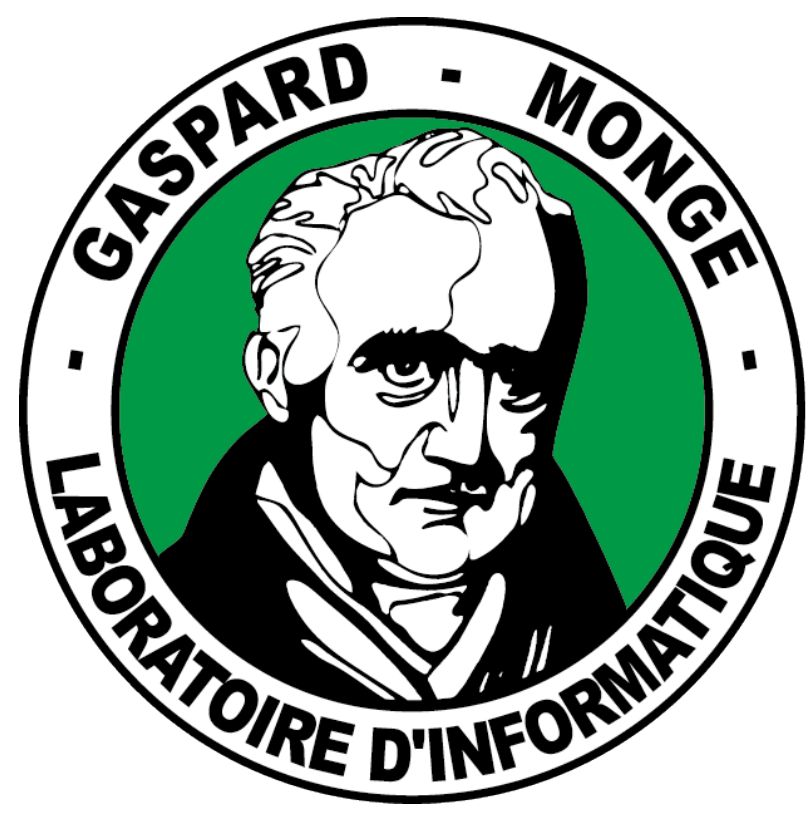




Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Foundations of Query Languages for Property Graphs

C. David, N. Francis, V. Marsault, A. Meyer

BAAM team – LIGM (UMR 8049)

Property graph databases return walks that match regular expressions

Reachability query

What are the *walks* from s to t using any number of **R**oad or **F**erry?

$$Q_1 = (\mathbf{R} + \mathbf{F})^*$$

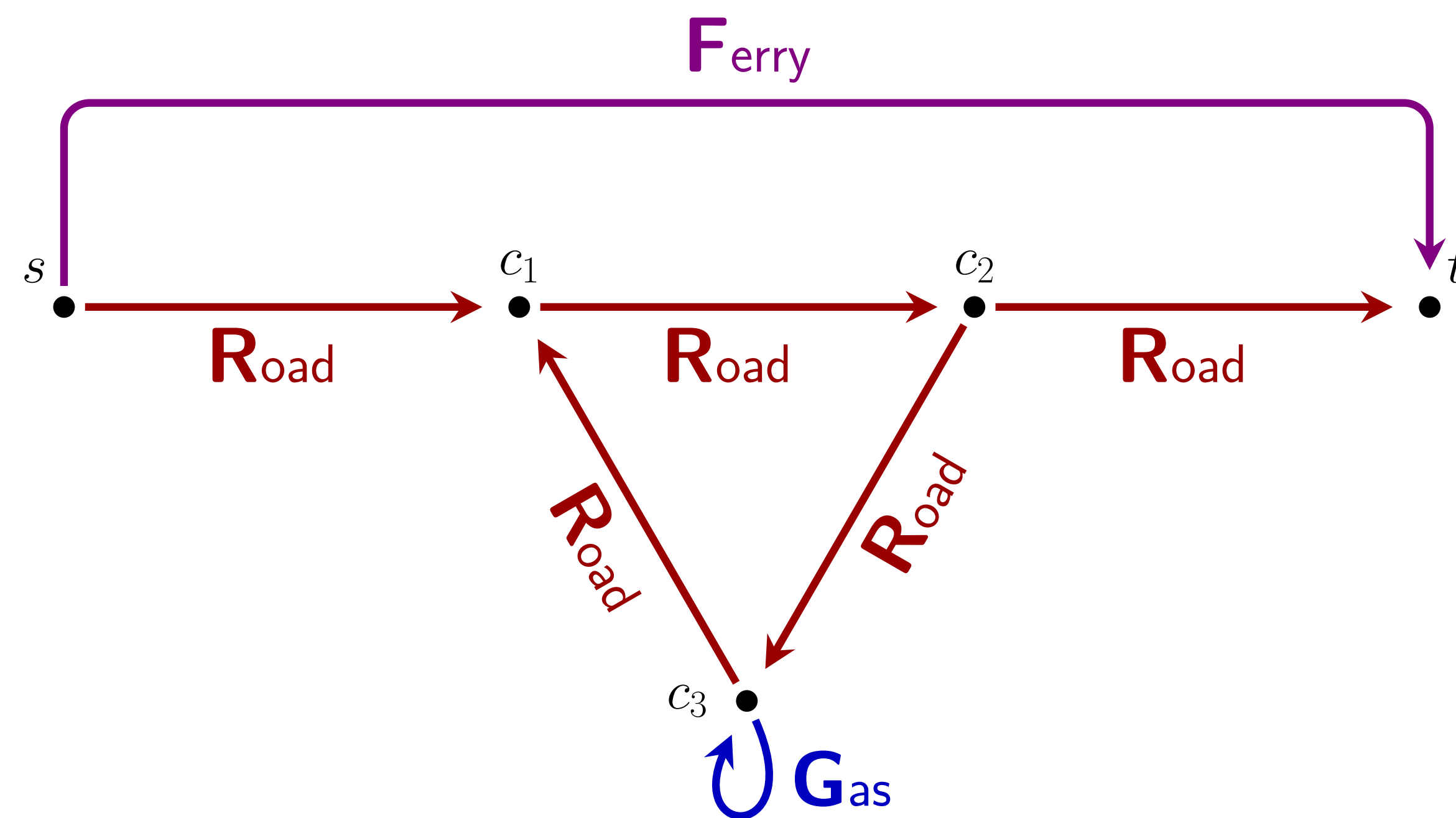
Infinitely many walks *match* Q_1 :

