On the Complexity of Frequent Subtree Mining in Very Simple Structures

Pascal Welke, Tamás Horváth, Stefan Wrobel

Fraunhofer IAIS

universitätbonn

Problem Setting

Given a database D of graphs and a frequency threshold,

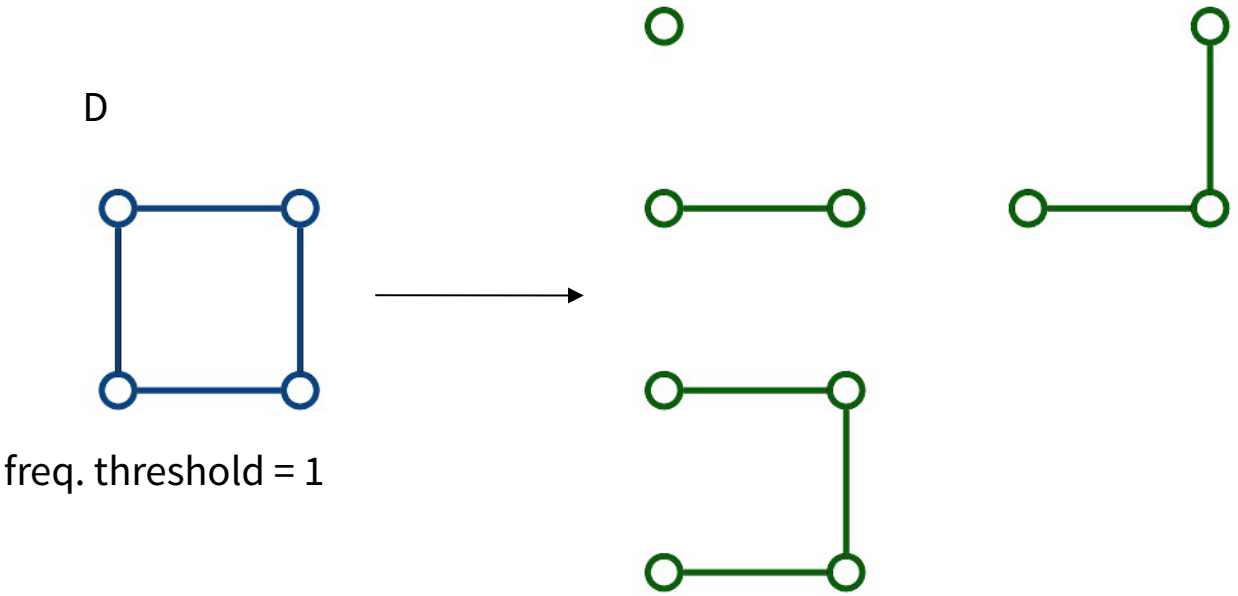
list the set of frequent connected subtrees

Problem Setting

Given a database D of graphs and a frequency threshold,

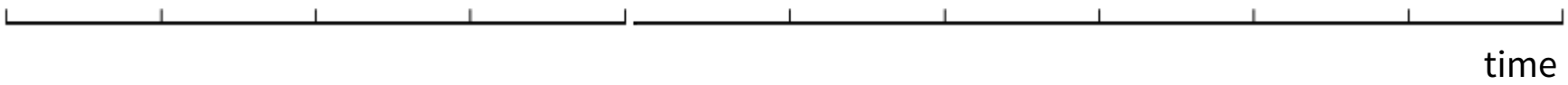
list the set of frequent connected subtrees

Example:



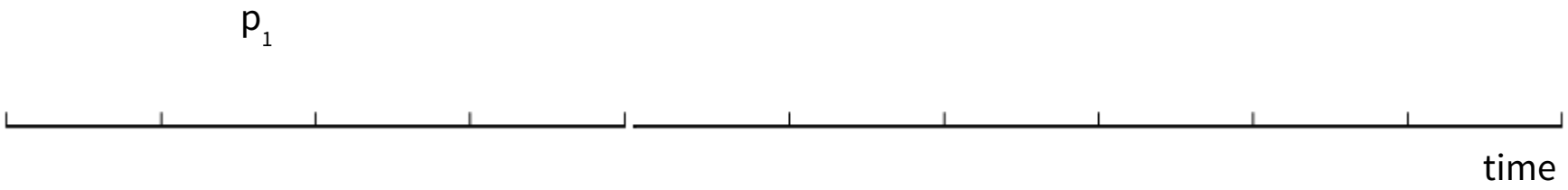
Efficiency Measures

Polynomial delay: Time between printing consecutive patterns is polynomial in the size of the database D



