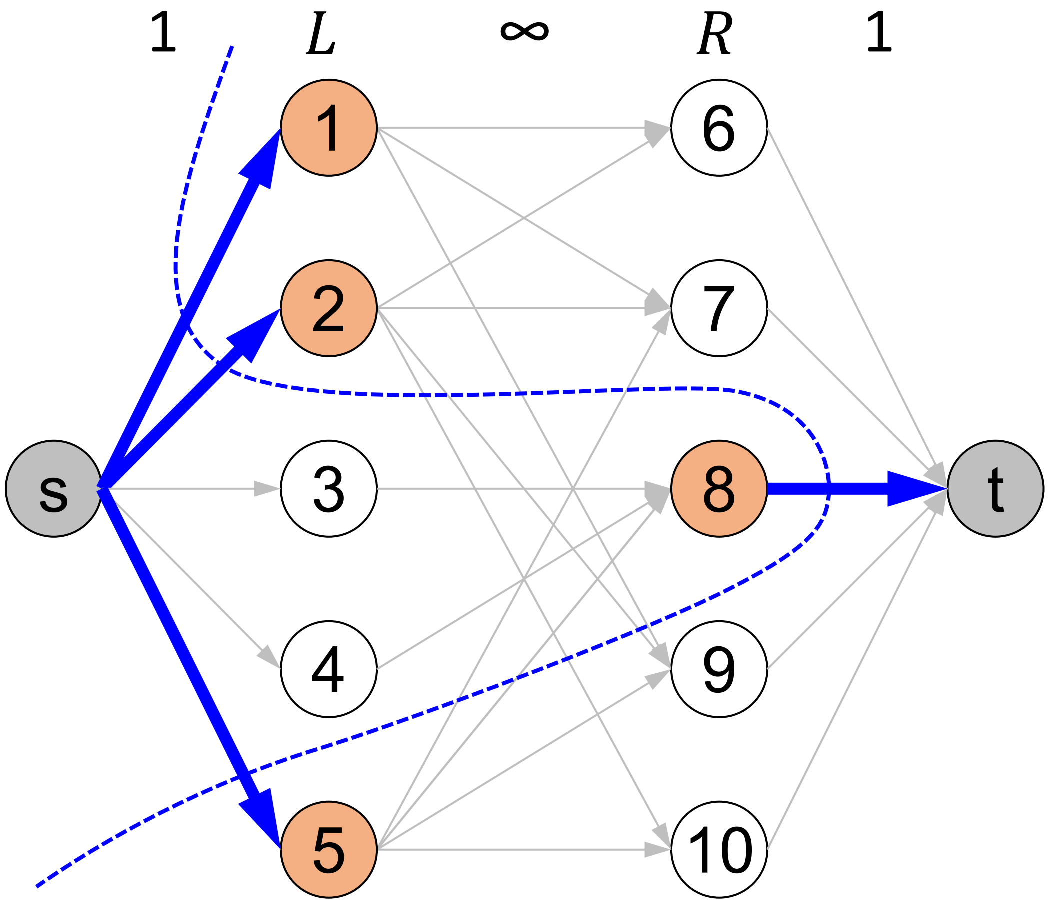
Let C be the VC, $C(L) = C \cap L$, $C(R) = C \cap R$.

Then define: $S \coloneqq \{s\} \cup \big(L - C(L)\big) \cup C(R)$

$T \coloneqq \{t\} \cup \big(R - C(R)\big) \cup C(L)$

$\#VC = c(S, T) = 4$

Proof Kőnig-Egeváry 2: VC = cut



1. A matching of size x corresponds to an integral flow of same value.

2. Any VC of size x defines a cut of same capacity.

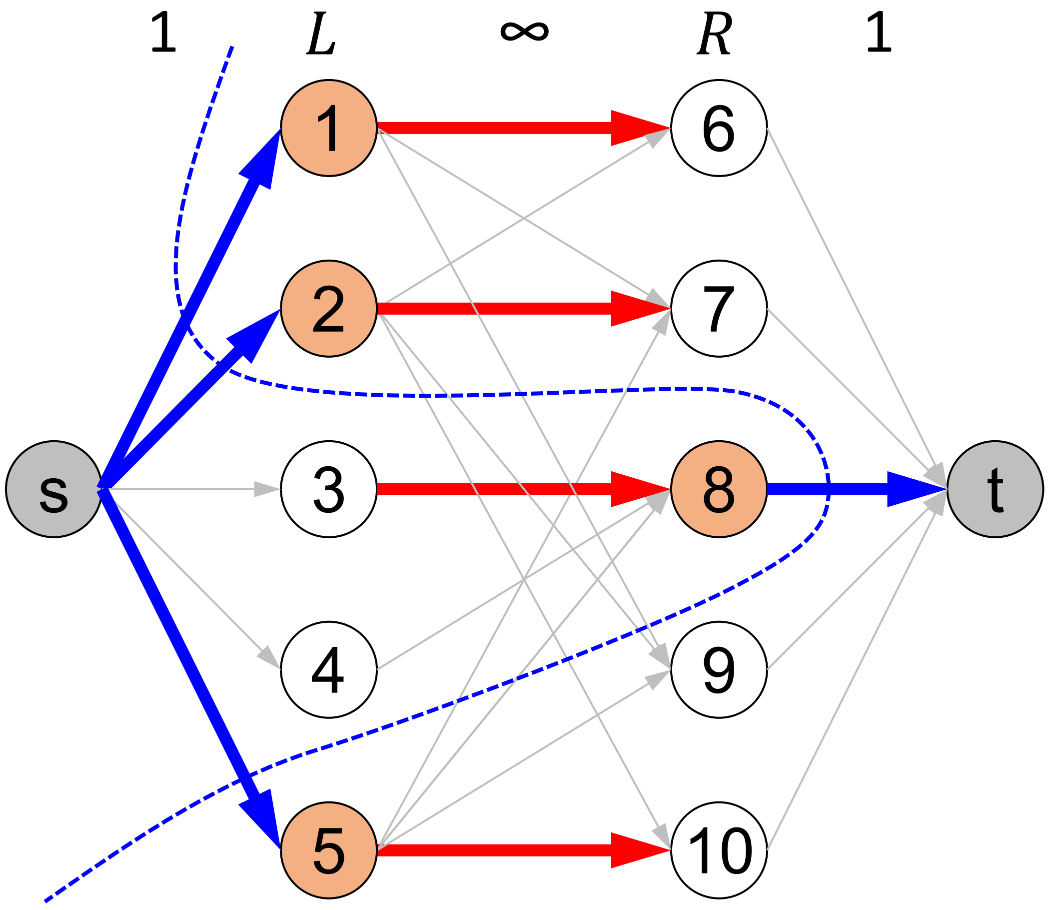
Let C be the VC, $C(L) = C \cap L$, $C(R) = C \cap R$.

Then define: $S \coloneqq \{s\} \cup \big(L - C(L)\big) \cup C(R)$

$T \coloneqq \{t\} \cup \big(R - C(R)\big) \cup C(L)$

#VC = $c(S,T) = 4$

Proof Kőnig-Egeváry 3: max-flow = min-cut $\Rightarrow$ max matching = min VC



1. A matching of size x corresponds to an integral flow of same value.

2. Any VC of size x defines a cut of same capacity.

   Let C be the VC, $C(L) = C \cap L$, $C(R) = C \cap R$.

   Then define: $S \coloneqq \{s\} \cup \big(L - C(L)\big) \cup C(R)$
   $T \coloneqq \{t\} \cup \big(R - C(R)\big) \cup C(L)$

3. Since max flow = min cut, therefore also max matching = min VC

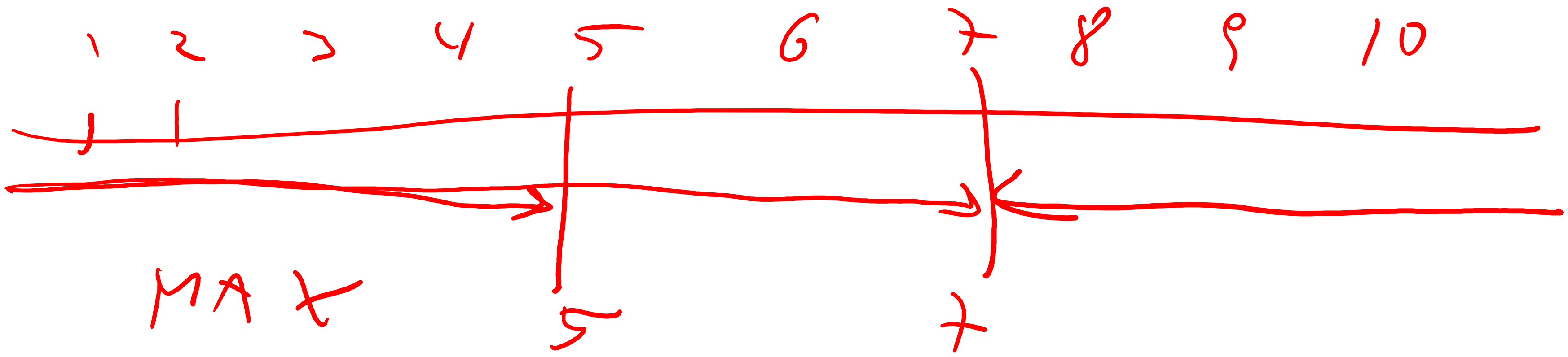
#matching = $|f|$ = 4
#VC = $c(S,T)$ = 4

LP (Linear Programming) and duality gaps

Dual Optimization Problem

(e.g, max independent set) (e.g., min edge cover)

- A maximization problem **M** and a minimization problem **N**, defined on the same instances (such as graphs, constraints) s.t.:
  1. for every candidate solution *M* to **M** and every candidate solution *N* to **N**, the value of *M* is less than or equal to the value of *N*
  2. obtaining candidate solutions *M* and *N* that have the same value proves that *M* and *N* are optimal solutions for that instance.



A quick primer on Duality in Linear Programming

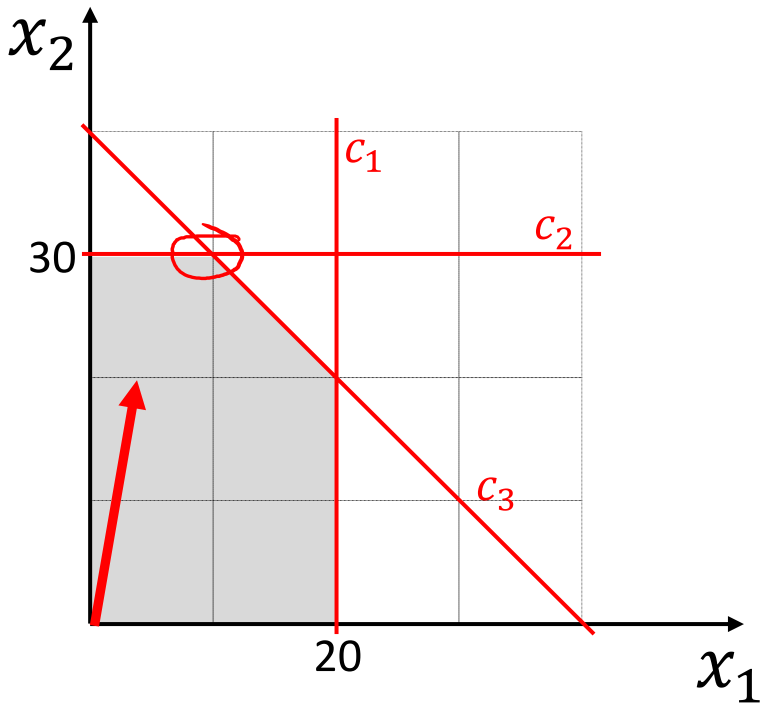
$$\max\ 1x_1 + 6x_2$$

c_1: $x_1 \le 20$

c_2: $x_2 \le 30$

c_3: $x_1 + x_2 \le 40$

$x_1, x_2 \ge 0$



Assume I give you the solution $(x_1, x_2) = (10,30)$ with objective value = 190.
How could you prove it is indeed the maximum feasible value?

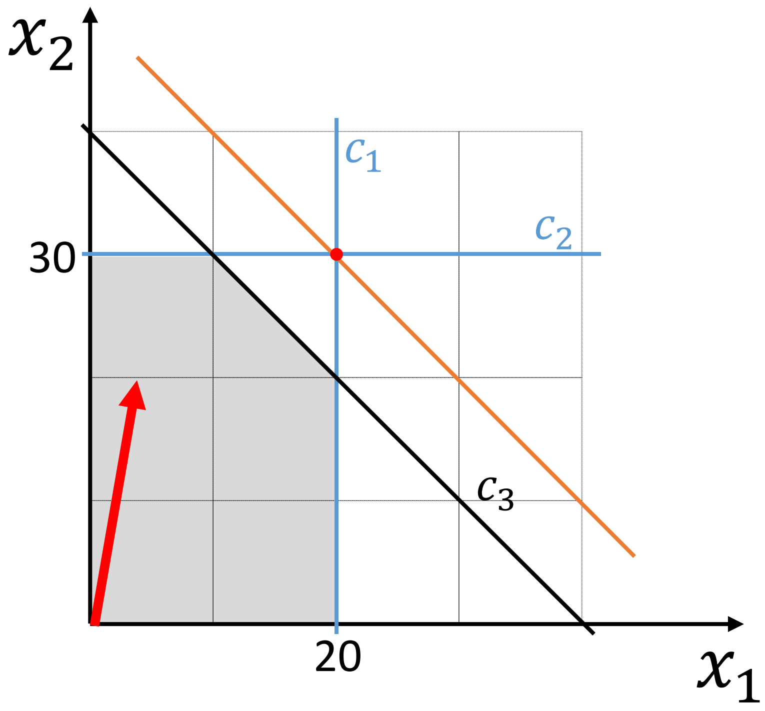
?

A quick primer on Duality in Linear Programming

non-negative multipliers!

$$\max\ 1x_1 + 6x_2$$

$$\begin{array}{lrl} c_1: & x_1 \leq 20 & \times 1 \\ c_2: & x_2 \leq 30 & \times 1 \\ c_3: & x_1 + x_2 \leq 40 & \times 0 \\ & x_1, x_2 \geq 0 & \\ \hline & 1x_1 + 1x_2 \leq 50 & \end{array}$$



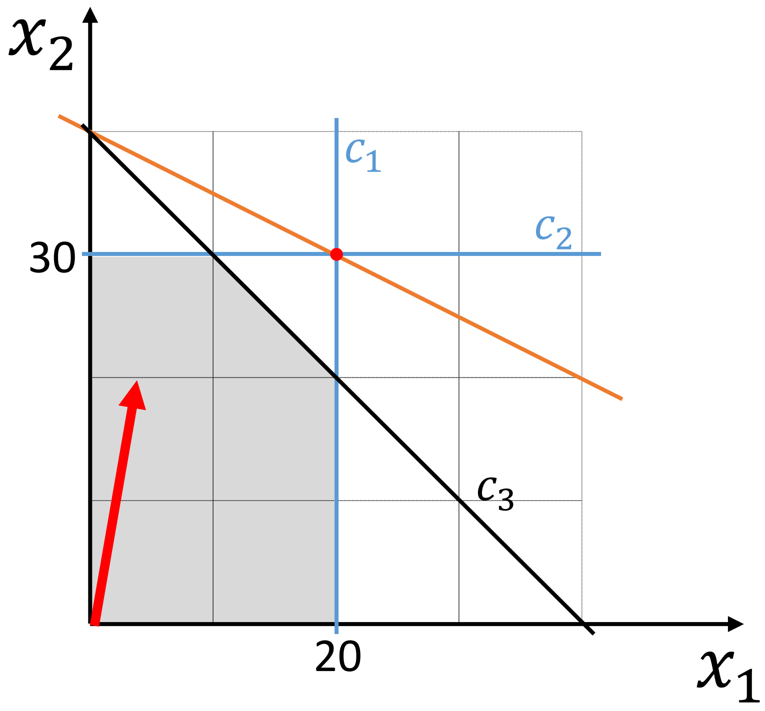
Assume I give you the solution $(x_1, x_2) = (10,30)$ with objective value = 190.
How could you prove it is indeed the maximum feasible value?

