Mining and Ranking of Generalized Multi-Dimensional Frequent Subgraphs

André Petermann[1], Giovanni Micale[2], Giacomo Bergami[3], Alfredo Pulvirenti[2] and Erhard Rahm[1]

1: UNIVERSITÄT LEIPZIG

2: University of Catania

3: University of Bologna

ScaDS DRESDEN LEIPZIG

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Förderkennzeichen: 01is14014

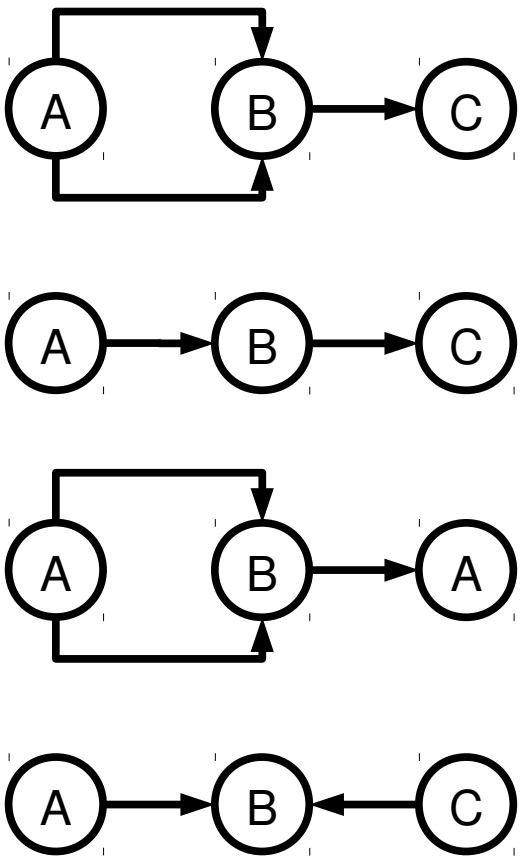
Contents

- Problem definition and motivation
- Mining algorithms
- Result ranking
- Experimental evaluation

Frequent Subgraph Mining (FSM)

- Input: Collection of graphs, threshold sup_{min}
- A graph supports a pattern if there is at least one subgraph isomorphic to the pattern
- Output: Set of frequent graph patterns
  where for all *sup*(pattern) >= sup_{min}

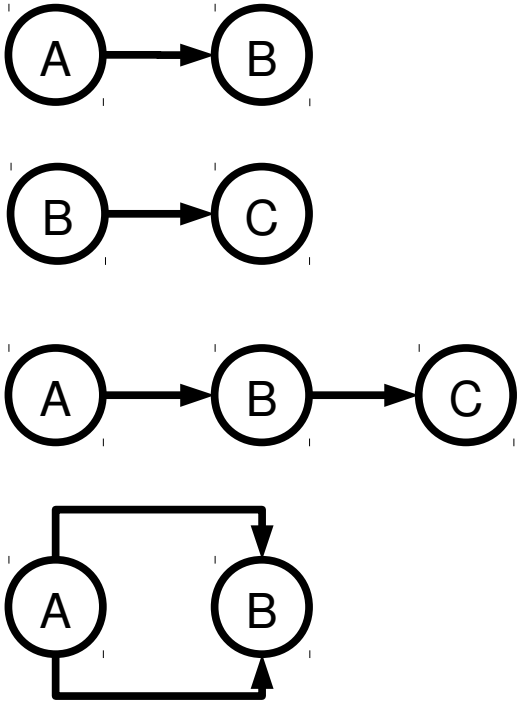
Frequent Subgraph Mining (FSM)