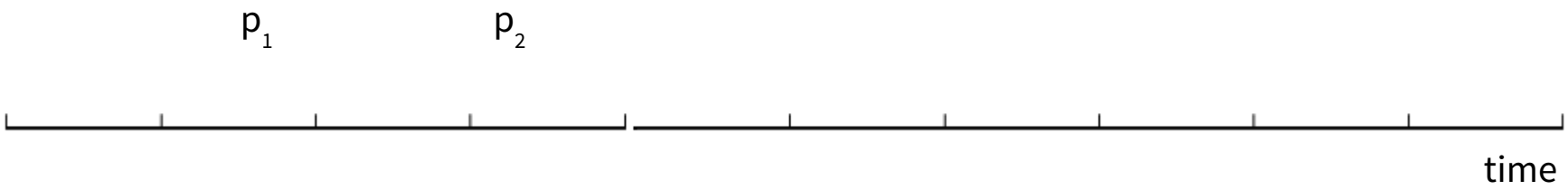
Efficiency Measures

Polynomial delay: Time between printing consecutive patterns is polynomial in the size of the database D



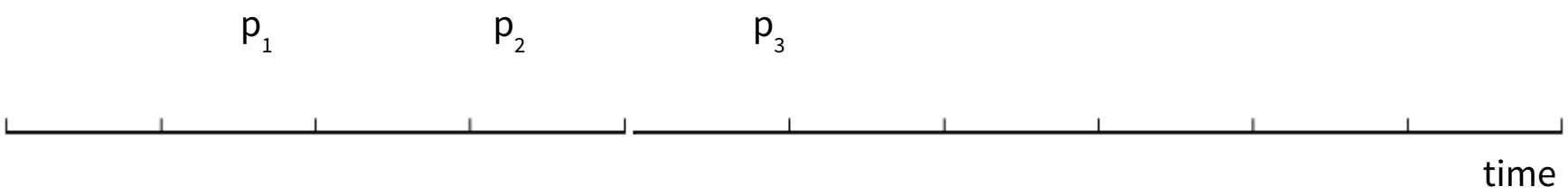
Efficiency Measures

Polynomial delay: Time between printing consecutive patterns is polynomial in the size of the database D



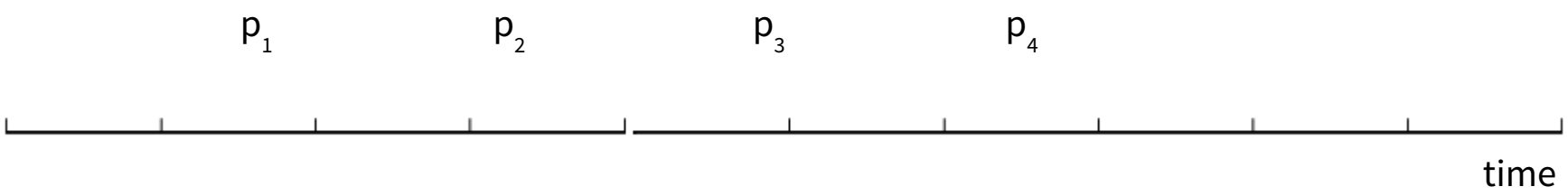
Efficiency Measures

Polynomial delay: Time between printing consecutive patterns is polynomial in the size of the database D