A quick primer on Duality in Linear Programming

non-negative multipliers!

$$\max\ 1x_1 + 6x_2$$

$$\begin{array}{lrl}
c_1: & x_1 \leq 20 & \times 1 \\
c_2: & x_2 \leq 30 & \times 2 \\
c_3: & x_1 + x_2 \leq 40 & \times 0 \\
 & x_1, x_2 \geq 0 & \\
\hline
 & 1x_1 + 2x_2 \leq 80 &
\end{array}$$



Assume I give you the solution $(x_1, x_2) = (10,30)$ with objective value = 190.
How could you prove it is indeed the maximum feasible value?

A quick primer on Duality in Linear Programming

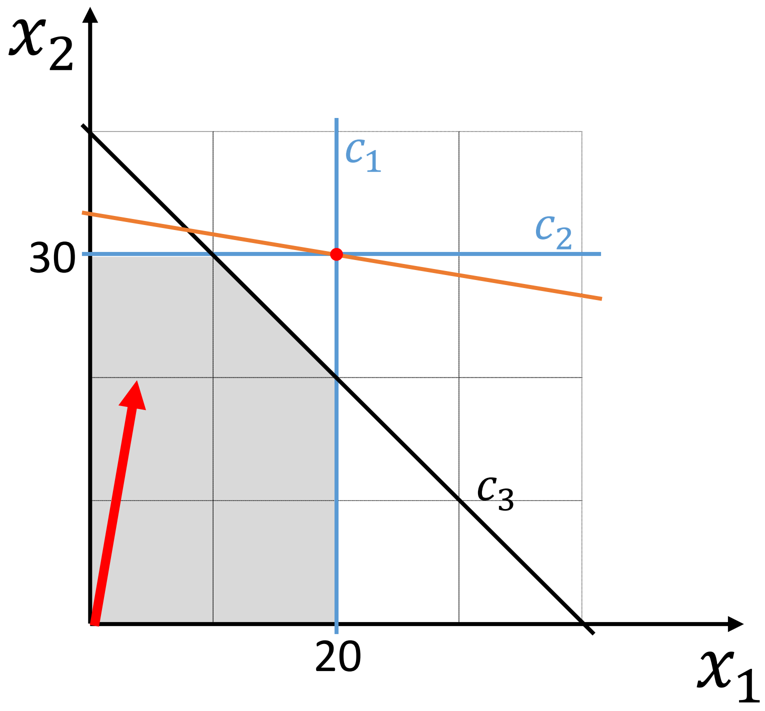
non-negative multipliers!

$$\max\ 1x_1 + 6x_2$$

$$\begin{aligned}
c_1: \quad & x_1 \le 20 & \times 1 \\
c_2: \quad & x_2 \le 30 & \times 6 \\
c_3: \quad & x_1 + x_2 \le 40 & \times 0 \\
& x_1, x_2 \ge 0 &
\end{aligned}$$

$$1x_1 + 6x_2 \le 200$$

upper bound to the objective function!



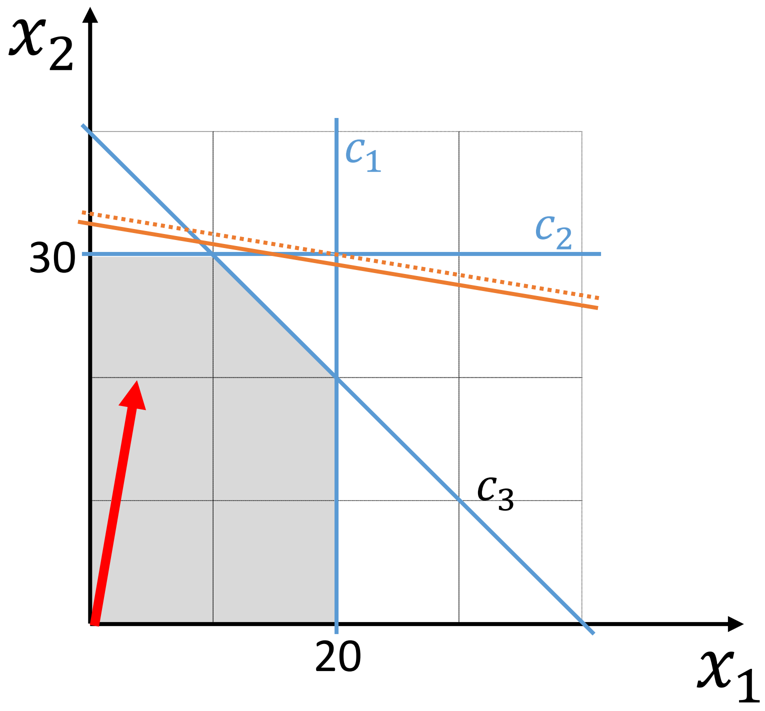
Assume I give you the solution $(x_1, x_2) = (10{,}30)$ with objective value = 190. How could you prove it is indeed the maximum feasible value?

A quick primer on Duality in Linear Programming

non-negative multipliers!

$$\max\ 1x_1 + 6x_2$$

$$\begin{array}{lrl} c_1: & x_1 \le 20 & \times 0.5 \\ c_2: & x_2 \le 30 & \times 5.5 \\ c_3: & x_1 + x_2 \le 40 & \times 0.5 \\ & x_1, x_2 \ge 0 & \\ \hline & 1x_1 + 6x_2 \le 195 & \end{array}$$



upper bound to the objective function!

Assume I give you the solution $(x_1, x_2) = (10{,}30)$ with objective value = 190.
How could you prove it is indeed the maximum feasible value?

A quick primer on Duality in Linear Programming

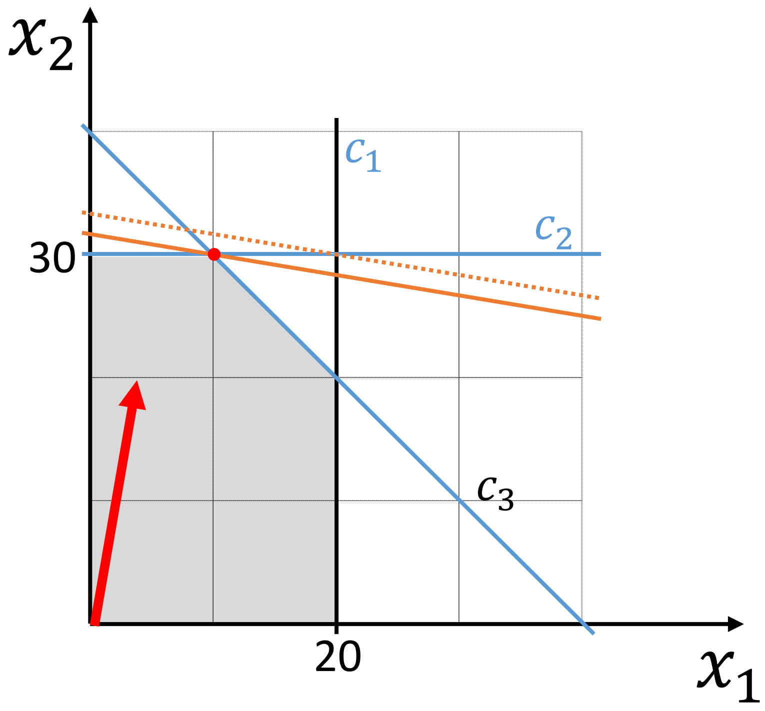
non-negative multipliers!

$$\max\ 1x_1 + 6x_2$$

$$\begin{array}{lrl} c_1: & x_1 \leq 20 & \times 0 \\ c_2: & x_2 \leq 30 & \times 5 \\ c_3: & x_1 + x_2 \leq 40 & \times 1 \\ & x_1, x_2 \geq 0 & \\ \hline & 1x_1 + 6x_2 \leq \boxed{190} & \end{array}$$

certificate of optimality

minimum upper bound to the objective function!



Assume I give you the solution $(x_1, x_2) = (10{,}30)$ with objective value = 190. How could you prove it is indeed the maximum feasible value?

A quick primer on Duality in Linear Programming

non-negative multipliers!

$$\max\ 1x_1 + 6x_2$$

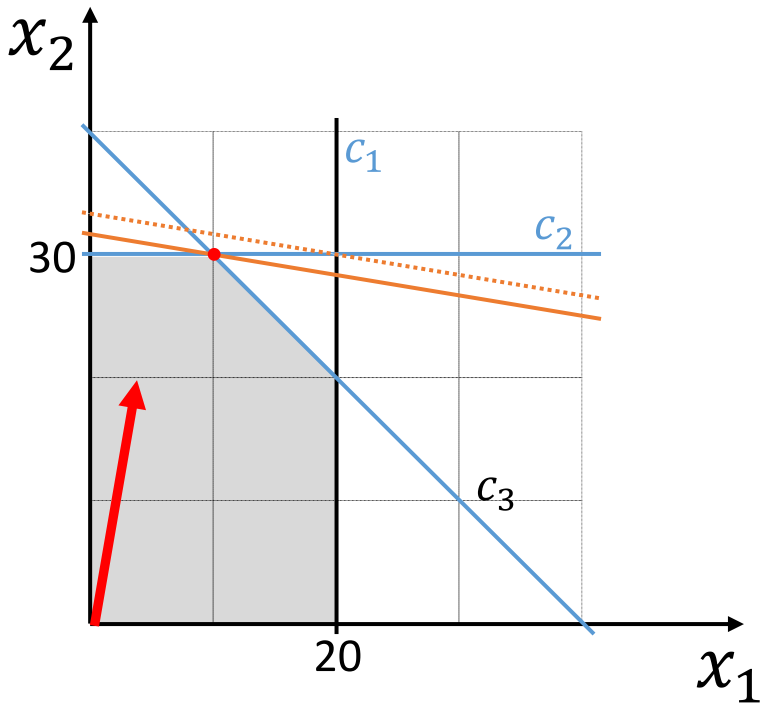
$$\begin{array}{lrl} c_1: & x_1 \le 20 & \times y_1 \\ c_2: & x_2 \le 30 & \times y_2 \\ c_3: & x_1 + x_2 \le 40 & \times y_3 \\ & x_1, x_2 \ge 0 & \end{array}$$

$$1x_1 + 6x_2 \le 20y_1 + 30y_2 + 40y_3$$

$\ge$ $\ge$

$y_1 + y_3$ $y_2 + y_3$

find a convex combination of the constraints
to get the minimum upper bound to the objective function!



Assume I give you the solution $(x_1, x_2) = (10, 30)$ with objective value = 190.
How could you prove it is indeed the maximum feasible value?

Example taken from: Dasgupta, Papadimitriou, Vazirani. Algorithms. 2006. http://algorithmics.lsi.upc.edu/docs/Dasgupta-Papadimitriou-Vazirani.pdf
Wolfgang Gatterbauer. Principles of scalable data management: https://northeastern-datalab.github.io/cs7240/

A quick primer on Duality in Linear Programming

$$\max\ 1x_1 + 6x_2$$

$$x_1 \leq 20 \qquad \times y_1$$

$$\geq \qquad \geq \qquad x_2 \leq 30 \qquad \times y_2$$

$$x_1 + x_2 \leq 40 \qquad \times y_3$$

$$x_1, x_2 \geq 0$$

$$(y_1 + y_3)x_1 + (y_2 + y_3)x_2 \leq 20y_1 + 30y_2 + 40y_3$$

$$\min\ 20y_1 + 30y_2 + 40y_3$$

$$y_1 + y_3 \geq 1$$

$$y_2 + y_3 \geq 6$$

$$y_1, y_2, y_3 \geq 0$$

Primal solution $(x_1, x_2) = (10, 30)$

Dual solution $(y_1, y_2, y_3) = (0,5,1)$

Example taken from: Dasgupta, Papadimitriou, Vazirani. Algorithms. 2006. http://algorithmics.lsi.upc.edu/docs/Dasgupta-Papadimitriou-Vazirani.pdf

A quick primer on Duality in Linear Programming

$$\max\ 1x_1 + 6x_2$$

$$\begin{aligned} x_1 &\le 20 && \times y_1 \\ x_2 &\le 30 && \times y_2 \\ x_1 + x_2 &\le 40 && \times y_3 \\ x_1, x_2 &\ge 0 \end{aligned}$$

$\ge$ $\ge$

$$(y_1{+}y_3)x_1 + (y_2{+}y_3)x_2 \le 20y_1 + 30y_2 + 40y_3$$

$$1x_1 + 6x_2 \le 20{\cdot}0 + 30{\cdot}5 + 40{\cdot}1$$

$$1{\cdot}10 + 6{\cdot}30 \le 190$$

$$\min\ 20y_1 + 30y_2 + 40y_3$$

$$\begin{aligned} y_1 + y_3 &\ge 1 && \times x_1 \\ y_2 + y_3 &\ge 6 && \times x_2 \\ y_1, y_2, y_3 &\ge 0 \end{aligned}$$

$\le$ $\le$ $\le$

$$x_1y_1 + x_2y_2 + (x_1{+}x_2)y_3 \ge 1x_1 + 6x_2$$

$$10y_1 + 30y_2 + (10{+}30)y_3 \ge 1{\cdot}10 + 6{\cdot}30$$

$$10{\cdot}y_1 + 30{\cdot}y_2 + 40{\cdot}y_3 \ge 190$$

Primal solution $(x_1, x_2) = (10, 30)$

Dual solution $(y_1, y_2, y_3) = (0,5,1)$

Example taken from: Dasgupta, Papadimitriou, Vazirani. Algorithms. 2006. http://algorithmics.lsi.upc.edu/docs/Dasgupta-Papadimitriou-Vazirani.pdf

LP in Canonical Form and Matrix-vector notation

$$\max\ 1x_1 + 6x_2$$
$$x_1 \leq 20$$
$$x_2 \leq 30$$
$$x_1 + x_2 \leq 40$$
$$x_1, x_2 \geq 0$$

$$\min\ 20y_1 + 30y_2 + 40y_3$$
$$y_1 + y_3 \geq 1$$
$$y_2 + y_3 \geq 6$$
$$y_1, y_2, y_3 \geq 0$$

Canonical form:

$$\max\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$$

objective vector

$$\min\ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$$

constraint matrix

$$\mathbf{A}\boldsymbol{x} \leq \boldsymbol{b}$$

constraint vector

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{y} \geq \boldsymbol{c}$$

$$\boldsymbol{x} \geq 0$$

$$\boldsymbol{y} \geq 0$$

A quick primer on Duality in Linear Programming

$$\max \begin{pmatrix}1\\6\end{pmatrix}^{\mathrm{T}} \boldsymbol{x}$$

$$\begin{pmatrix}1 & 0\\0 & 1\\1 & 1\end{pmatrix}\boldsymbol{x} \leq \begin{pmatrix}20\\30\\40\end{pmatrix}$$

$$\boldsymbol{x} \geq 0$$

$$\min \begin{pmatrix}20\\30\\40\end{pmatrix}^{\mathrm{T}} \boldsymbol{y}$$

$$\begin{pmatrix}1 & & 1\\ & 1 & 1\end{pmatrix}\boldsymbol{y} \geq \begin{pmatrix}1\\6\end{pmatrix}$$

$$\boldsymbol{y} \geq 0$$

$$\max\ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$$

$$\mathbf{A}\boldsymbol{x} \leq \boldsymbol{b}$$

$$\boldsymbol{x} \geq 0$$

$$\min\ \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}$$

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{y} \geq \boldsymbol{c}$$

$$\boldsymbol{y} \geq 0$$

A quick primer on Duality in Linear Programming

Figure 7.10 A generic primal LP in matrix-vector form, and its dual.

