Graphs can be succinctly indexed for pattern matching in $O(|E|^2 + |V|^{5/2})$ time

Nicola Cotumaccio

GSSI, L'Aquila, Italy

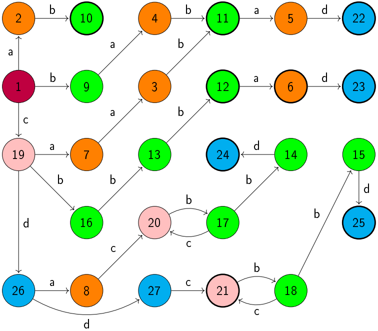
Wheeler graphs

Wheeler graphs generalize the nice properties of De Bruijn graphs.

- A Wheeler graph on the alphabet Σ with n nodes and e edges can be stored using only $2(e+n) + e\log|\Sigma| + |\Sigma|\log e + o(n + e\log|\Sigma|)$ bits.
- This representation allows to decide whether a string α matches the graph in only $O(|\alpha|\log|\Sigma|)$ time.

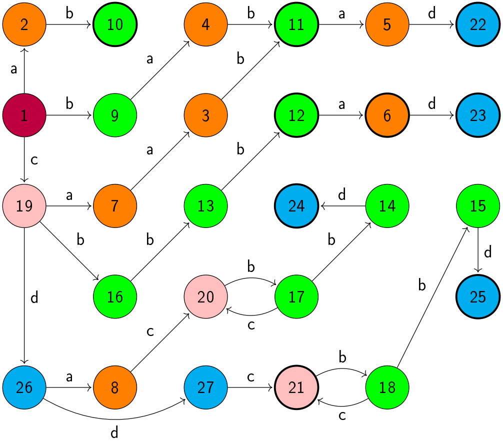
Wheeler graphs

- Wheeler graphs are graphs endowed with a TOTAL order $\leq$ on the set of all nodes.
- We assume that all edges entering the same node have the same label.
- Here are the properties that the total order $\leq$ must satisfy.



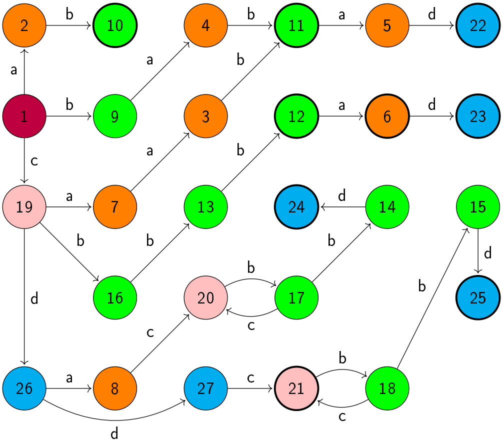
Wheeler graphs

Nodes without incoming edges come first.



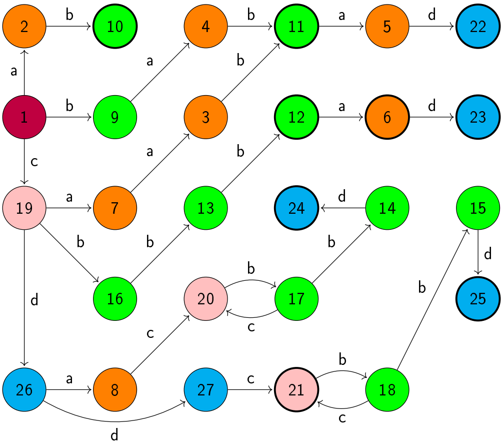
Wheeler graphs

All nodes reached by a come before all nodes reached by b, which come before all nodes reached by c...



Wheeler graphs

Equally-labeled edges must respect the total order (think of $(7, 3, a)$, $(9, 4, a)$, $(6, 23, d)$, $(15, 25, d)$).

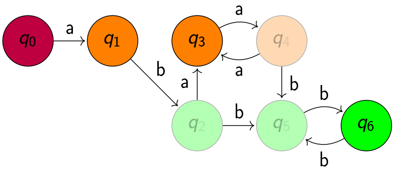


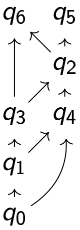
Limitations

- Expressive power: most graphs are not Wheeler.
- Tractability: deciding whether a graph admits a total order with the desired properties is an NP-hard problem.

Expressive power

- Even if most graphs do not admit a total order with the desired properties, every graph admits a PARTIAL order with the desired properties (a *co-lex order*).
- A string α can be matched in $O(p^2|\alpha| \log |\Sigma|)$ time, where p is the width of the partial order.
- Again intractability: finding the minimum p is NP-hard.