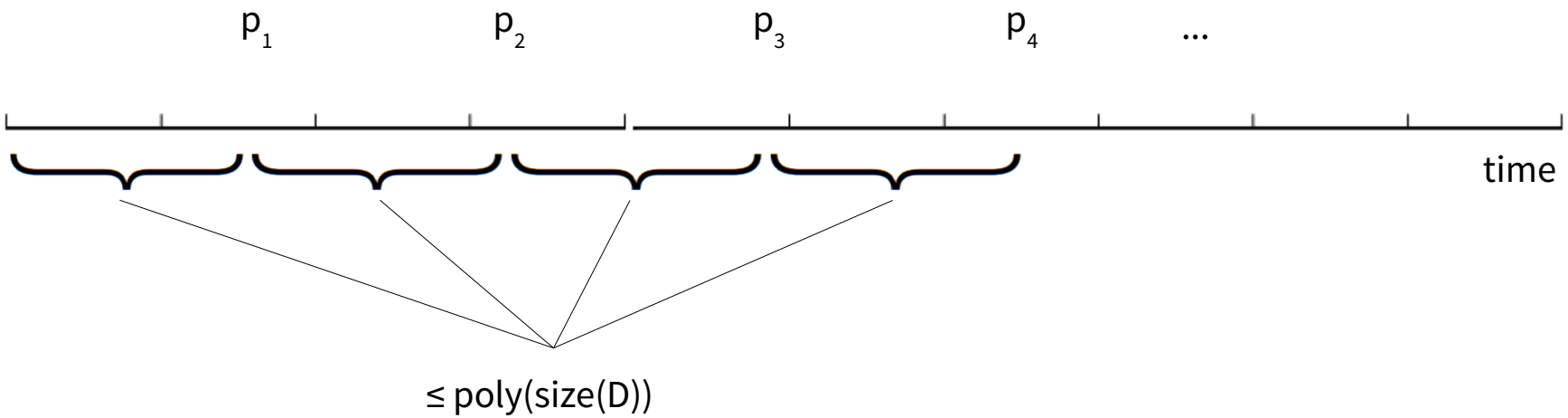
Efficiency Measures

Polynomial delay: Time between printing consecutive patterns is polynomial in the size of the database D



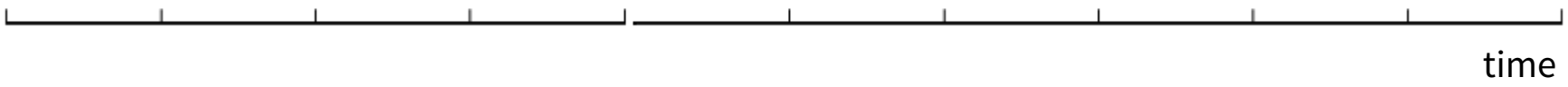
Efficiency Measures

Polynomial delay: Time between printing consecutive patterns is polynomial in the size of the database D



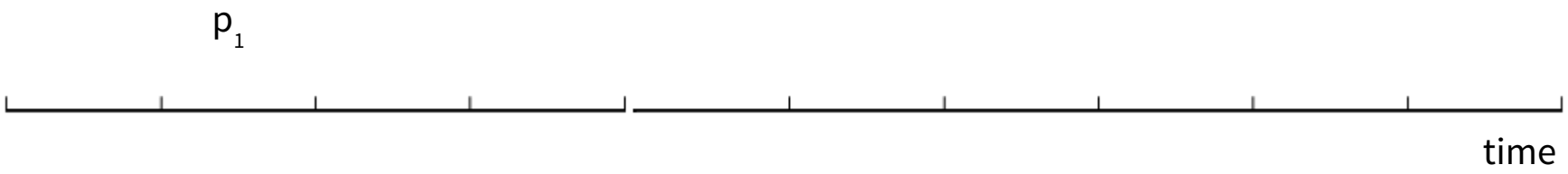
Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far