$$f = s \rightarrow t$$

$$r_0 = s \rightarrow c_1 \rightarrow c_2 \rightarrow t$$

⋮

$$r_k = s \rightarrow c_1 (\rightarrow c_2 \rightarrow c_3 \rightarrow c_1)^k \rightarrow c_2 \rightarrow t$$



Gas stop query

Same as Q_1 , but with a mandatory gas stop.

$$Q_2 = (\mathbf{R} + \mathbf{F})^* \mathbf{G} (\mathbf{R} + \mathbf{F})^*$$

All walks matching Q_2 use the edge $c_1 \rightarrow c_2$ more than once.

How to efficiently extract a finite and meaningful answer?

Usual approaches

Endpoint semantics

- ▷ Main theoretical semantics
- ▷ Return endpoints of matches

Trail semantics

- ▷ Major semantics in practice
- ▷ Return non-repeating walks

Shortest semantics

- ▷ Main theoretical semantics
- ▷ Return walks of min. length

If ③ is not required

All shortest enumeration [1]

Query evaluation under **shortest semantics** can be enumerated with $O(|Q| \times \lambda)$ delay after a $O(|Q| \times |D|)$ preprocessing, where λ is the size of the next output.

Fails ①
 Q_1 and Q_2 both return (s, t)

Fails ②, attempts ③
 Q_2 is NP-hard and returns $\emptyset$

Fails ③
 Q_1 returns $\{f\}$

Desirable properties

- ① Returned objects are walks
 - ② Evaluation is efficient
 - ③ Allows rich aggregation
- ... Others?

Trail is the most used, hence ② is not the most important criterion

Comparison framework (work in progress)

- ▷ Classify semantics
- ▷ Find desirable properties
- ▷ Compatibilities of those properties

New approaches

Binding-trail semantics [2]

- ▷ Return walks resulting from non-repeating computations
- ▷ Evaluation is poly-delay
- ▷ Guarantees subwalk coverage

Promising semantics (work in progress)

- ▷ Shortest coverage semantics *keeps a witness for each vertex*
- ▷ Subwalk minimal semantics *returned walks must not have matching subwalks*
- ▷ Atom coverage semantics *keeps a witness for each letter in the regular expression Q*

Takeaway: no best semantics, only good compromises!

[1] C. David, N. Francis, and V. Marsault. "Distinct Shortest Walk Enumeration for RPQs". In: *Proceedings of PODS*. ACM, 2024.

[2] C. David, N. Francis, and V. Marsault. "Run-Based Semantics for RPQs". In: *Proceedings of KR*. 2023.

[3] V. Marsault. *Matching walks that are minimal with respect to edge inclusion*. ArXiv preprint:2408.14048. 2024.