Primal LP:

$$\begin{aligned} \max\ \mathbf{c}^T\mathbf{x} \\ \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ \mathbf{x} \geq 0 \end{aligned}$$

Dual LP:

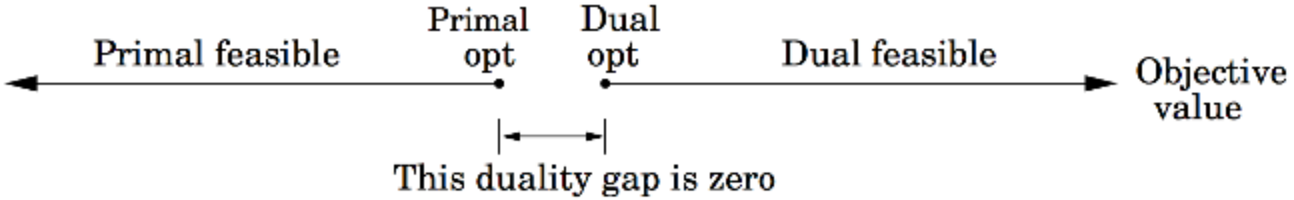
$$\begin{aligned} \min\ \mathbf{y}^T\mathbf{b} \\ \mathbf{y}^T\mathbf{A} \geq \mathbf{c}^T \\ \mathbf{y} \geq 0 \end{aligned}$$

Primal LP:

$$\begin{aligned} \max\ c_1x_1 + \cdots + c_nx_n & \\ a_{i1}x_1 + \cdots + a_{in}x_n \leq b_i & \quad \text{for } i \in I \\ a_{i1}x_1 + \cdots + a_{in}x_n = b_i & \quad \text{for } i \in E \\ x_j \geq 0 & \quad \text{for } j \in N \end{aligned}$$

Dual LP:

$$\begin{aligned} \min\ b_1y_1 + \cdots + b_my_m & \\ a_{1j}y_1 + \cdots + a_{mj}y_m \geq c_j & \quad \text{for } j \in N \\ a_{1j}y_1 + \cdots + a_{mj}y_m = c_j & \quad \text{for } j \notin N \\ y_i \geq 0 & \quad \text{for } i \in I \end{aligned}$$



Updated 4/7/2023

Topic 3: Efficient query evaluation
Unit 2: Cyclic query evaluation
Lecture 24

Wolfgang Gatterbauer

CS7240 Principles of scalable data management (sp23)

https://northeastern-datalab.github.io/cs7240/sp23/

4/7/2023

Pre-class conversations

- Last class summary
- Project: comments on comments (think rapid prototyping https://en.wikipedia.org/wiki/Rapid_application_development)
- Please prepare written comments for the class feedback phase

Today:
  - End of efficient query evaluation for cycles
  - Pointers to recorded tutorial on optimization problems & top-k
- Next time:
  - last class by me, on graphs
  - then you present

Suggestion for final report: iterate on my comments

Wolfgang: why do you need to join the two tables, and not just filter the second one on the IDs directly? There is no information in the user table that you need (as you would in say: posts from usernames < "Alice") [redacted] This is a very simple example and incidentally the foreign key is the one used for filtering the first table as well. We want to generalize the problem a bit such that the list of users may have been derived based on a column other than the User ID, and we need the user IDs to filter the posts. In such a case, we would need a join to execute a single query.

Although all the queries in Figure 2 and their equivalent SQL queries shown in Figure 3 vary significantly in implementation, they all produce the same result in “allPosts”. Thus, it is important to evaluate the performance of these queries in terms of execution time, and heap utilization in order to identify the best implementation for a particular use case. To evaluate the queries, we construct a test database, execute the queries against this database, and measure their performance.

Wolfgang Gatterbauer

Thus to make this example self-contained (and not redundant) my suggestion to select on user names, like only those whose names starts with letters A, B, or C

relational join, respectively, are more interesting cases. The have almost identical evaluation times, with the where clause being marginally faster. Wolfgang: did you run this experiment many times and took the median and showed 90% confidence interval? Without that it is all within margin of error [redacted]: We ran the experiment 100 times for each query and considered the average of those 100 runs. We can switch to using the median and perform some statistical significance test instead. These queries scale the best overall, being on par with the best case of the N+1 query on the lower end, and performing marginally better than the naive transformation on the higher end. We also observe that the N+1 query performs better than than naive transformation for N less than 20 on this particular database.

Figure 5 shows the comparison of heap utilization for different reformulations of query against different number of records requested from the database. The X-axis shows the number of records requested from 1001(1% of the database) to 100100 (100% of the database), and the Y-axis shows

Wolfgang Gatterbauer

I strongly suggest to use the median and then the 90% confidence interval. Thus for each x-point, you sort the times in increasing order, report the AVG of 5th and 6th as the lower bound, the AVG of 50th and 51st as the median, and AVG of 95 and 96 as the upper bound (of 90% confidence)

Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

Topic Duality in Linear Programming (LP)

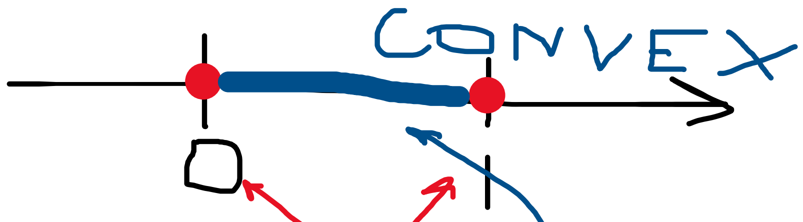
- Subtopics
  - Connections between (max) set packing and (min) set covers in graphs
  - Linear Programming (LP) and duality gaps
  - LP relaxations of ILP problems (Integer Linear Programming)
  - Duality b/w independent vertex sets and edge covers

Duality in linear programming: Intuitively, every Linear Program has a dual problem with the same optimal solution, but the variables in the dual problem correspond to constraints in the primal problem and vice versa.
But the notion of duality is more general:

- "Over and over again, it turns out that one can associate with a given mathematical object a related, 'dual' object that helps one to understand the properties of the object one started with." [The Princeton Companion to Mathematics, 2008]
- "Fundamentally, duality gives two different points of view of looking at the same object.[Michael Atiyah 2007]

LP relaxations of ILP problems (Integer Linear Programming)

Example: Minimal (Fractional) Vertex Cover in k-clique



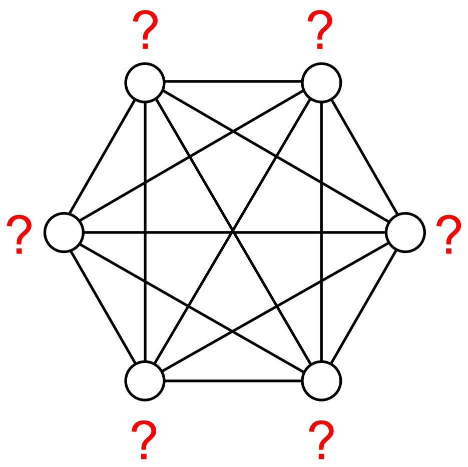
Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

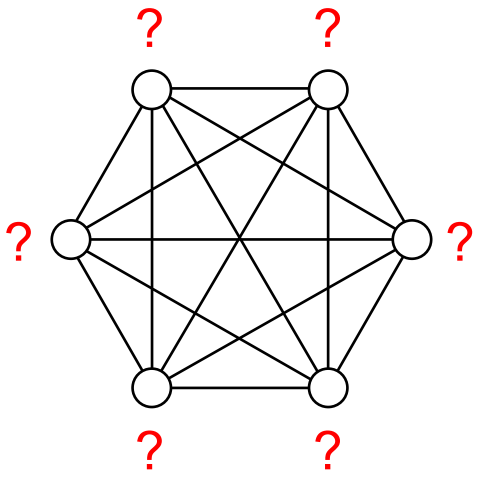
a fractional & convex relaxation

Minimal Integral Vertex Cover:



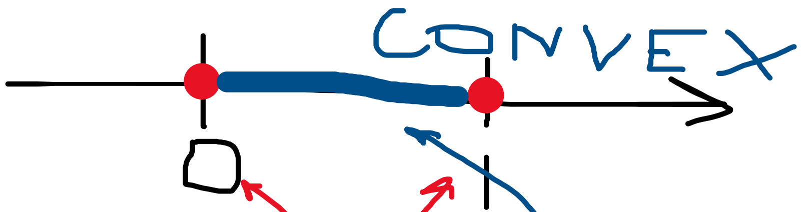
ILP: ?

Minimal Fractional Vertex Cover:



LP: ?

Example: Minimal (Fractional) Vertex Cover in k-clique

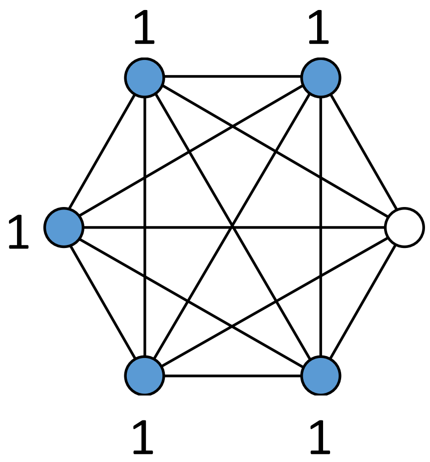


Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

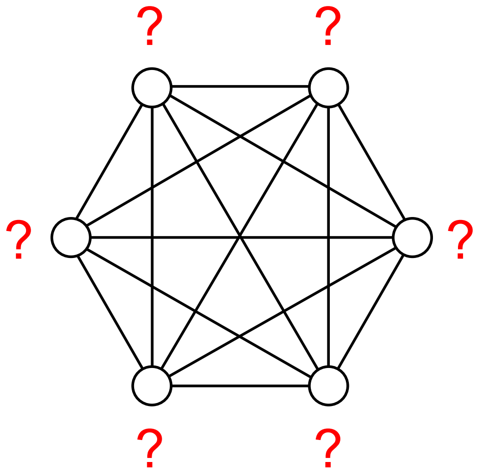
Minimal Integral Vertex Cover:



ILP: 5 = k-1

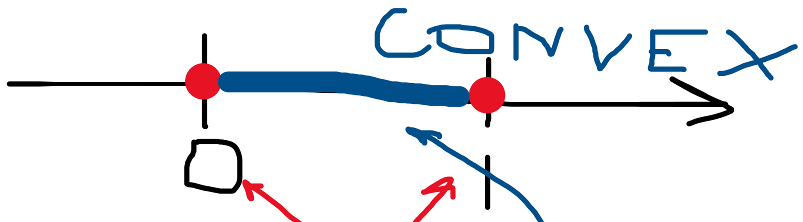
for k-clique

Minimal Fractional Vertex Cover:



LP: ?

Example: Minimal (Fractional) Vertex Cover in k-clique

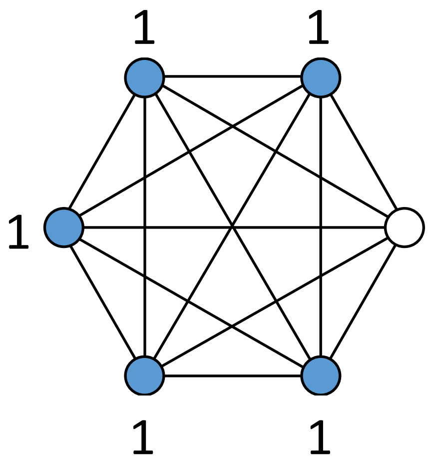


Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

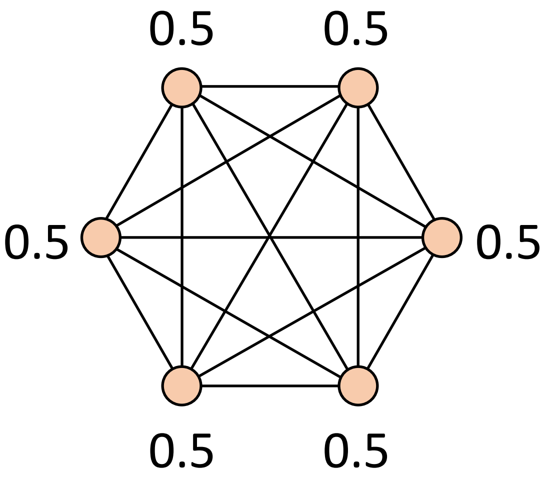
Minimal Integral Vertex Cover:



ILP: 5 = k-1

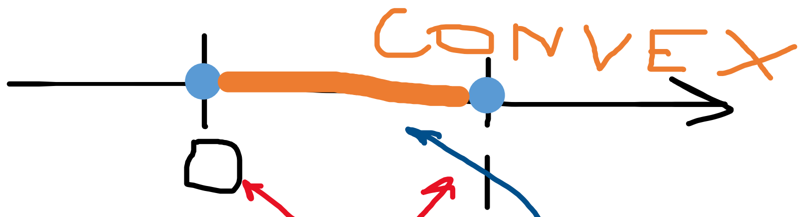
for k-clique

Minimal Fractional Vertex Cover:



LP: 3 = k/2

Example: Minimal (Fractional) Vertex Cover in even k-cycle

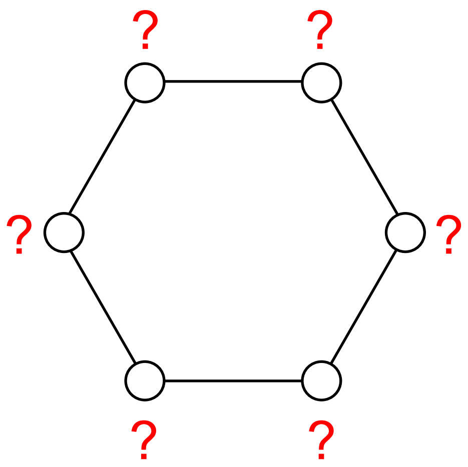


Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

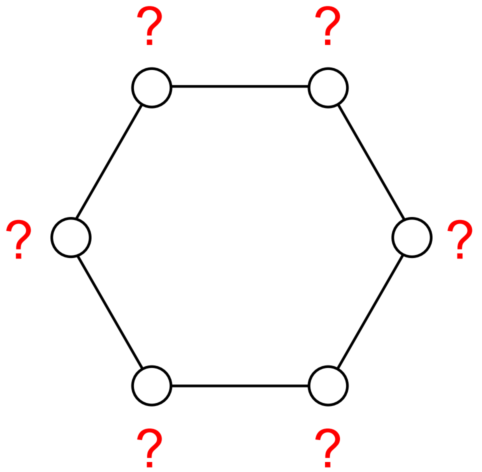
or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

Minimal Integral Vertex Cover:



ILP: ?

Minimal Fractional Vertex Cover:



LP: ?