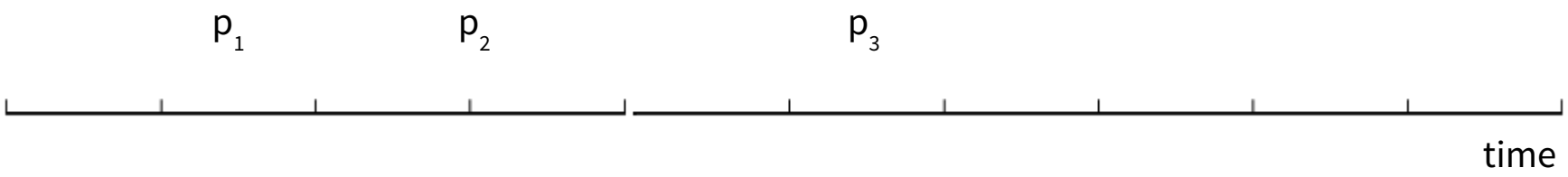


Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far

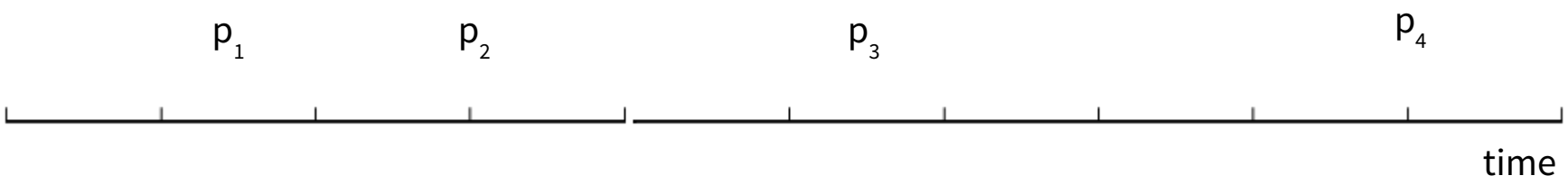
Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



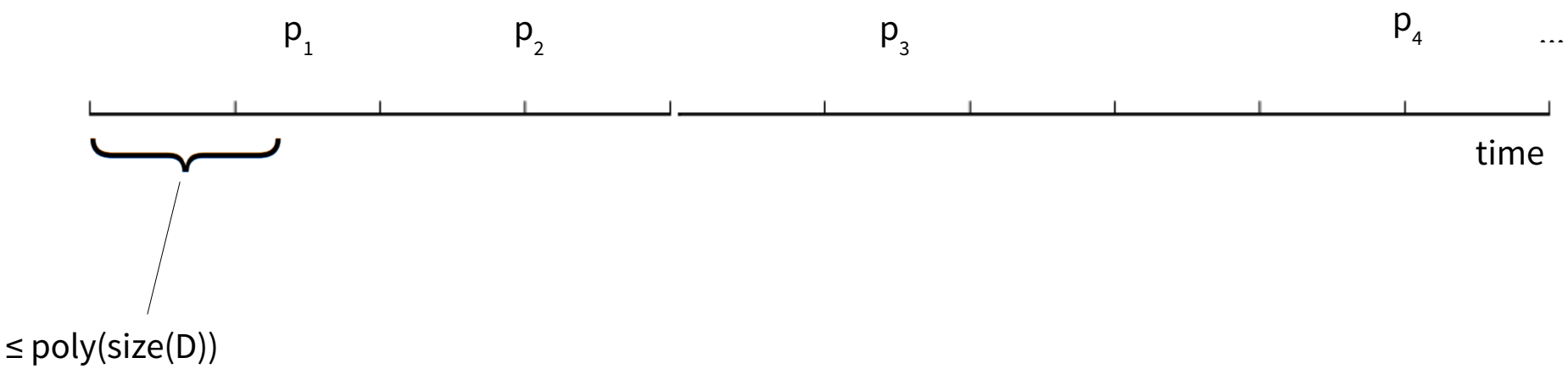
Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