Input collection

FSM

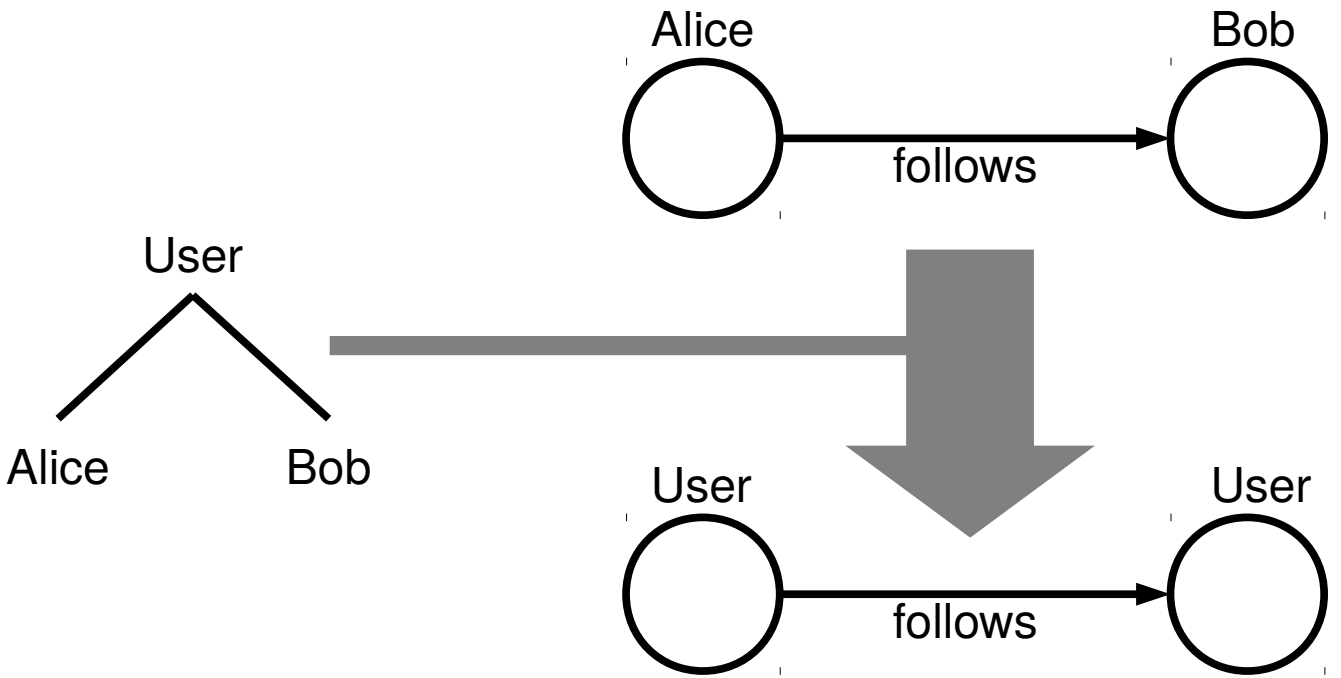
$sup_{min} = 50\%$



Output collection

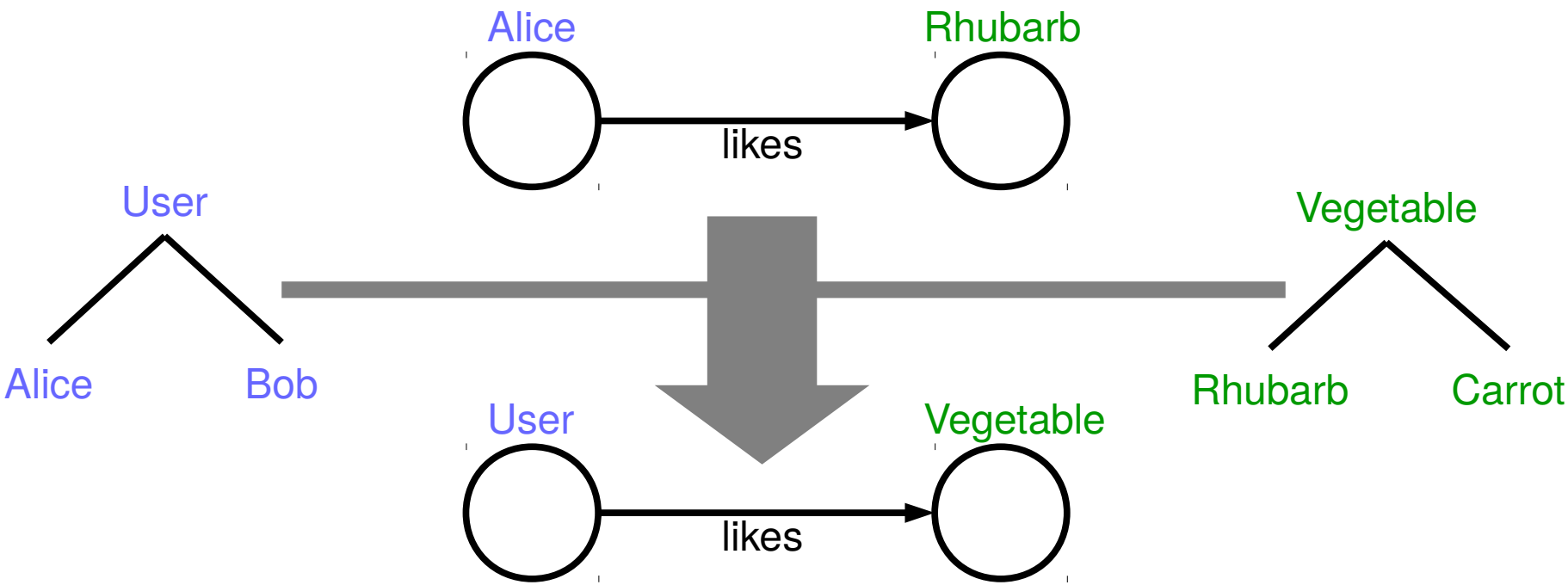
Generalized FSM

- Vertices can be attached to a single taxonomy

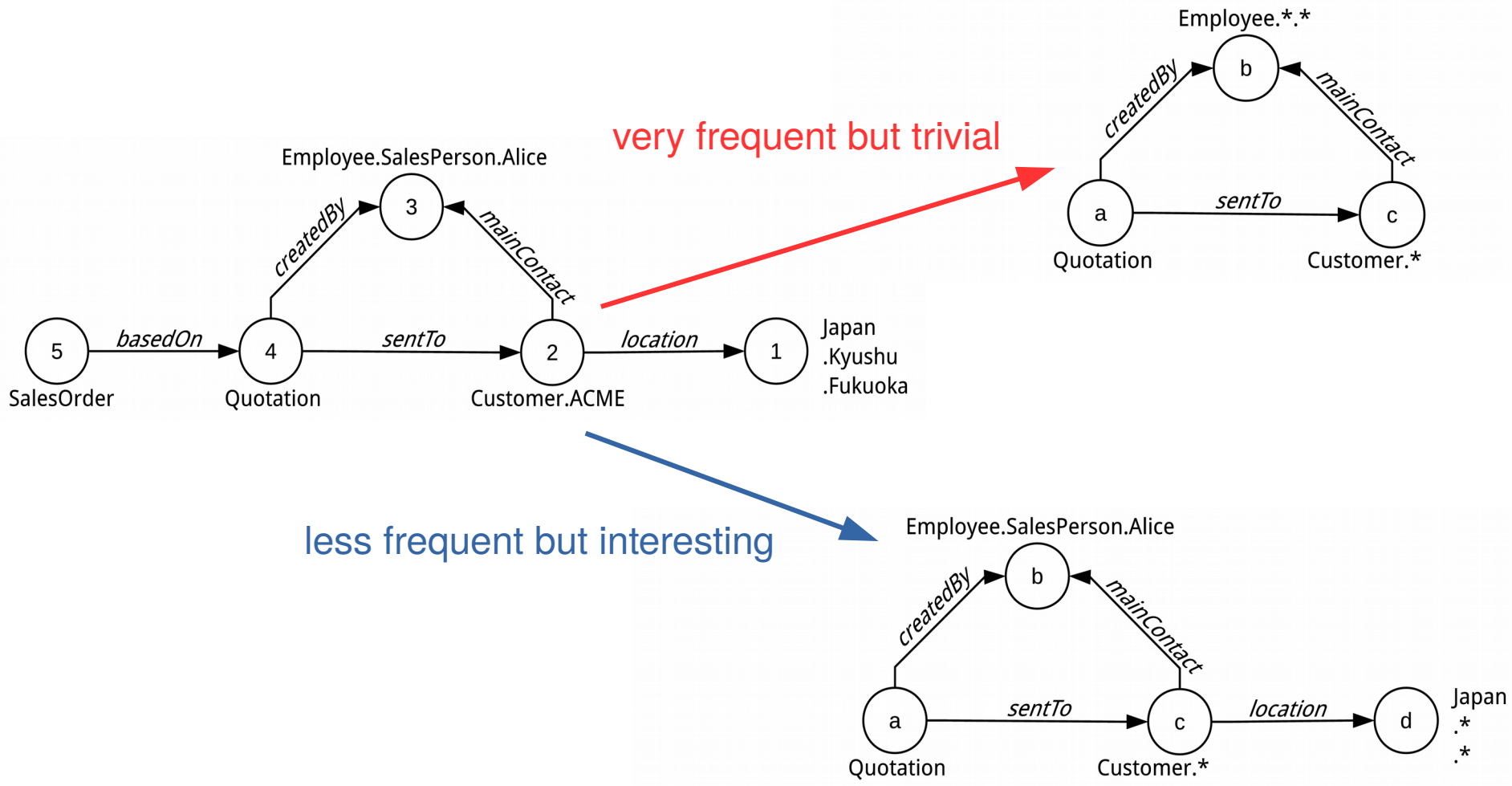