Example: Minimal (Fractional) Vertex Cover in even k-cycle

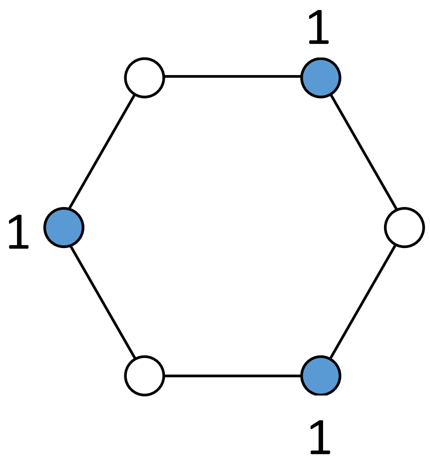
CONVEX

Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

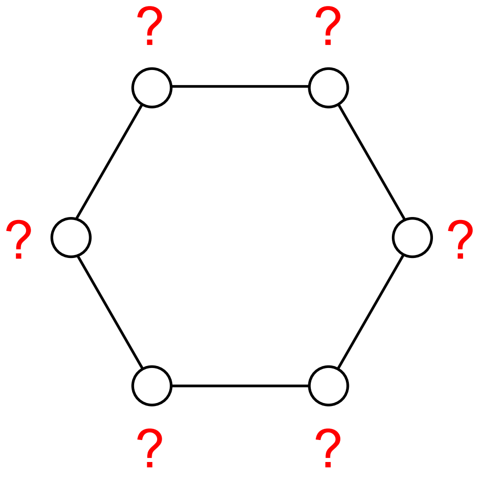
Minimal Integral Vertex Cover:



ILP: 3 = k/2

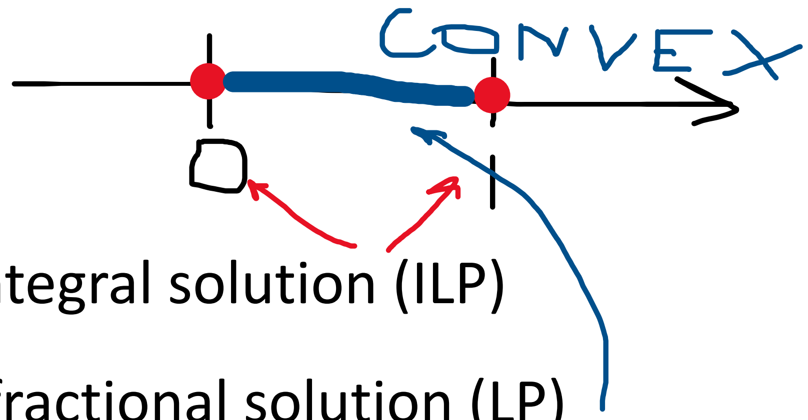
for even cycle of length k

Minimal Fractional Vertex Cover:



LP: ?

Example: Minimal (Fractional) Vertex Cover in even k-cycle

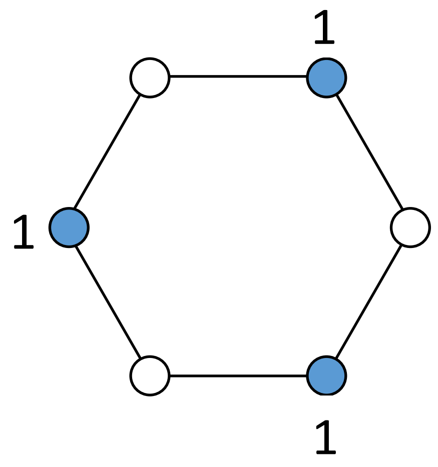


Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

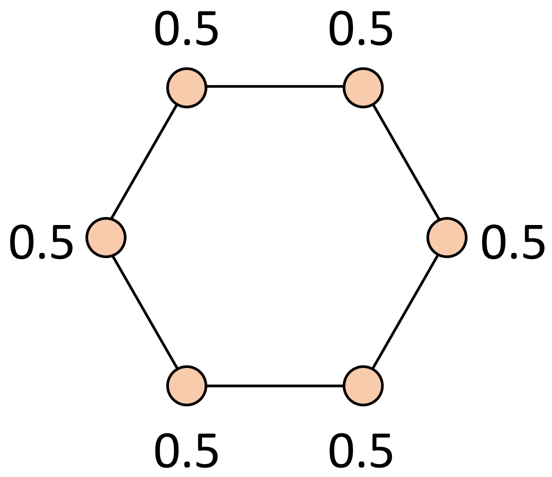
Minimal Integral Vertex Cover:



ILP: 3 = k/2

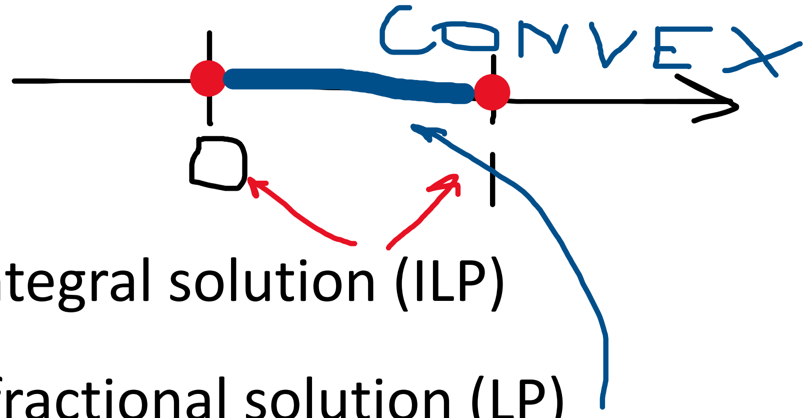
for even cycle
of length k

Minimal Fractional Vertex Cover:



LP: 3 = k/2

Example: Minimal (Fractional) Vertex Cover in odd k-cycle

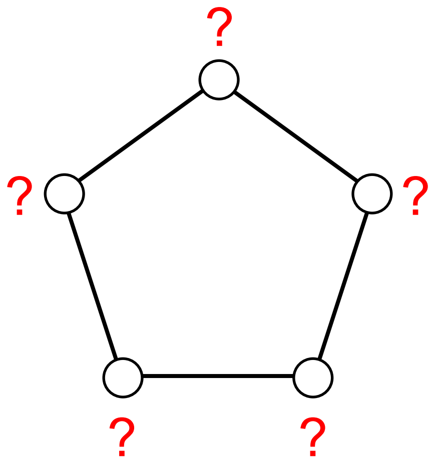


Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

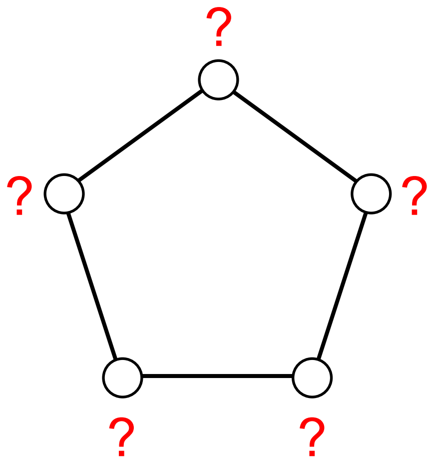
or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

Minimal Integral Vertex Cover:



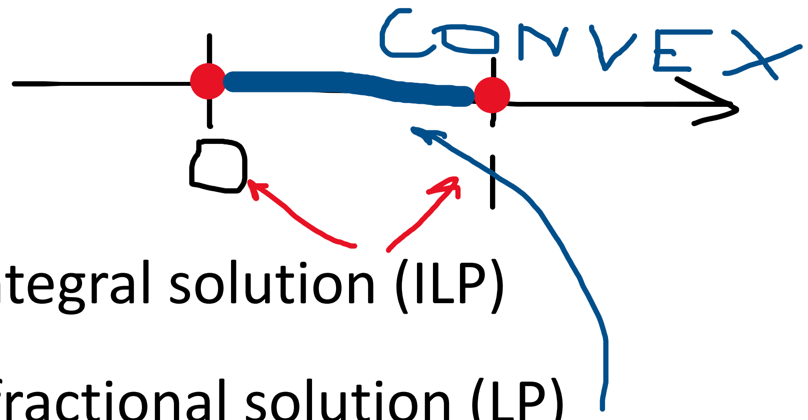
ILP: ?

Minimal Fractional Vertex Cover:



LP: ?

Example: Minimal (Fractional) Vertex Cover in odd k-cycle

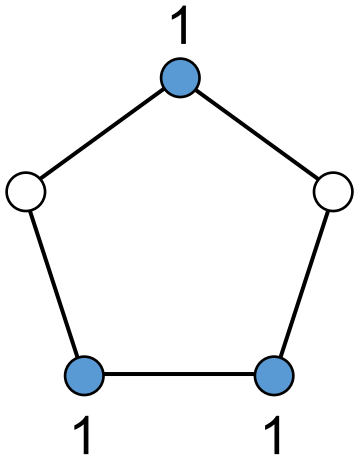


Objective: $\min \sum_{v \in V} w_v$ s.t. $w_v + w_u \geq 1$ for each edge

and $w_v \in \{0,1\}$ for each node for integral solution (ILP)

or $0 \leq w_v \leq 1$ for each node for fractional solution (LP)

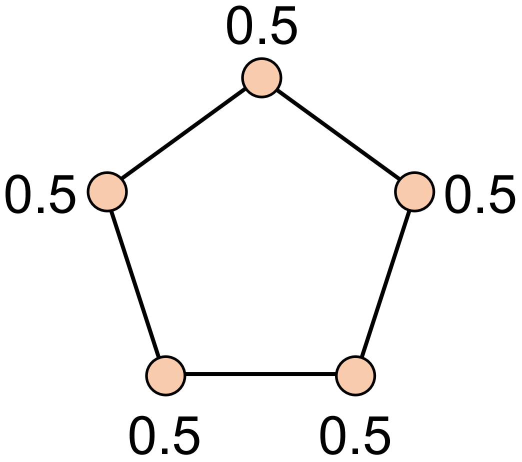
Minimal Integral Vertex Cover:



ILP: 3 = (k+1)/2

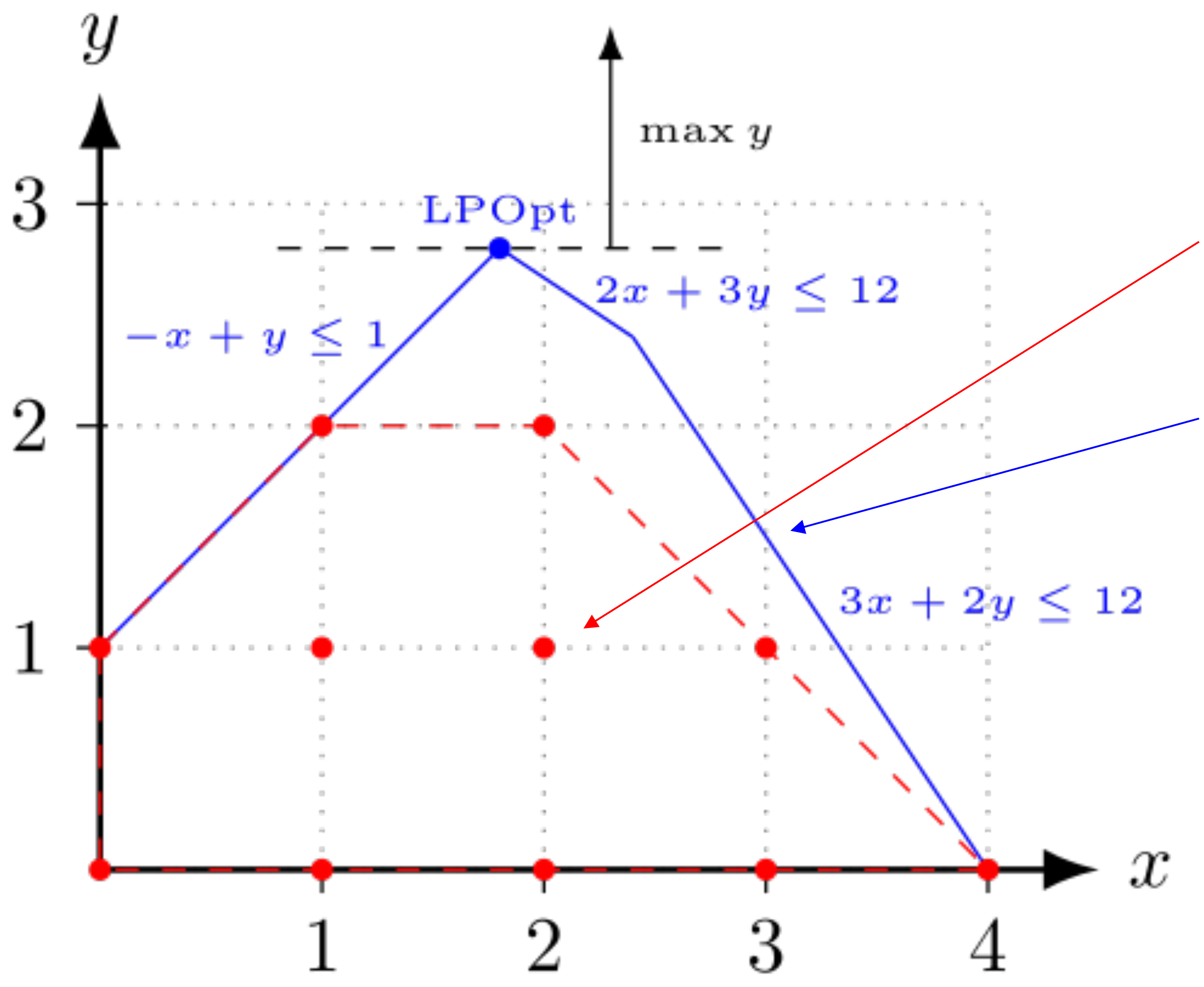
for odd cycle
of length k

Minimal Fractional Vertex Cover:



LP: 2.5 = k/2

ILP and its LP relaxation



ILP (Integer program or Integer Linear program)

LP-relaxation obtained from an ILP by relaxing the integrality constraints for variables x and y

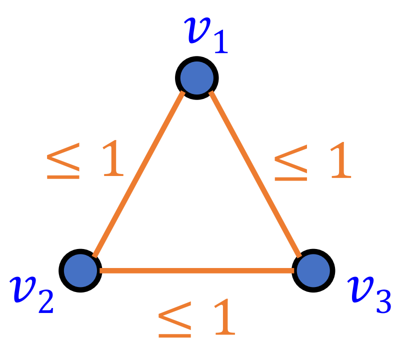
Notice the search space gets enlarged and becomes convex. Contrast with GHD vs HD: there the search space got restricted...