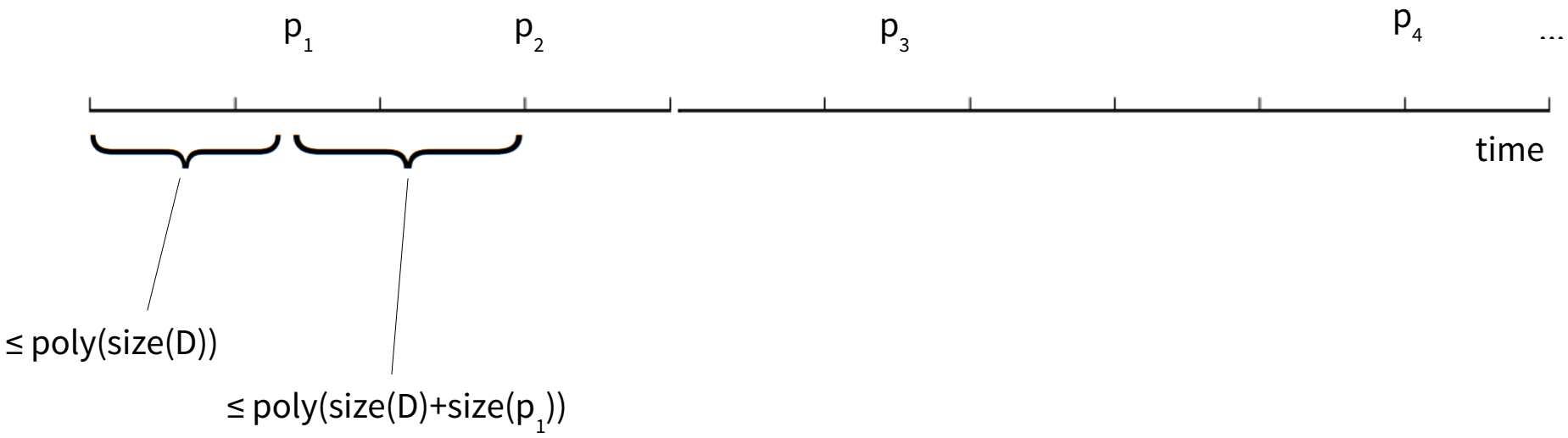
Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



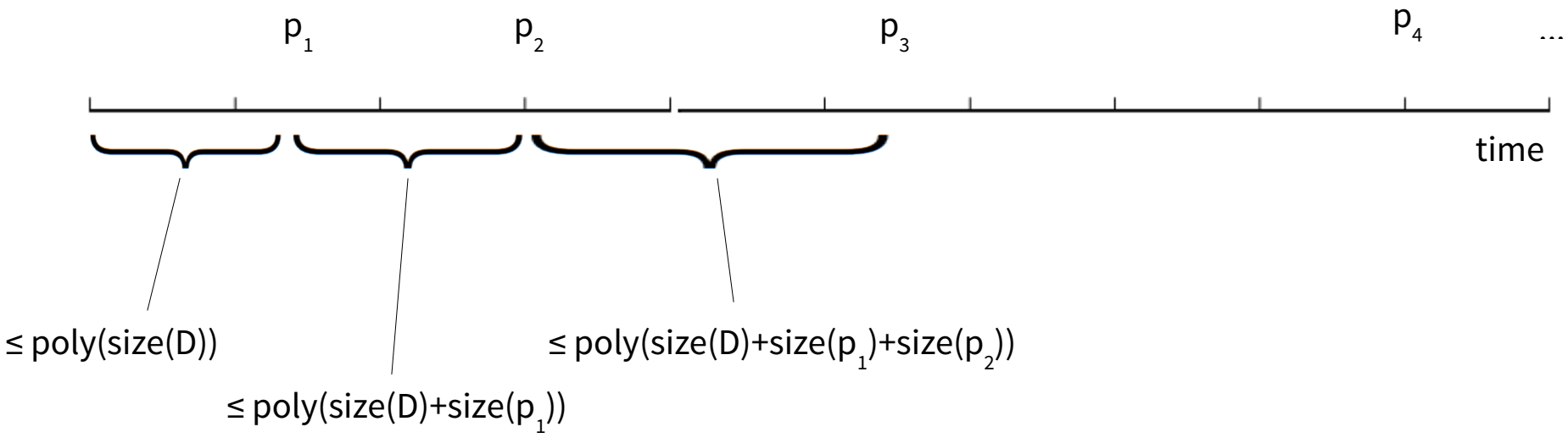
Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