Multi-Dimensional Generalized FSM

- Vertices can be attached to multiple taxonomies



Motivation

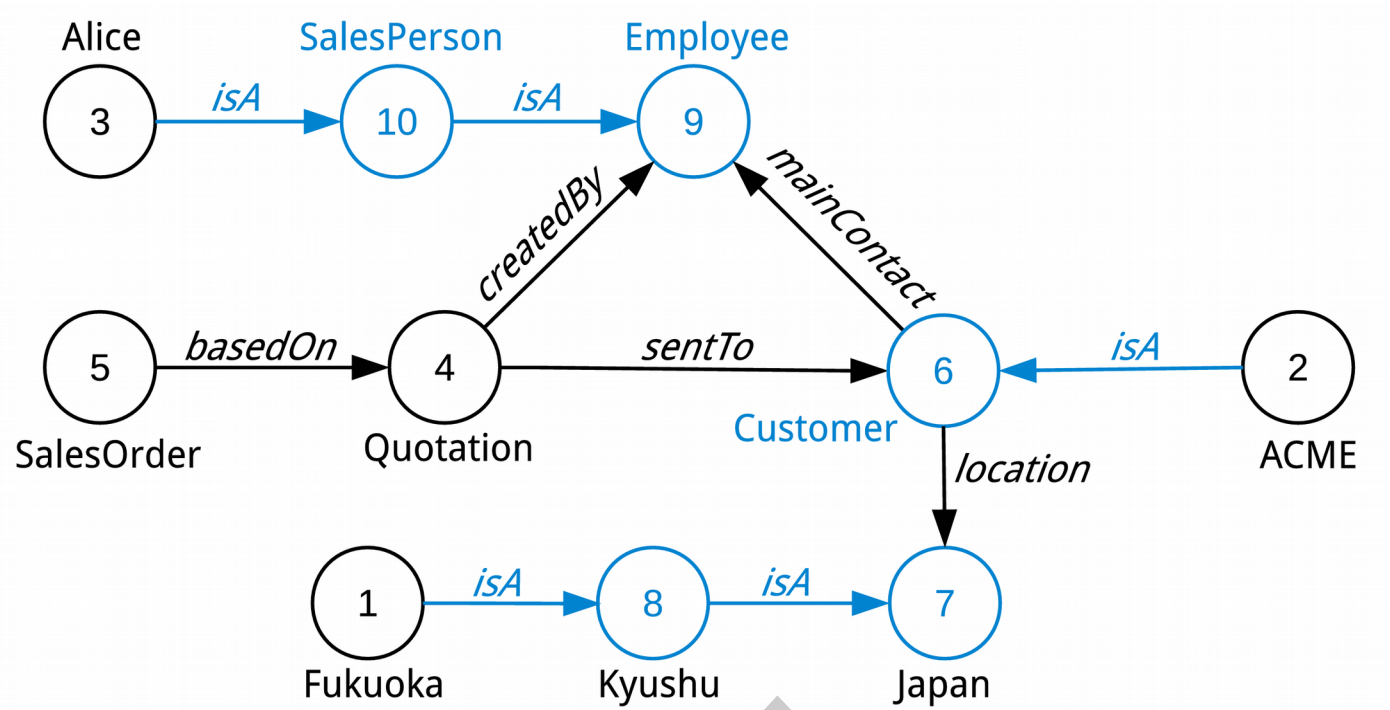

very frequent but trivial
Employee.*.*
createdBy
b
mainContact
a
sentTo
c
Quotation
Customer.*
Employee.SalesPerson.Alice
createdBy
3
mainContact
5