Source: https://en.wikipedia.org/wiki/Linear_programming_relaxation

Duality b/w independent vertex sets and edge covers

A quick primer on Duality in Linear Programming

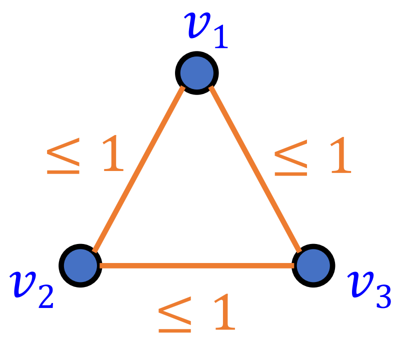
Primal: Max Independence (Vertex) set



$$\begin{aligned} \max\ & v_1 + v_2 + v_3 \text{ , s.t.} \\ & v_1 + v_2 \leq 1 \\ & v_1 + v_3 \leq 1 \\ & v_2 + v_3 \leq 1 \end{aligned}$$

A quick primer on Duality in Linear Programming

Primal: Max Independence (Vertex) set



non-negative multiplier per edge

$$
\begin{aligned}
\max\ & v_1 + v_2 + v_3 \text{ , s.t.} \\
& v_1 + v_2 \leq 1 \qquad u_1 \\
& v_1 + v_3 \leq 1 \qquad u_2 \\
& v_2 + v_3 \leq 1 \qquad u_3
\end{aligned}
$$

$$(u_1 + u_2)v_1 + (u_1 + u_3)v_2 + (u_2 + u_3)v_3 \leq u_1 + u_2 + u_3$$

if ≥ 1 $\qquad \geq 1$ $\qquad \geq 1$

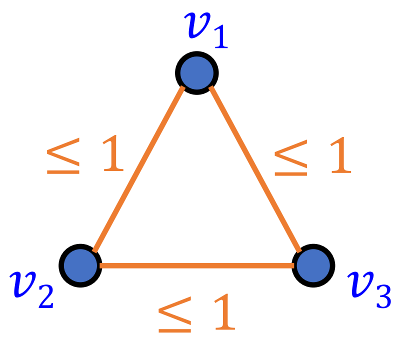
for each vertex

then the right side $\sum_j u_j$ is an upper bound for the primal objective $\sum_i v_i$

A quick primer on Duality in Linear Programming

Primal: Max Independence (Vertex) set

What is this dual problem ?



$$
\begin{aligned}
\max\ & v_1 + v_2 + v_3 \text{, s.t.} \\
& v_1 + v_2 \leq 1 \\
& v_1 + v_3 \leq 1 \\
& v_2 + v_3 \leq 1
\end{aligned}
\qquad\qquad
\begin{aligned}
\min\ & u_1 + u_2 + u_3 \text{, s.t.} \\
& u_1 + u_2 \geq 1 \\
& u_1 + u_3 \geq 1 \\
& u_2 + u_3 \geq 1
\end{aligned}
$$

$$(u_1 + u_2)v_1 + (u_1 + u_3)v_2 + (u_2 + u_3)v_3 \leq u_1 + u_2 + u_3$$

if ≥ 1 $\qquad$ ≥ 1 $\qquad$ ≥ 1

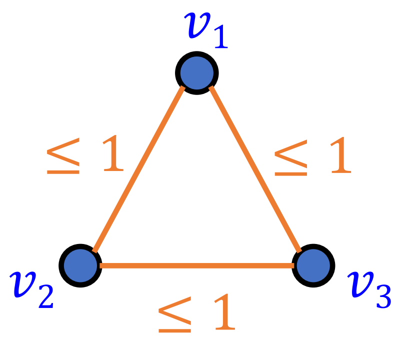
for each vertex

then the right side $\sum_j u_j$ is an upper bound for the primal objective $\sum_i v_i$

A quick primer on Duality in Linear Programming

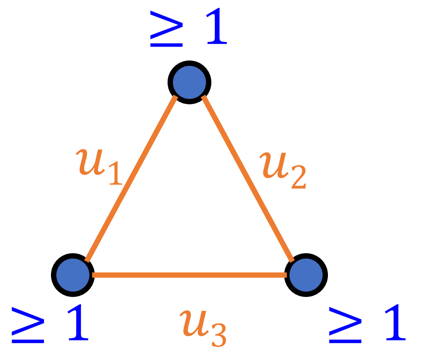
Primal: Max Independence (Vertex) set

Dual: Min Edge cover



$$\begin{aligned}
\max\ v_1 + v_2 + v_3\ &, \text{s.t.} \\
v_1 + v_2 \quad\quad &\leq 1 \\
v_1 \quad\quad + v_3 &\leq 1 \\
v_2 + v_3 &\leq 1
\end{aligned}$$

$$\begin{aligned}
\min\ u_1 + u_2 + u_3\ &, \text{s.t.} \\
u_1 + u_2 \quad\quad &\geq 1 \\
u_1 \quad\quad + u_3 &\geq 1 \\
u_2 + u_3 &\geq 1
\end{aligned}$$



$$(u_1 + u_2)v_1 + (u_1 + u_3)v_2 + (u_2 + u_3)v_3 \leq u_1 + u_2 + u_3$$

if ≥ 1 $\quad \geq 1 \quad \geq 1$

for each vertex

then the right side $\sum_j u_j$ is an upper bound for the primal objective $\sum_i v_i$

Example taken from: Dasgupta, Papadimitriou, Vazirani. Algorithms. 2006. http://algorithmics.lsi.upc.edu/docs/Dasgupta-Papadimitriou-Vazirani.pdf

Independent Sets & Edge covers in the Triangle

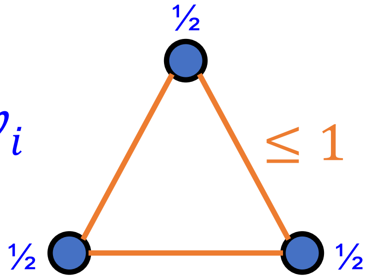
Fractional independence number (α^*)

max sum of weights $v_1, v_2, \ldots\ v_k\ \geq 0$ on vertices (variables)

s.t. for all $E(i,j)$: $v_i + v_j \leq 1$

$$\max v_1 + v_2 + v_3 \text{ , s.t.}$$
$$v_1 + v_2 \leq 1$$
$$v_1 + v_3 \leq 1$$
$$v_2 + v_3 \leq 1$$

$$\alpha^* = \max \textstyle\sum_i v_i$$



$$\boxed{\alpha^* = \rho^*}$$

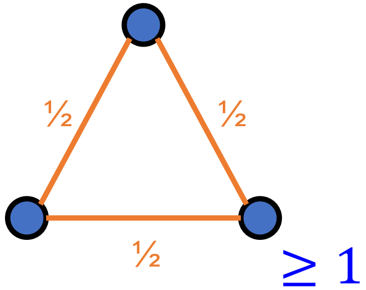
Fractional edge cover number (ρ^*)

min sum of weights $u_1, u_2, \ldots\ u_\ell \geq 0$ on edges (relations)

s.t. for all x_i: $\sum_{j:x_i \in E_j} u_j \geq 1$

$$\min u_1 + u_2 + u_3 \text{ , s.t.}$$
$$u_1 + u_2 \geq 1$$
$$u_1 + u_3 \geq 1$$
$$u_2 + u_3 \geq 1$$

$$\rho^* = \min \textstyle\sum_j u_j$$



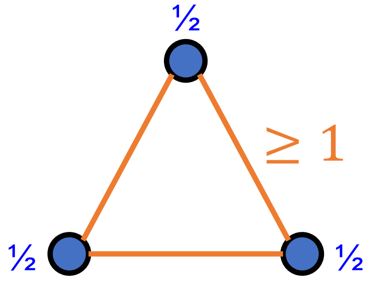
Independent Sets & Edge covers in the Triangle

Fractional vertex cover number (τ^*)

min sum of weights $v_1, v_2, \ldots\ v_k\ \geq 0$ on vertices (variables)

s.t. for all *E(i,j)*: $v_i + v_j \geq 1$

$\tau^* = \min \sum_i v_i$



min $v_1 + v_2 + v_3$, s.t.

$v_1 + v_2 \geq 1$

$v_1 + v_3 \geq 1$

$v_2 + v_3 \geq 1$

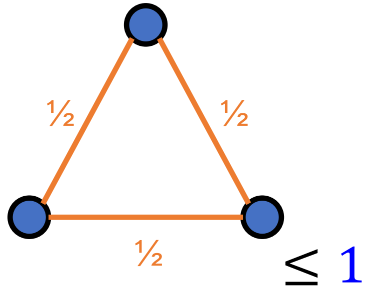
$\tau^* = \nu^*$

Fractional matching (edge packing) number (ν^*)

max sum of weights $u_1, u_2, \ldots\ u_\ell \geq 0$ on edges (relations)

s.t. for all x_i: $\sum_{j: x_i \in R_j} u_j \leq 1$

$\nu^* = \max \sum_j u_j$



max $u_1 + u_2 + u_3$, s.t.

$u_1 + u_2 \leq 1$

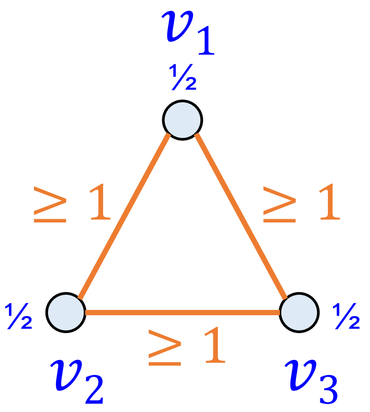
$u_1 + u_3 \leq 1$

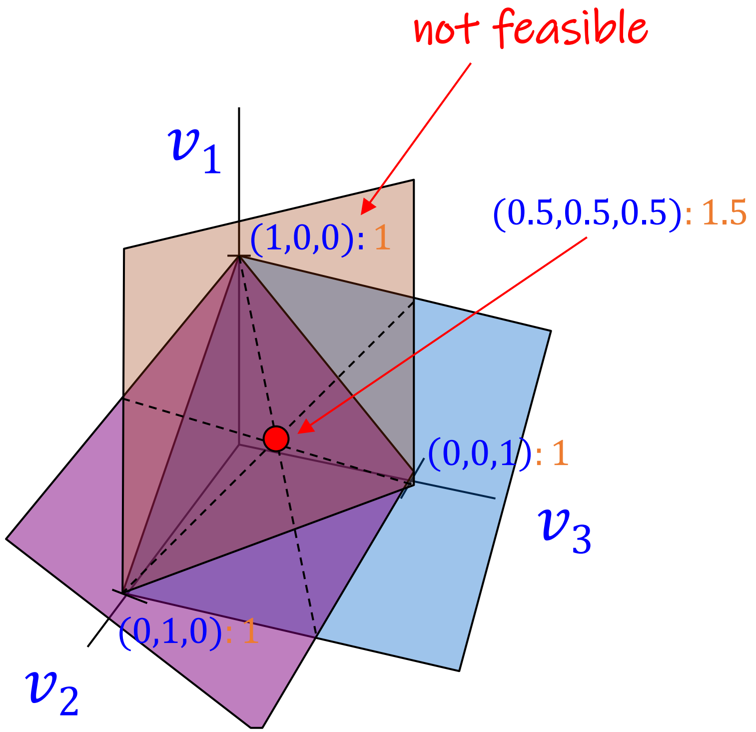
$u_2 + u_3 \leq 1$

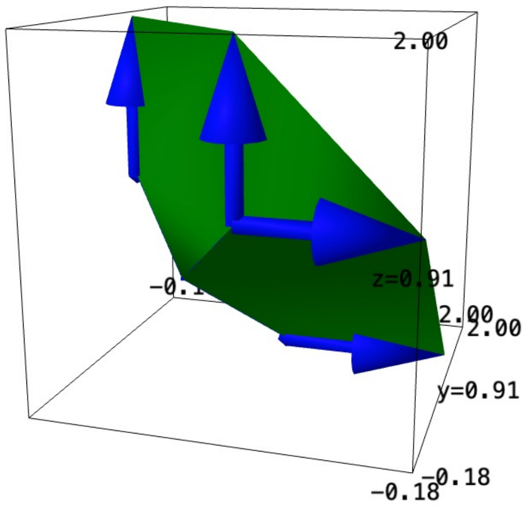
Fractional vertex cover in the triangle

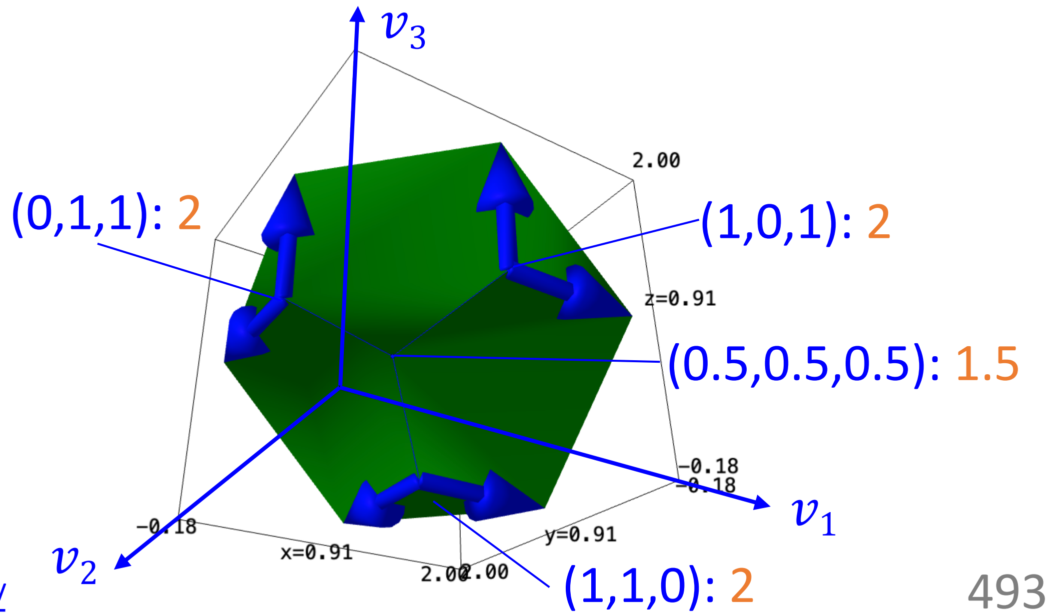
```
# https://sagecell.sagemath.org/
# inequalities: -1+v1+v2>=0, -1+v2+v3>=0, -1+v1+v3>=0, 1-v1>=0, 1-v2>=0, 1-v3>=0
p = Polyhedron(ieqs = [[-1,1,1,0],[-1,0,1,1],[-1,1,0,1],[0,1,0,0],[0,0,1,0],[0,0,0,1]])
p.plot()
```

$$\begin{aligned} \min\ & v_1 + v_2 + v_3\text{, s.t.} \\ & v_1 + v_2 \geq 1 \\ & v_1 + v_3 \geq 1 \\ & v_2 + v_3 \geq 1 \end{aligned}$$