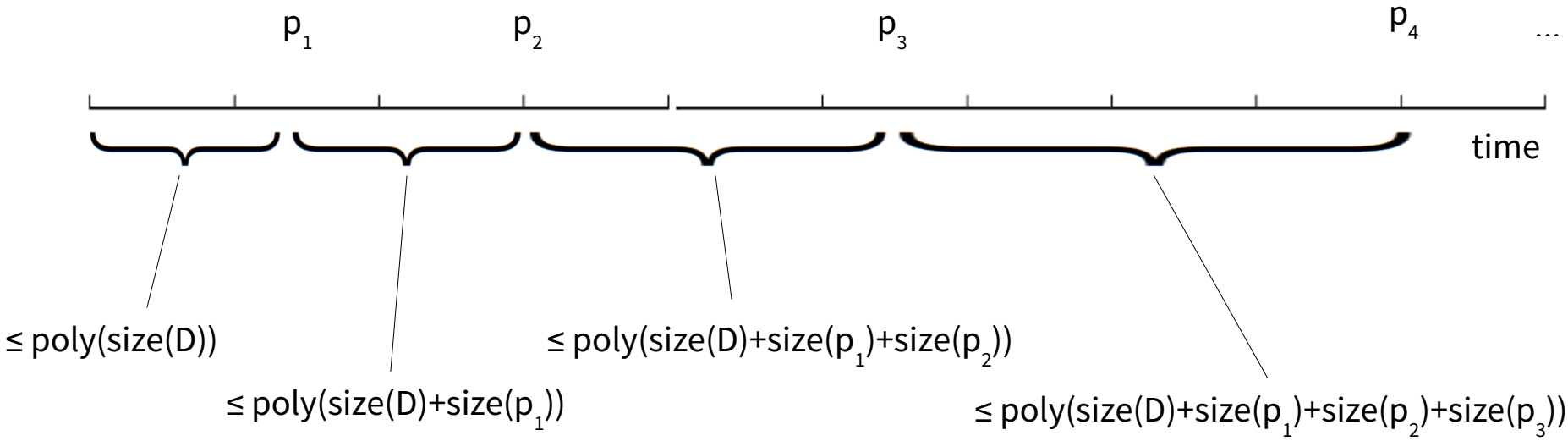
Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



Efficiency Measures

Incremental polynomial time: Time between printing consecutive patterns is polynomial in the size of the database D and the output so far



Research Question

- Research so far:

  Identification of graph classes allowing incremental polynomial time enumeration

Bounded Treewidth

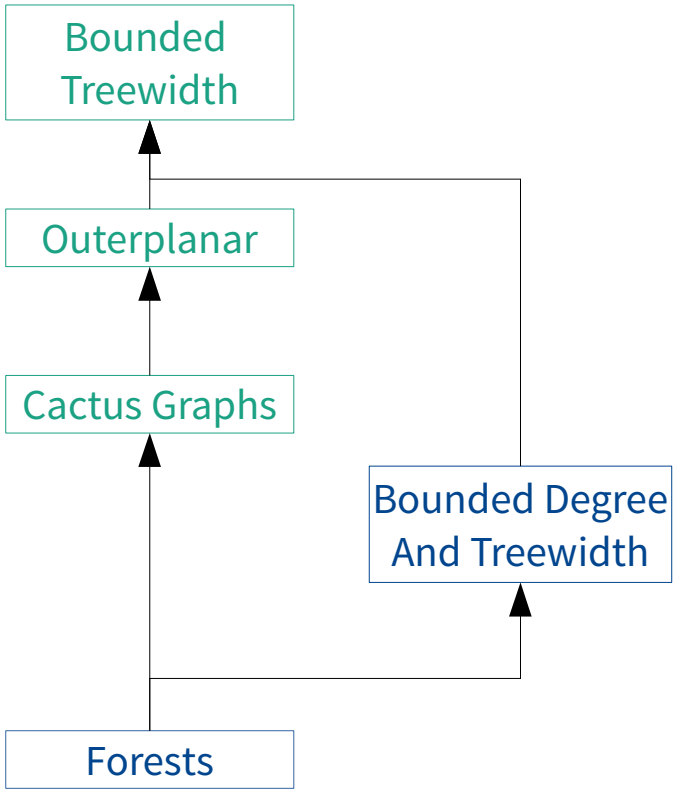
Research Question

- Research so far:

  Identification of graph classes allowing incremental polynomial time enumeration

- Question for this work:

  Identification of graph classes allowing polynomial delay enumeration

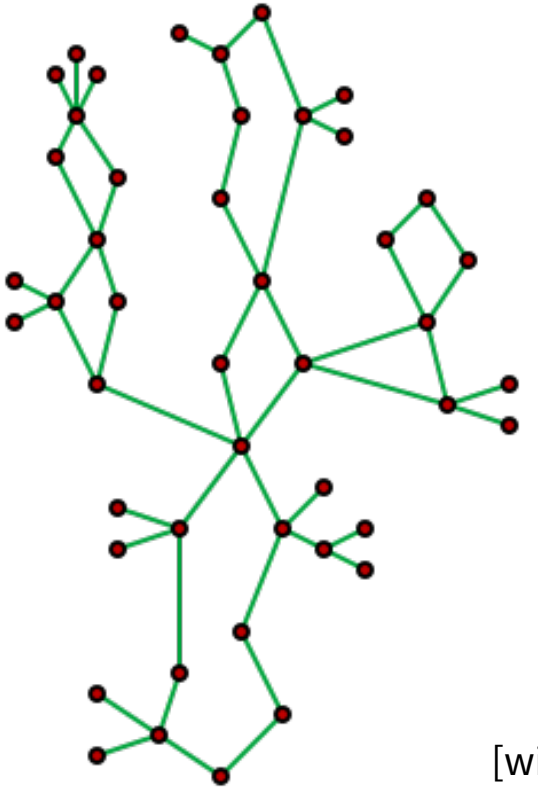


Main Result

Theorem: Frequent subtree mining is possible in cactus graph databases with bounded cycle degree with polynomial delay.

Main Result

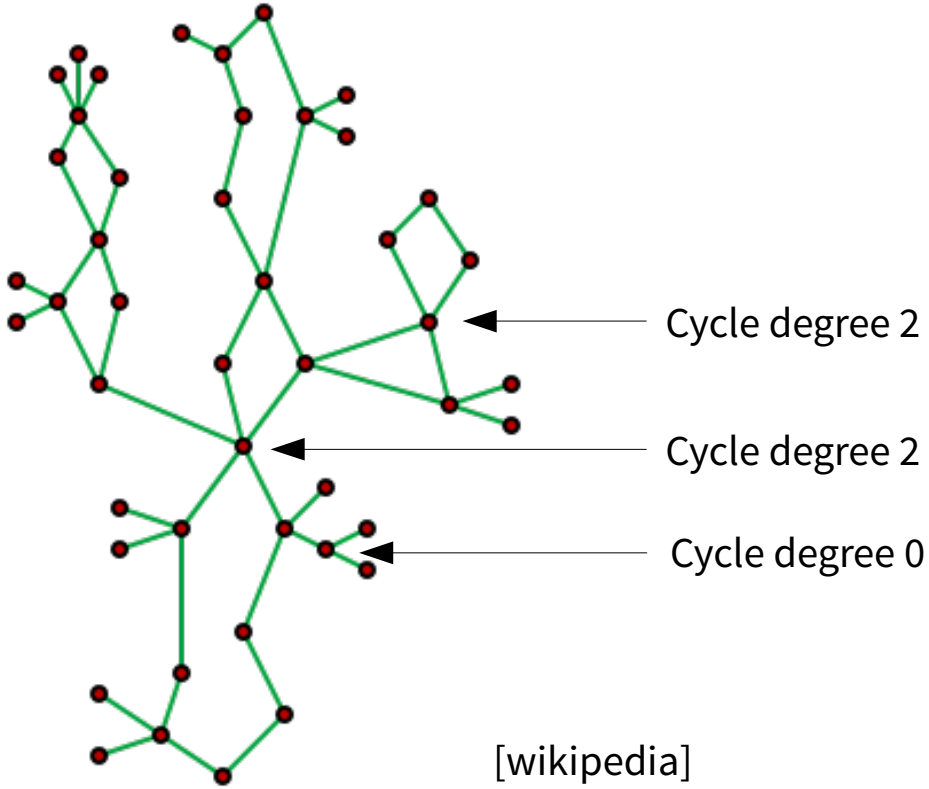
Theorem: Frequent subtree mining is possible in cactus graph databases with bounded cycle degree with polynomial delay.