basedOn
4
sentTo
2
location
1
Japan
.Kyushu
.Fukuoka
SalesOrder
Quotation
Customer.ACME
less frequent but interesting
Employee.SalesPerson.Alice
createdBy
b
mainContact
a
sentTo
c
location
d
Japan
.*
.*
Quotation
Customer.*

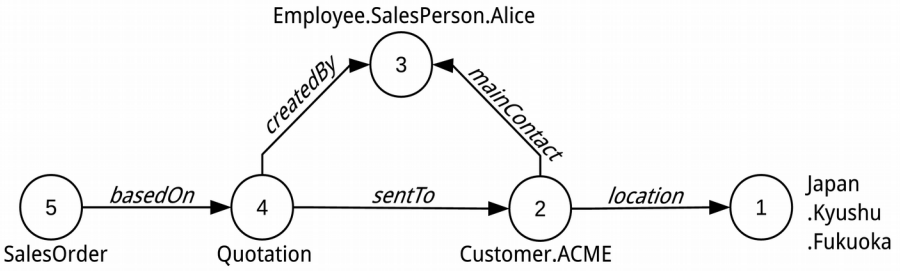
Mining algorithms

- Based on gSpan for FSM [1]
- Two methods:
  - Path substitution
  - Decomposition into FSM and Generalized Frequent Vector Mining

[1] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In Proc. IEEE Int. Conf. on Data Mining (ICDM), pages 721–724, 2002.