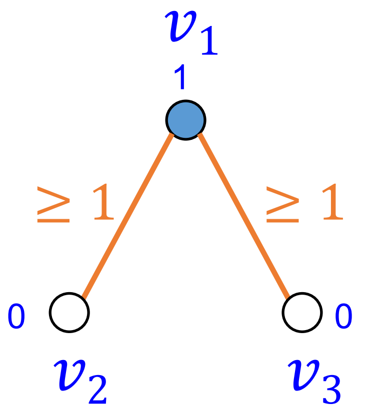


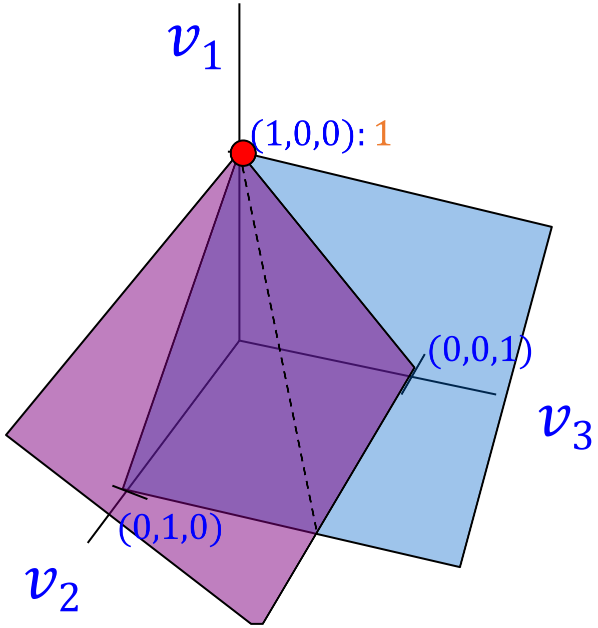


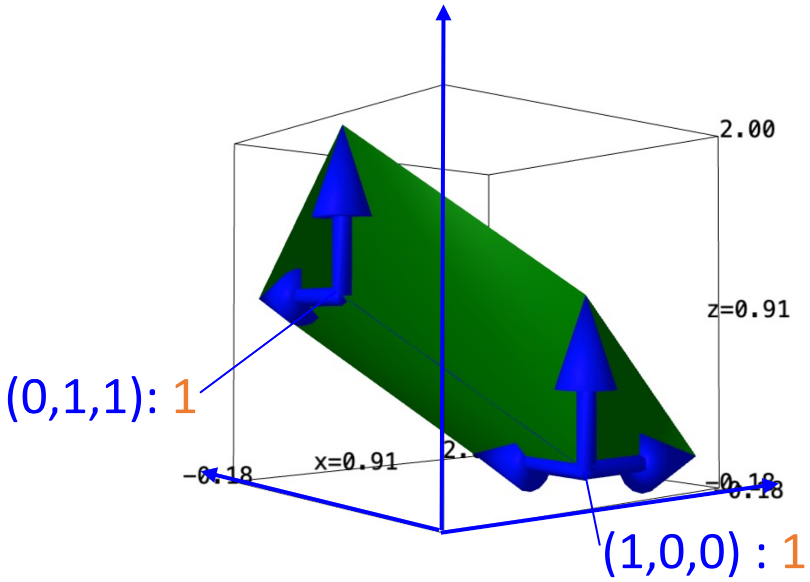
Fractional vertex cover in bipartite graph

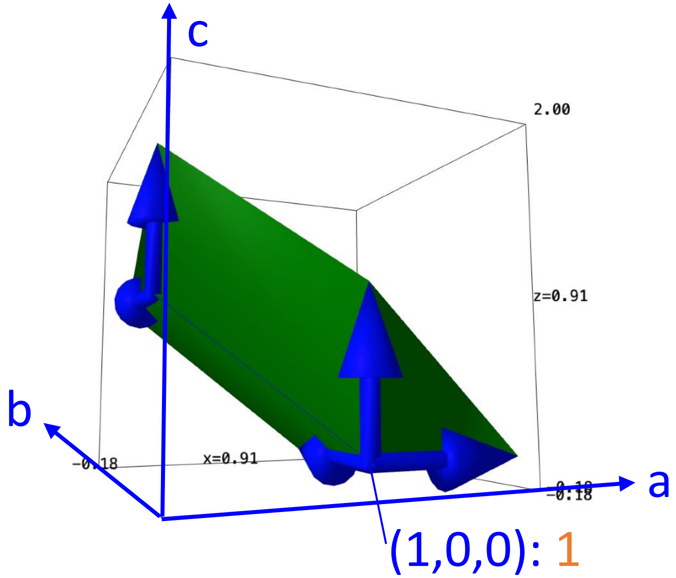
```
# https://sagecell.sagemath.org/
# inequalities: -1+v1+v2>=0, -1+v2+v3>=0, 1-v1>=0, 1-v2>=0, 1-v3>=0
p = Polyhedron(ieqs = [[-1,1,1,0],[-1,0,1,1],[0,1,0,0],[0,0,1,0],[0,0,0,1]])
p.plot()
```

$$\min v_1 + v_2 + v_3 \text{ , s.t.}$$
$$v_1 + v_2 \geq 1$$
$$v_1 + v_3 \geq 1$$









Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

What do we know about bounding the size of the answer?

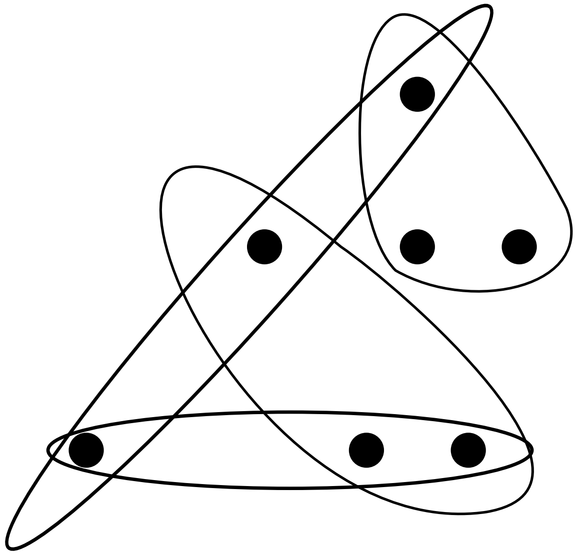
(. . .and enumerating all solutions)

Example by: Marx. "Graphs, hypergraphs, and the complexity of conjunctive database queries", ICDT 2017. http://edbticdt2017.unive.it/marx-icdt2017-talk.pdf

Upper bound

minimal edge cover

Observation: If the hypergraph has $\boxed{\text{edge cover number } \rho}$ and every relation has size at most N, then there are at most N^ρ tuples in the answer.

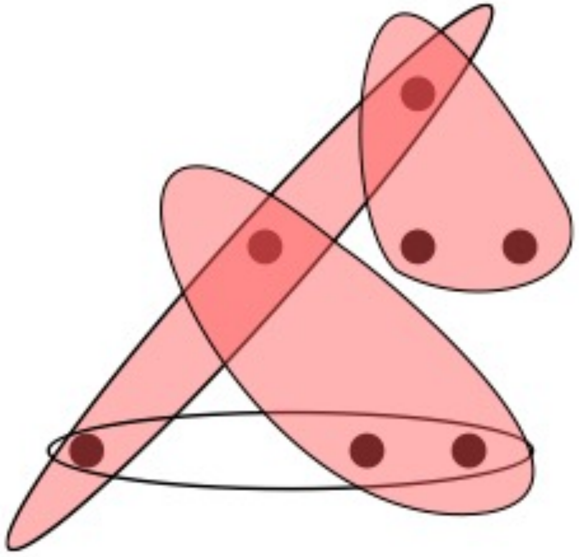


Upper bound

minimal edge cover

Observation: If the hypergraph has $\boxed{\text{edge cover number } \rho}$ and every relation has size at most N, then there are at most N^ρ tuples in the answer.

N^3

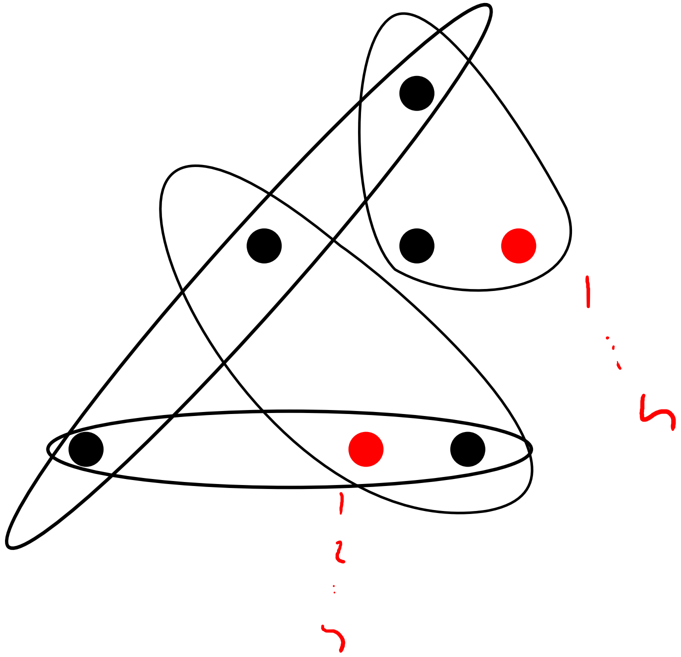


Example by: Marx. "Graphs, hypergraphs, and the complexity of conjunctive database queries", ICDT 2017. http://edbticdt2017.unive.it/marx-icdt2017-talk.pdf

Lower bound

maximal independent set

Observation: If the hypergraph has $\boxed{\text{independence number } \alpha}$, then one can construct an instance where every relation has size N and the answer has size N^{α}.

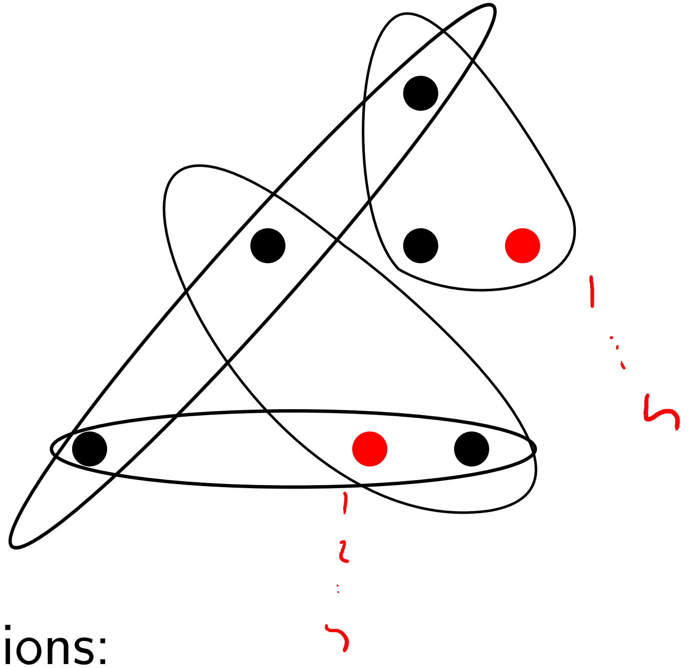


Example by: Marx. "Graphs, hypergraphs, and the complexity of conjunctive database queries", ICDT 2017. http://edbticdt2017.unive.it/marx-icdt2017-talk.pdf

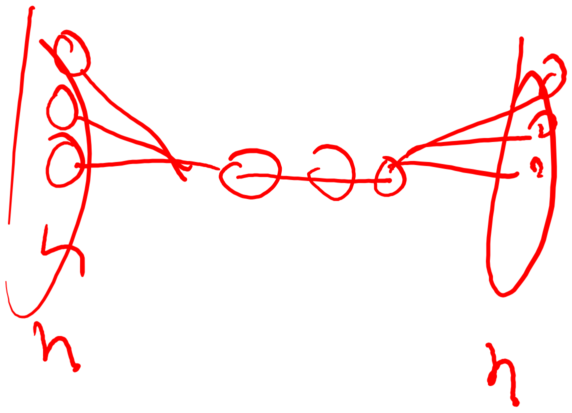
Lower bound

maximal independent set

Observation: If the hypergraph has $\boxed{\text{independence number } \alpha}$, then one can construct an instance where every relation has size N and the answer has size N^α.

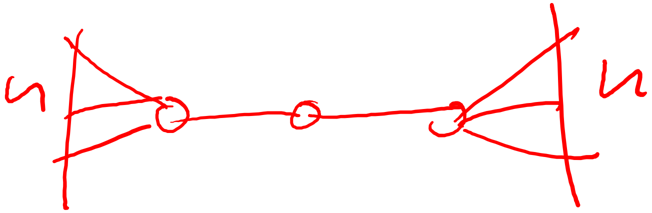


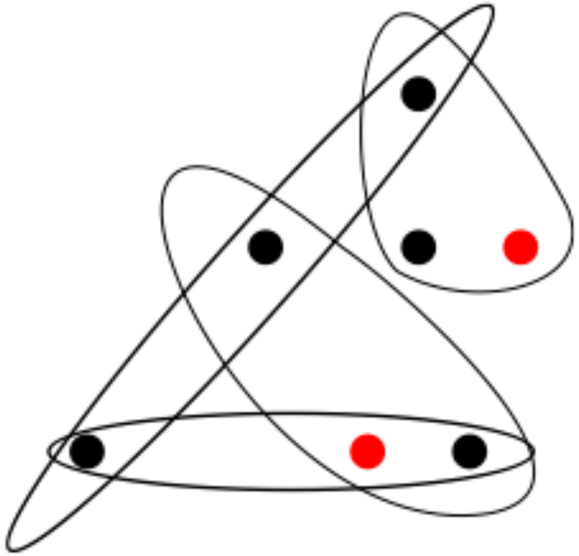
N^2



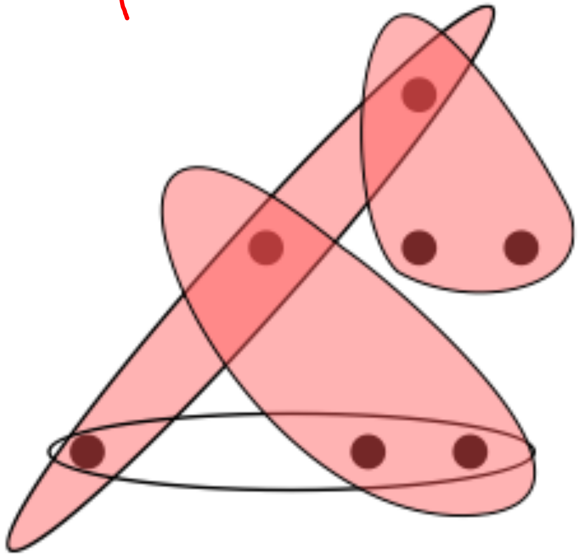
Definition of the relations:

- If variable A is in the independent set, then it can take any value in $[N]$.
- Otherwise it is forced to 1.





N^2



N^3

Which is tight: the upper bound or the lower bound?

Fractional values

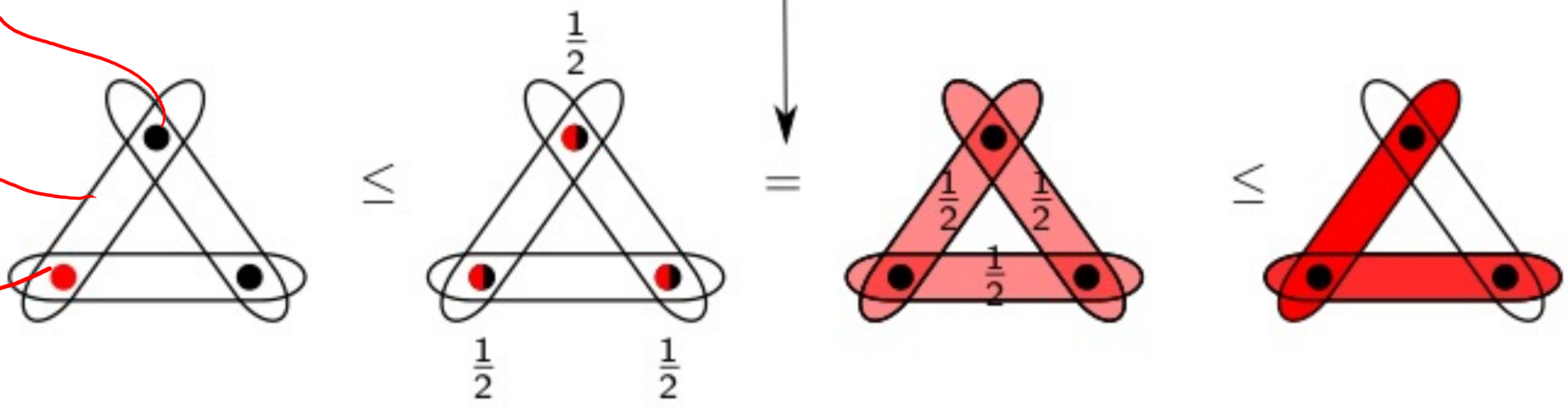
- α: independence number
- α^*: fractional independence number
  (max. weight of vertices s.t. each edge contains weight ≤ 1)
- ρ^*: fractional edge cover number
  (min. weight of edges s.t. each vertex receives weight ≥ 1)
- ρ: edge cover number

LP duality!