[wikipedia]

Main Result

Theorem: Frequent subtree mining is possible in cactus graph databases with bounded cycle degree with polynomial delay.



[wikipedia]

Main Result

Theorem: Frequent subtree mining is possible in cactus graph databases with bounded cycle degree with polynomial delay.

Proof Sketch: Based on a generalization of [Shamir, Tsur] for pattern matching and a generic subgraph mining framework.

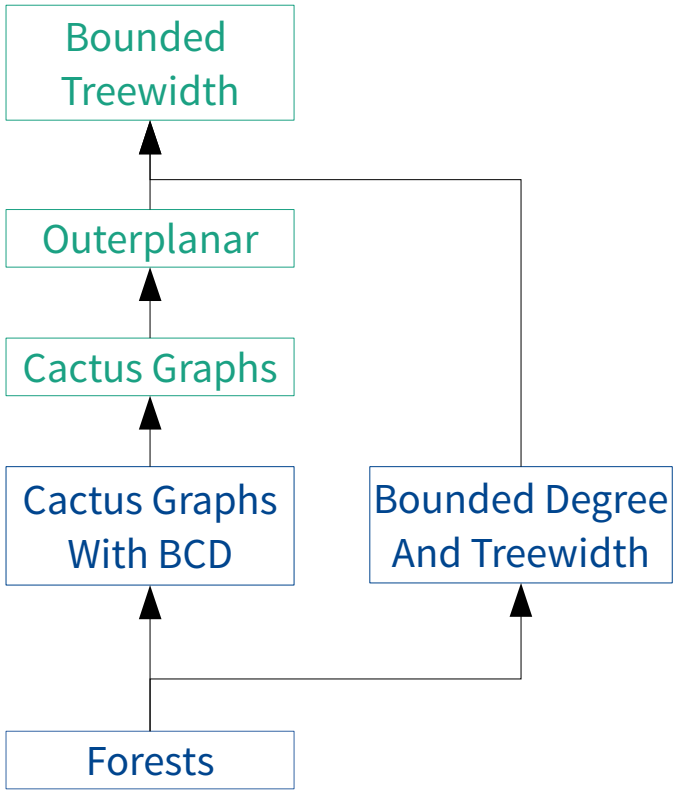
Why is this interesting?

- Many molecular graphs are cactus graphs
- Molecular graphs have very small cycle degrees

Data Set	Size	Max.CycD	Med.CycD	Cactus
NCI-HIV	42,687	4	1	50.08%
NCI-2012	249,533	3	1	63.81%
ZINC-leadlike	8,946,757	2	1	72.84%

Discussion of the Result

- Cycle degree seems a crucial parameter for polynomial delay enumeration
- If removed, there are two options:
  - a) Polynomial delay mining is not possible → P ≠ NP
  - b) Polynomial delay mining is possible → polynomial delay mining for NP-complete matching operators is possible



References

- T. Akutsu. *A polynomial time algorithm for finding a largest common subgraph of almost trees of bounded degree*. IEICE transactions on fundamentals of electronics, communications and computer sciences, 76(9):1488–1493, 1993.
- Y. Chi, Y. Yang, and R. R. Muntz. *Indexing and mining free trees.* In ICDM, pages 509–512. IEEE Computer Society, 2003
- T. Horváth and J. Ramon. *Efficient frequent connected subgraph mining in graphs of bounded tree-width.* Theor. Comput. Sci., 411(31-33):2784–2797, 2010.
- R. Shamir and D. Tsur. *Faster subtree isomorphism*. In Theory of Computing and Systems, 1997, pages 126–131. IEEE, 1997.
- M. M. Syslo. *Subgraph isomorphism problem for outerplanar graphs*. Theor. Comput. Sci., 17(1):91–97, 1982.