Tractability

- In this paper, we show that we can make the problem tractable, while retaining the expressive power (and improving space and time bounds).
- Why deciding whether a graph is Wheeler is difficult? Intuitively, because 2-SAT is an easy problem, while 3-SAT is NP-complete!

$$
\begin{cases}
x_{u<v} & \forall u \neq v \text{ with } \lambda(u) \prec \lambda(v) \text{ (Axiom 1)} \\
x_{u'<v'} \implies x_{u<v} & \forall (u', u, a), (v', v, a) \in E, u' \neq v', u = v \\
 & \text{(Axiom 2)} \\
x_{u<v} \vee x_{v<u} & \forall u \neq v \text{ (Comparability)} \\
x_{u<v} \implies \neg x_{v<u} & \forall u \neq v \text{ (Antisymmetry)} \\
(x_{u<v} \wedge x_{v<z}) \implies x_{u<z} & \forall u \neq v, v \neq z, u \neq z \text{ (Transitivity)}
\end{cases}
$$

Tractability

- Do we really need antisymmetry? The answer is no.
- Do we really need transitivity? The answer is no.
- The solution is to consider ARBITRARY RELATIONS with the desired properties (a *co-lex relation*).

Co-lex relations

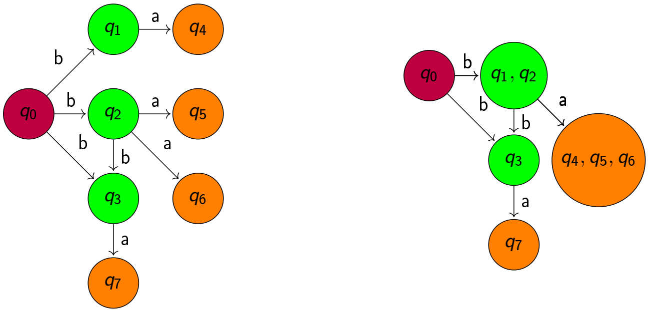
- Now algebraically everything becomes cleaner.
- Every graph admits a co-lex relation containing all co-lex relations on the graph (the *maximum co-lex relation*).
- Most importantly, the maximum co-lex relation can be computed in polynomial time (in $O(|E|^2)$ time).
- Furthermore, the maximum co-lex relation is always transitive (so it is a preorder), but in general it is not antisymmetric.

Co-lex relations

- But can we still solve pattern matching queries?
- Not only the answer the yes, but we can also compress the graph!
- The idea is the following: starting from the graph, collapse some nodes in such a way that:
  - There is a correspondence between patterns on the original graph and patterns on the quotient graph.
  - Apply the results on co-lex orders (that is, PARTIAL ORDERS) in the quotient graph.
  - Prove that the problem of determining the minimum width p is easy on a quotient graph (we know that it is NP-complete on general graphs).

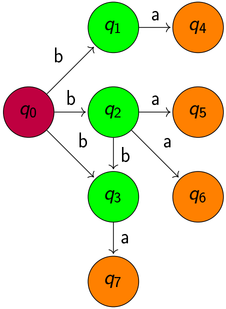
Co-lex relations

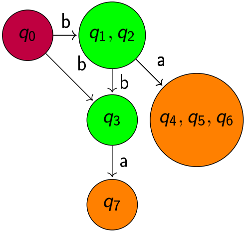
- We need a quotient graph because the maximum co-lex relation need not be antisymmetric.
- Intuitively, if two nodes are comparable in both directions in the maximum co-lex relation R, then we can read the same strings when we proceed backward (for example both $(q_4, q_5) \in R$ and $(q_5, q_4) \in R$).



Co-lex relations

- As a consequence, from a pattern matching perspective we can simply collapse two nodes comparable in both directions.
- It can be proved that every node obtained by collapsing two or more nodes hat at most one ingoing edge in the quotient graph.





Co-lex relations

- Quotient graphs admit a co-lex order (a PARTIAL ORDER) which contains all co-lex orders on the graphs, the *maximum co-lex order* (this is NOT true for general graphs: every graph admits the maximum co-lex *relation* but in general a graph does not admit the maximum co-lex *order*).
- The maximum co-lex order is automatically the best co-lex order: the one yielding the minimum width p.
- Such a best co-lex order can be computed in polynomial time on quotient graphs, because it is naturally induced by the maximum co-lex relation on the original graph (while determing a best co-lex order on a general graph is NP-hard).

The polynomial-time algorithm

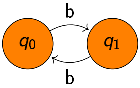
We can index a graph G for pattern matching as follows.

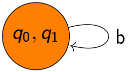
- Compute the maximum co-lex relation on G (which always exists).
- Build a quotient graph by collapsing the nodes of G comparable in both directions.
- Compute the maximum co-lex order on the quotient graph (which always exists on quotient graphs, and it is induced by the maximum co-lex relation on G).
- Apply the previously known results on co-lex orders to the quotient graph.
- Map a pattern matching query on G to the quotient graph.

A final remark

The class of Wheeler graphs is (strictly) contained in the class of all graphs such that the maximum co-lex relation has width equal to 1.

	Max. co-lex relation with $p = 1$	*Wheeler*
Representation	succinct	succinct
Pattern matching	$O(\lvert\alpha\rvert \log \lvert\Sigma\rvert)$	$O(\lvert\alpha\rvert \log \lvert\Sigma\rvert)$
Decision problem	$O(\lvert E\rvert^2)$	NP-complete





Graphs can be succinctly indexed for pattern matching in $O(|E|^2 + |V|^{5/2})$ time

Nicola Cotumaccio

GSSI, L'Aquila, Italy