$\in \{0,1\}$

Σ

$\in \{0,1\}$

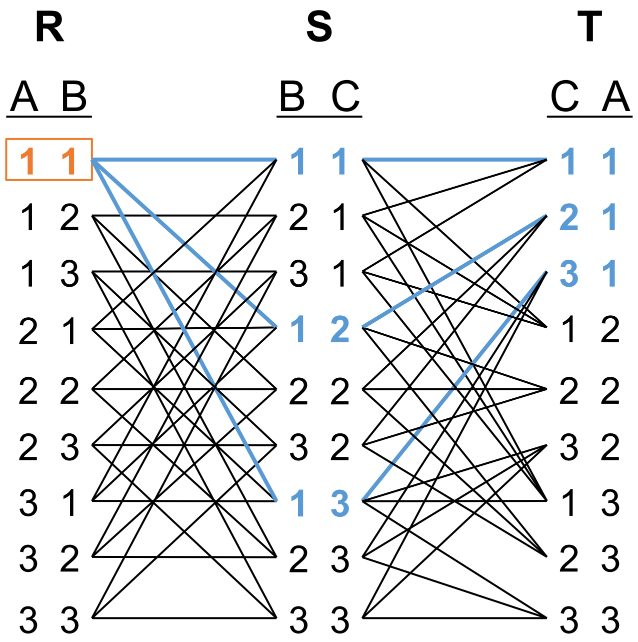


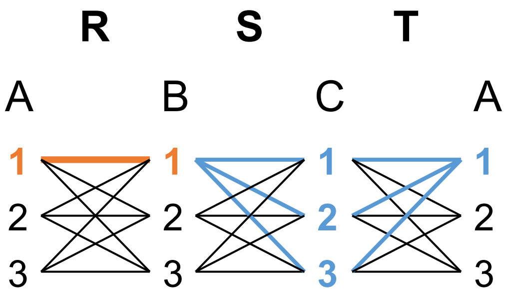
$\alpha = 1$ $\alpha^* = 3/2$ $\rho^* = 3/2$ $\rho = 2$

A tight example for AGM bound $O(n^{1.5})$ for triangle $\mathbf{Q}_\Delta$

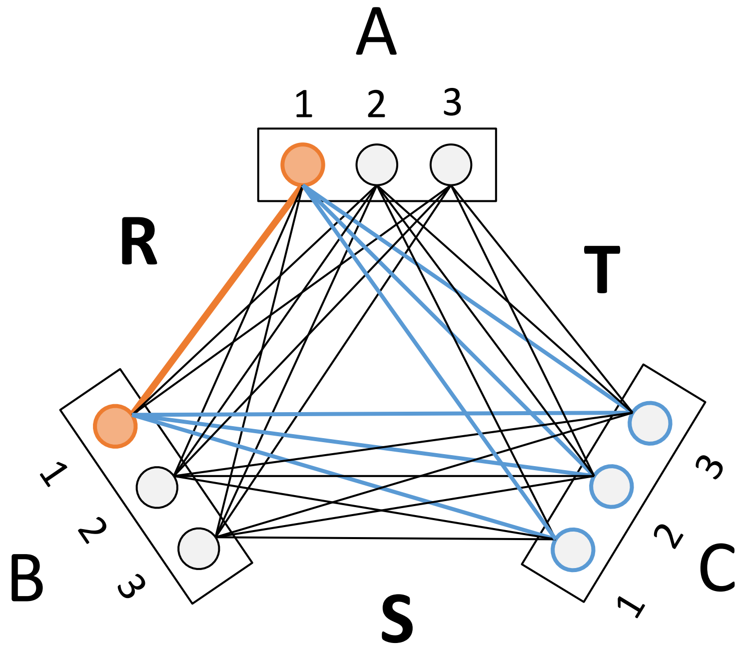
$Q_\Delta(A,B,C) = R(A,B) \bowtie S(B,C) \bowtie T(C,A)$

Q(x,y,z) :- R(x,y), S(y,z), T(z,x).





Notice every tuple is part of 3 join results, e.g. shown here for R(1,1)



$n \;=\;$ 9 number of tuples per relation (= DB size for self-joins)

$m \;=\; \sqrt{n} = 3$ domain size

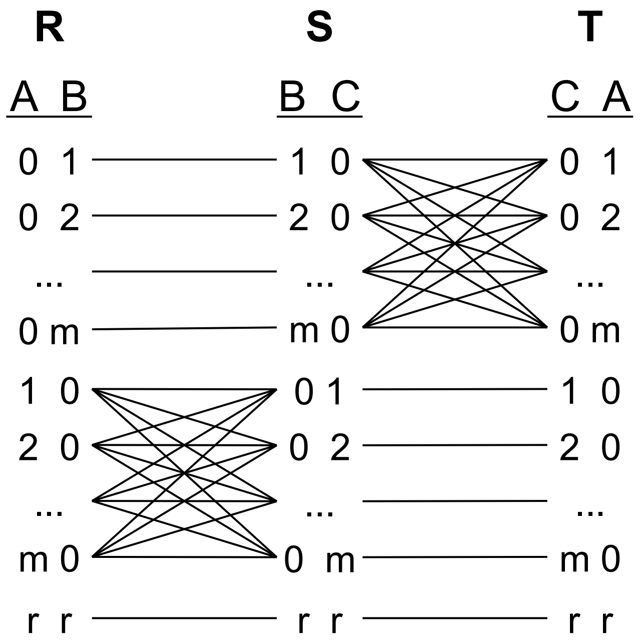
$|OUT| \;= n^{1.5} = 27$ output tuples

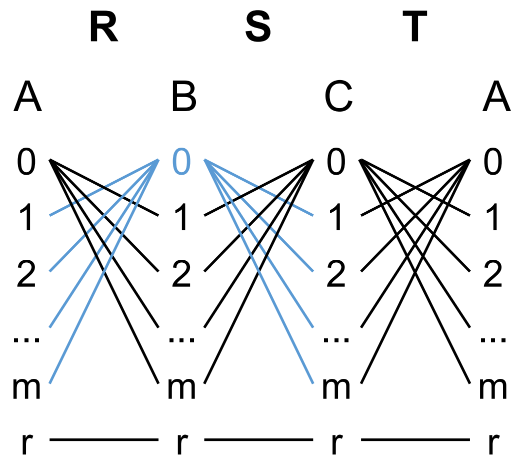
Q(x,y,z) :- R(x,y), R(y,z), R(z,x).

When binary joins give $O(n^2)$ intermediate sizes for $\mathbf{Q}_\Delta$

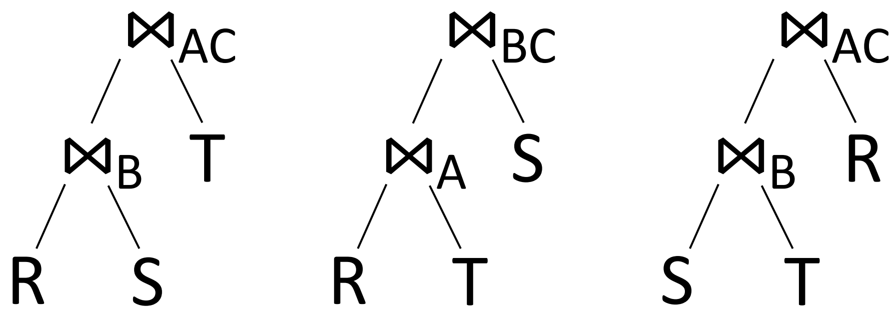
$Q_\Delta(A,B,C) = R(A,B) \bowtie S(B,C) \bowtie T(C,A)$

Q(x,y,z) :- R(x,y), S(y,z), T(z,x).





In whatever sequence we join the three tables, the size of the first join will always be $\Theta(n^2)$



$|R \bowtie_B S| = |R \bowtie_A T| = |S \bowtie_B T| = m^2 = \Theta(n^2)$

$n = 2m + 1$ tuples per relation

$m + 2$ domain size

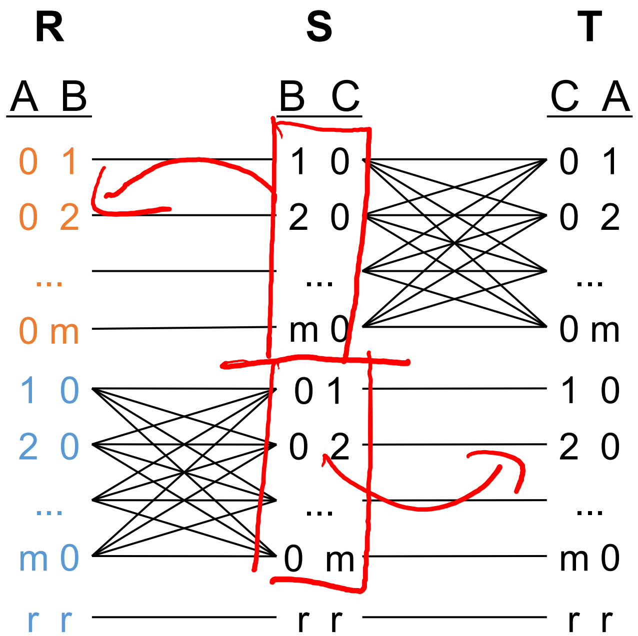
$|OUT| = 1$ output tuple

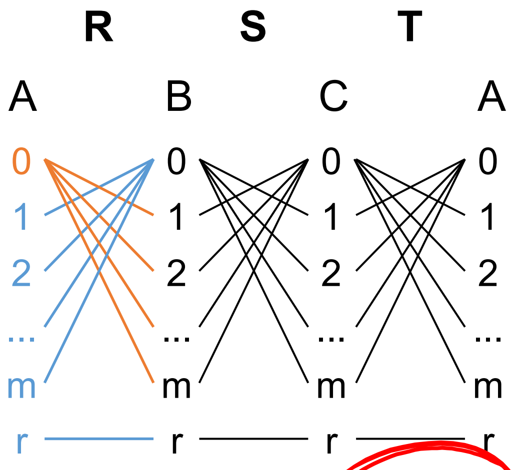
Solution: partition the data

Q_Δ (A,B,C) = R(A,B) ⋈ S(B,C) ⋈ T(C,A)

Q(x,y,z) :- R(x,y), S(y,z), T(z,x).

Trick: partition by outdegree, and use two plans in parallel!



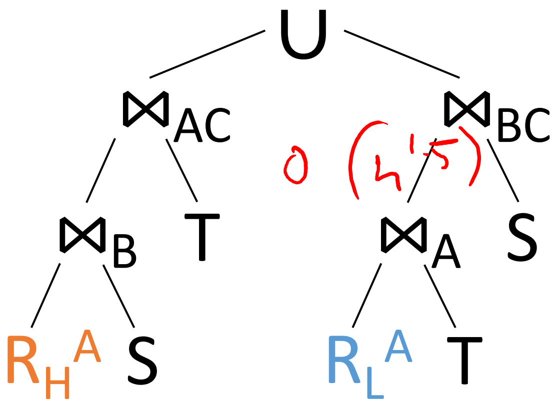


$n^{0.5} \cdot n^1 = n^{1.5}$

$R = R_H^A \cup R_L^A$

$R_H^A = \{(a,b) \in R : |\sigma_{A=a}R| > n^{0.5}\}$

$R_L^A = \{(a,b) \in R : |\sigma_{A=a}R| \leq n^{0.5}\}$



$n = 2m + 1$ tuples per relation

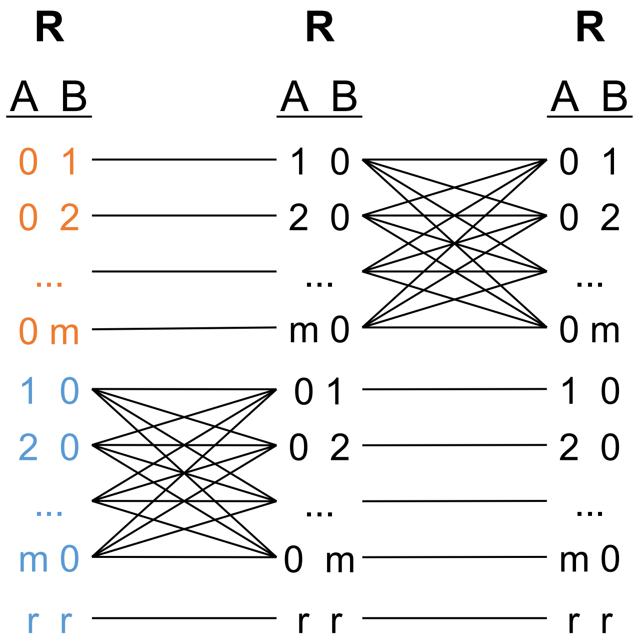
$m + 2$ domain size

$|OUT| = 1$ output tuple

Solution: partition the data

602

Q(x,y,z) :- R(x,y), R(y,z), R(z,x).



m+2 = domain size

(2m+1) database size

1 = output size

```
-----------------------------
-- Query 1
-----------------------------
select count(*)
into record1
from R R1, R R2, R R3
where R1.B=R2.A
and R2.B=R3.A
and R3.B=R1.A;
```

```
-----------------------------
-- Query 2
-----------------------------
With Cutoff as
    (select sqrt(count(*)) as C from R),
DomainH as
    (select R.A from R
    group by R.A
    having count(*) > (Select C from Cutoff)),
RH as
    (SELECT R.A, R.B
    FROM R, DomainH
    WHERE R.A=DomainH.A),
RL as
    (select * from R
    Except
    select * from RH)
select count(*) into record2
from
    (select T.A, T.B, T.C
    from    (select RH.A, RH.B, R2.B as C
        from RH, R R2
        where RH.B = R2.A) T, R R3
    where T.A = R3.B and T.C = R3.A
    union
    select T.A, T.B, T.C
    from    (select RL.A, RL.B, R3.A as C
        from RL, R R3
        where RL.A = R3.B) T, R R2
    where T.B = R2.A and T.C = R2.B) X;
```

m=2000: t_{Q1}=2409 msec t_{Q2}=7 msec

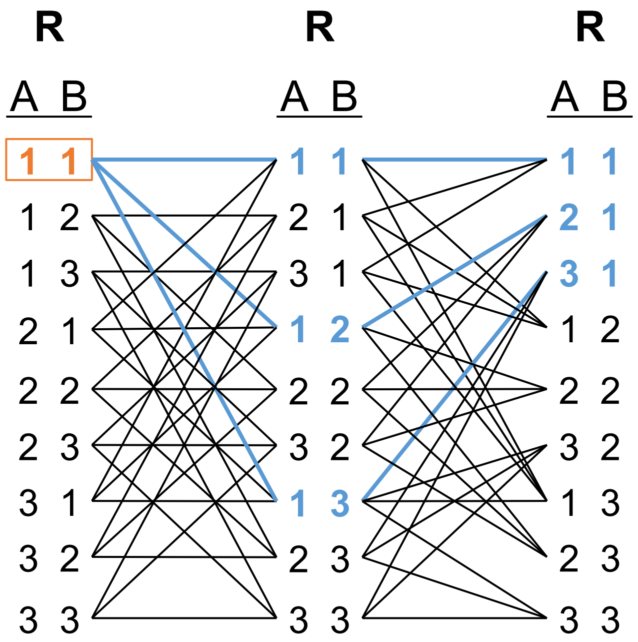
m=4000: t_{Q1}=8912 msec t_{Q2}=14 msec

SQL example available at: https://github.com/northeastern-datalab/cs3200-activities/tree/master/sql

Solution: partition the data

603

$Q(x,y,z)$:- $R(x,y), R(y,z), R(z,x)$.



m=domain size

m^2 database size

m^3 output size

```
-----------------------------
-- Query 1
-----------------------------
select count(*)
into record1
from R R1, R R2, R R3
where R1.B=R2.A
and R2.B=R3.A
and R3.B=R1.A;
```

```
-----------------------------
-- Query 2
-----------------------------
With Cutoff as
    (select sqrt(count(*)) as C from R),
DomainH as
    (select R.A from R
    group by R.A
    having count(*) > (Select C from Cutoff)),
RH as
    (SELECT R.A, R.B
    FROM R, DomainH
    WHERE R.A=DomainH.A),
RL as
    (select * from R
    Except
    select * from RH)
select count(*) into record2
from
    (select T.A, T.B, T.C
    from    (select RH.A, RH.B, R2.B as C
        from RH, R R2
        where RH.B = R2.A) T, R R3
    where T.A = R3.B and T.C = R3.A
    union
    select T.A, T.B, T.C
    from    (select RL.A, RL.B, R3.A as C
        from RL, R R3
        where RL.A = R3.B) T, R R2
    where T.B = R2.A and T.C = R2.B) X;
```

m=100: t_{Q1}=0.60 sec t_{Q2}=0.94 sec

m=200: t_{Q1}=5.8 sec t_{Q2}=10.3 sec

SQL example available at: https://github.com/northeastern-datalab/cs3200-activities/tree/master/sql

Algorithmica (1997) 17: 209–223

Algorithmica

Finding and Counting Given Length Cycles[1]

N. Alon,[2] R. Yuster,[2] and U. Zwick[2]

Abstract. We present an assortment of methods for finding and counting simple cycles of a given length in directed and undirected graphs. Most of the bounds obtained depend solely on the number of edges in the graph in question, and not on the number of vertices. The bounds obtained improve upon various previously known results.

k=2 for triangle

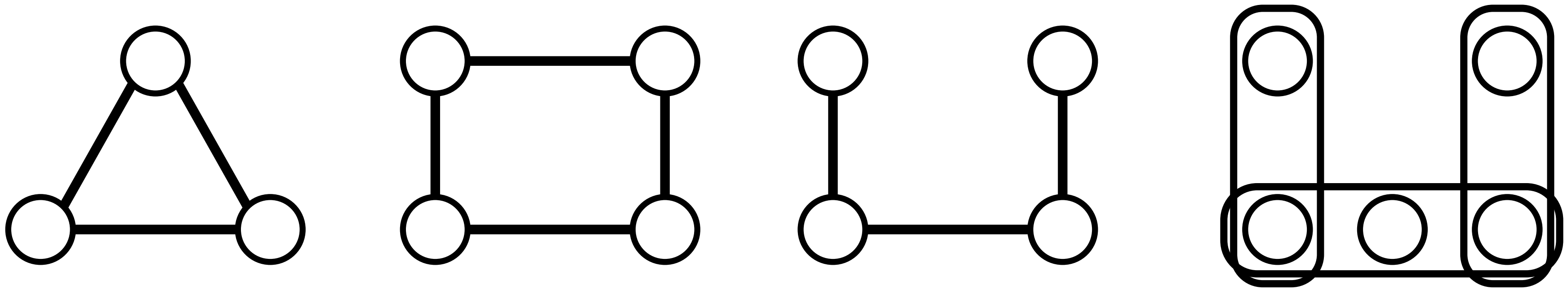
THEOREM 3.4. *Deciding whether a directed or undirected graph $G = (V, E)$ contains simple cycles of length exactly $2k - 1$ and of length exactly $2k$, and finding such cycles if it does, can be done in $O(E^{2-1/k})$ time.*

PROOF. We describe an $O(E^{2-1/k})$-time algorithm for finding a C_{2k} in a directed graph $G = (V, E)$. The details of all the other cases are similar. Let $\Delta = E^{1/k}$. A vertex in G whose degree is at least Δ is said to be of *high degree*. The graph $G = (V, E)$ contains at most $2E/\Delta = O(E^{1-1/k})$ high-degree vertices. We check, using Monien's

= "heavy": $\Delta = E^{1/2}$ for triangle

Source: Alon, Yuster, Zwick. "Finding and counting given length cycles", Algorithmica, 1997. https://doi.org/10.1007/BF02523189

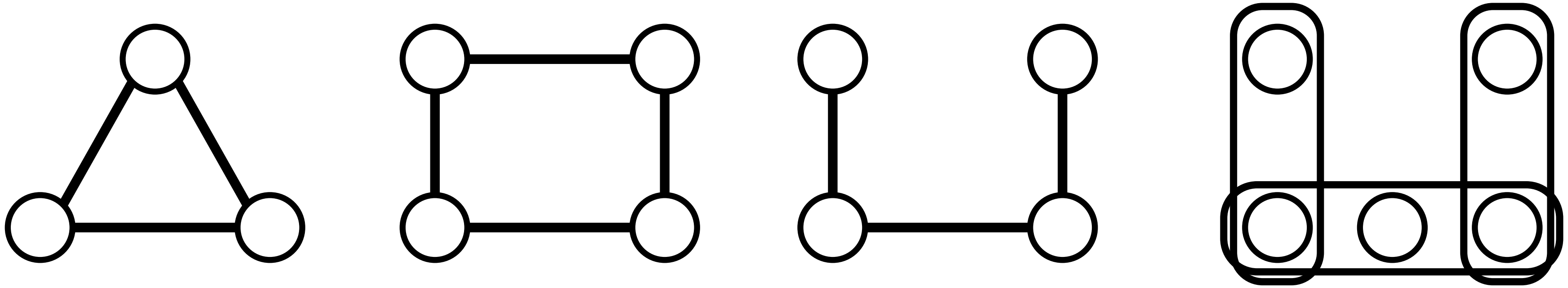
Examples



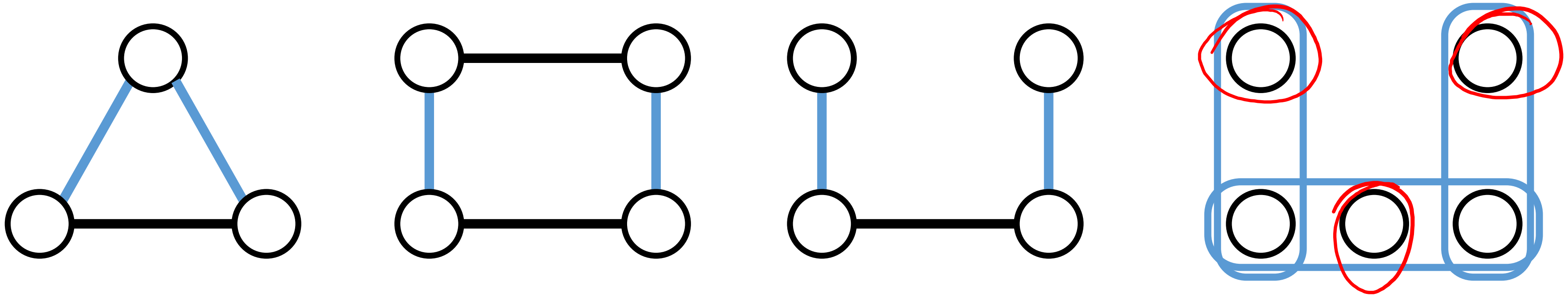
Min edge cover (α) : ? ? ? ?

Max independ. (vertex) set (ρ): ? ? ? ?

Examples



Min edge cover (α) :

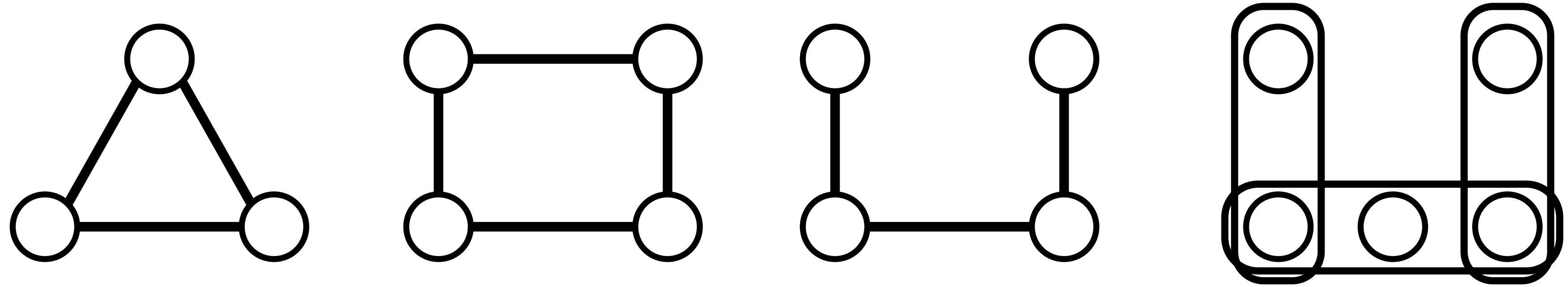


2 2 2 3

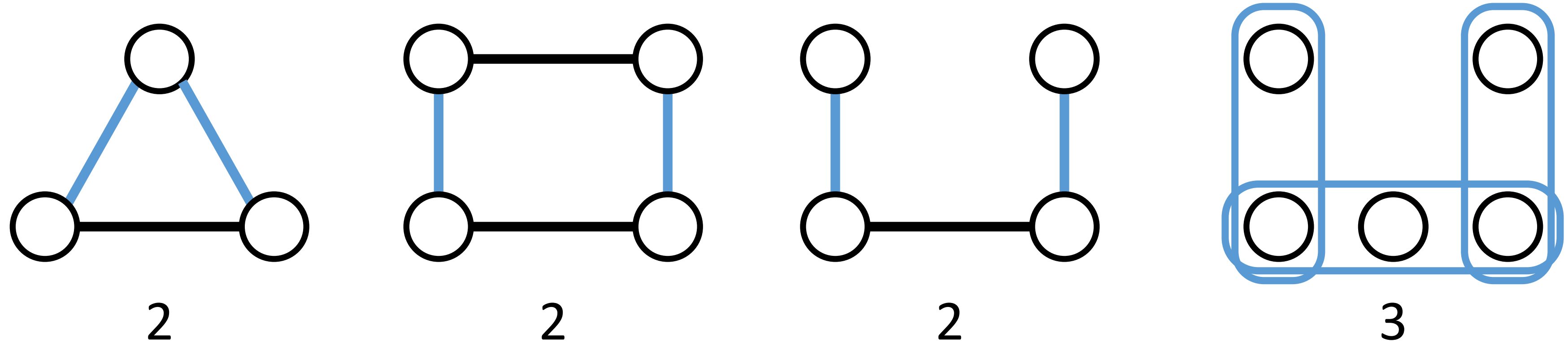
Max independ. (vertex) set (ρ):

? ? ? ?

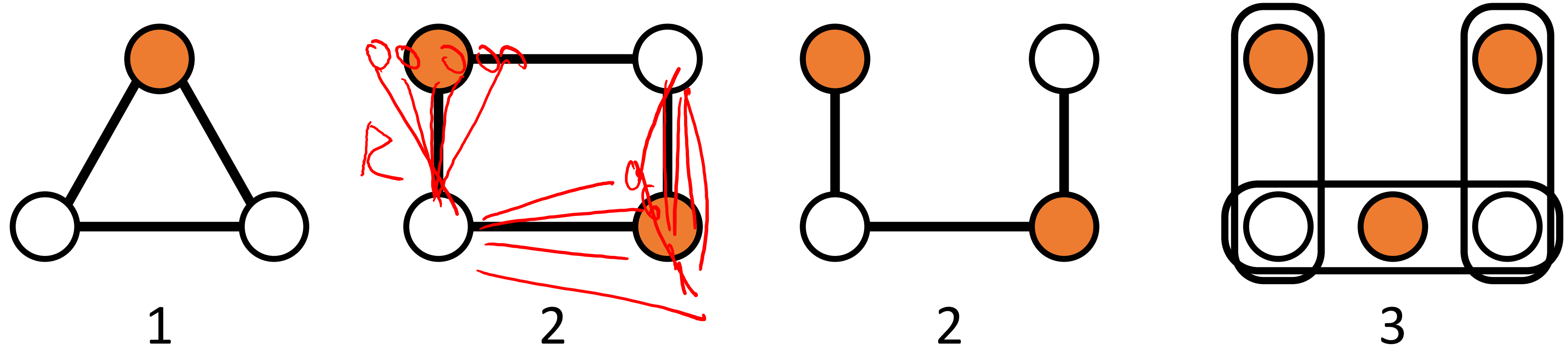
Examples



Min edge cover (α) :



Max independ. (vertex) set (ρ):



Pointers to some related work

- "*AGM bound*": Atserias, Grohe, Marx. *Size bounds and query plans for relational joins*. SIAM J. Comput. 2013. https://doi.org/10.1137/110859440 (also FOCS 2008)
- "*Worst-Case Optimal (WCO) joins*": Ngo, Porat, Re, Rudra. *Worst-case optimal join algorithms*. JACM 2018. https://doi.org/10.1145/3180143 (also PODS 2012)
- "*FAQ paper*": Khamis, Ngo, Rudra. *FAQ: Questions Asked Frequently*. PODS 2016. https://doi.org/10.1145/2902251.2902280 (see also SIGMOD record 2017).
- Khamis, Ngo, Suciu. *What do Shannon-type inequalities, submodular width, and disjunctive Datalog have to do with one another?* PODS 2017. https://doi.org/10.1145/3034786.3056105
- Robertson, Seymour. *Graph minors. II. Algorithmic aspects of tree-width*. Journal of Algorithms. 1986. https://doi.org/10.1016/0196-6774(86)90023-4
- Chekuri, Rajaraman. *Conjunctive query containment revisited*. Elsevier Theoretical Computer Science 2000. https://doi.org/10.1016/S0304-3975(99)00220-0
- Gottlob, Leone, Scarcello. *Hypertree Decompositions and Tractable Queries*. JCSS 2002. https://doi.org/10.1006/jcss.2001.1809
- Grohe, Marx. *Constraint Solving via Fractional Edge Covers*. ACM Trans. Algorithms 2014. https://doi.org/10.1145/2636918
- Marx. *Tractable Hypergraph Properties for Constraint Satisfaction and Conjunctive Queries*. JACM 2014. https://doi.org/10.1145/2535926
- Alon, Yuster, Zwick. *Finding and counting given length cycles*. Algorithmica 1997. https://doi.org/10.1007/BF02523189

Outline: T3-2: Cyclic conjunctive queries

- T3-1: Acyclic conjunctive queries
- T3-2: Cyclic conjunctive queries
  - 2SAT (a detour)
  - Tree decompositions
  - Decompositions of hypertrees
  - Duality in Linear programming (a quick primer)
  - AGM bound (maximal result size for full CQs) and Worst-case optimal joins for the triangle query
  - Worst-case optimal joins & the 4-cycle
  - Optimal joins & the 4-cycle

not covered this year!