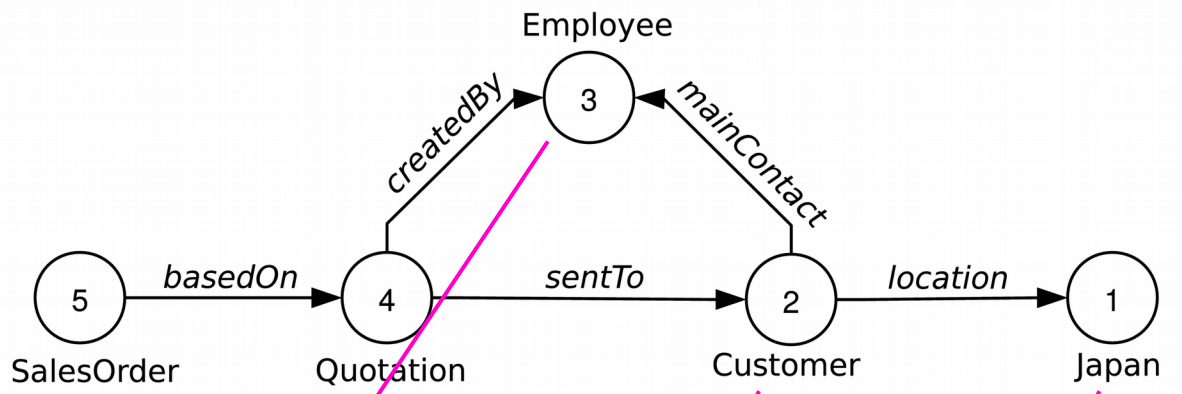
Method 1: Path substitution





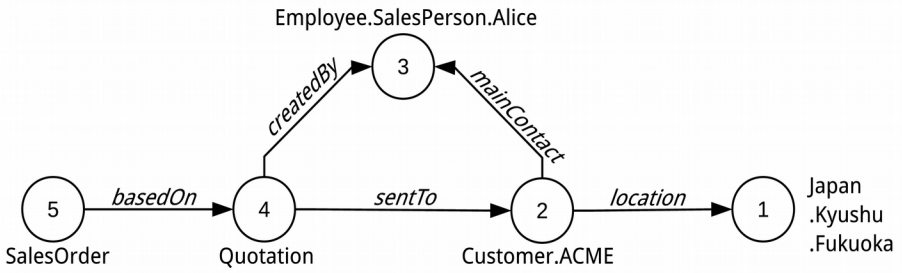
Method 2: Decomposition

Top-level pattern



Mapping

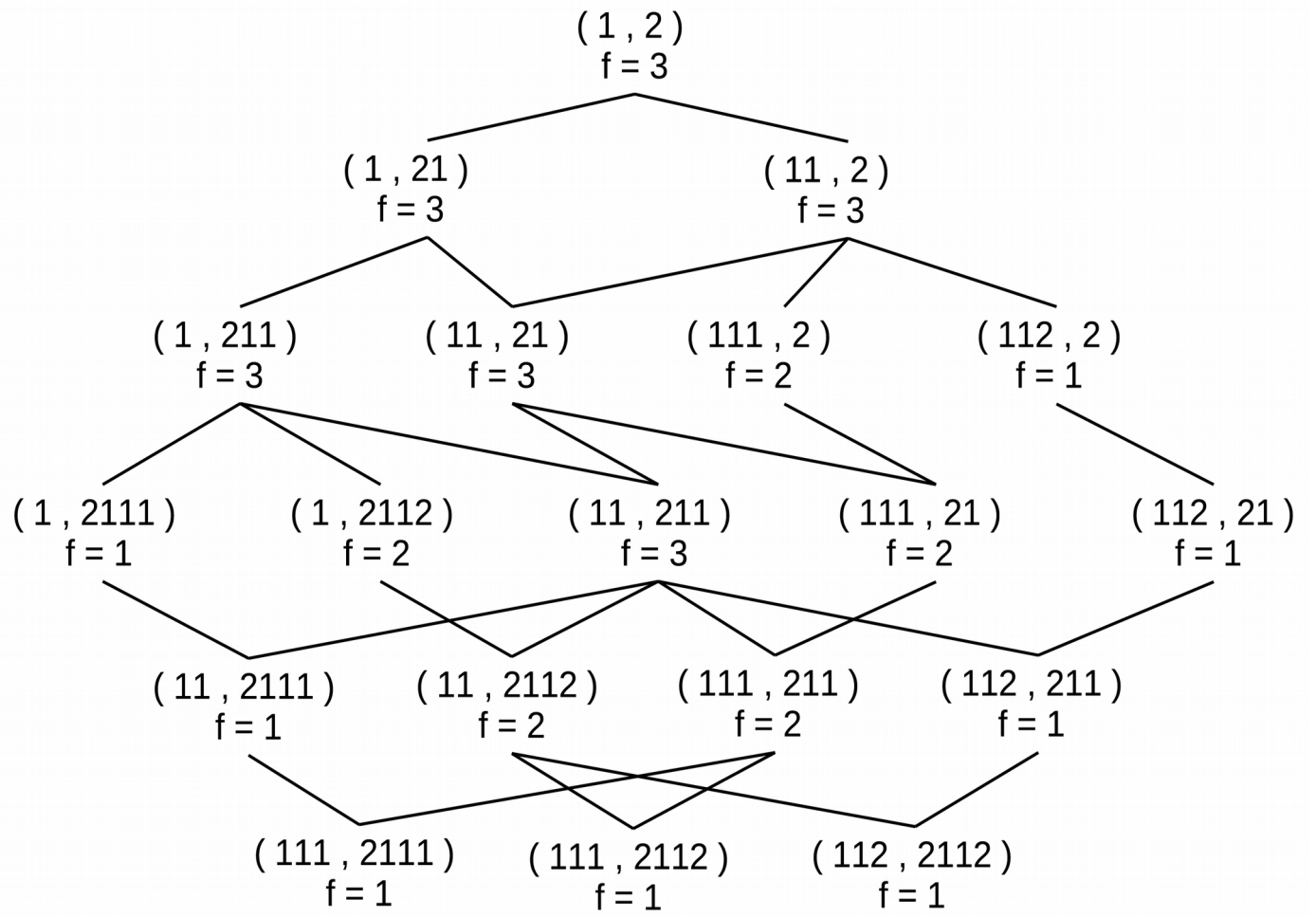
Lower-level vector `(SalesPerson.Alice, ACME, Kyushu.Fukuoka)`



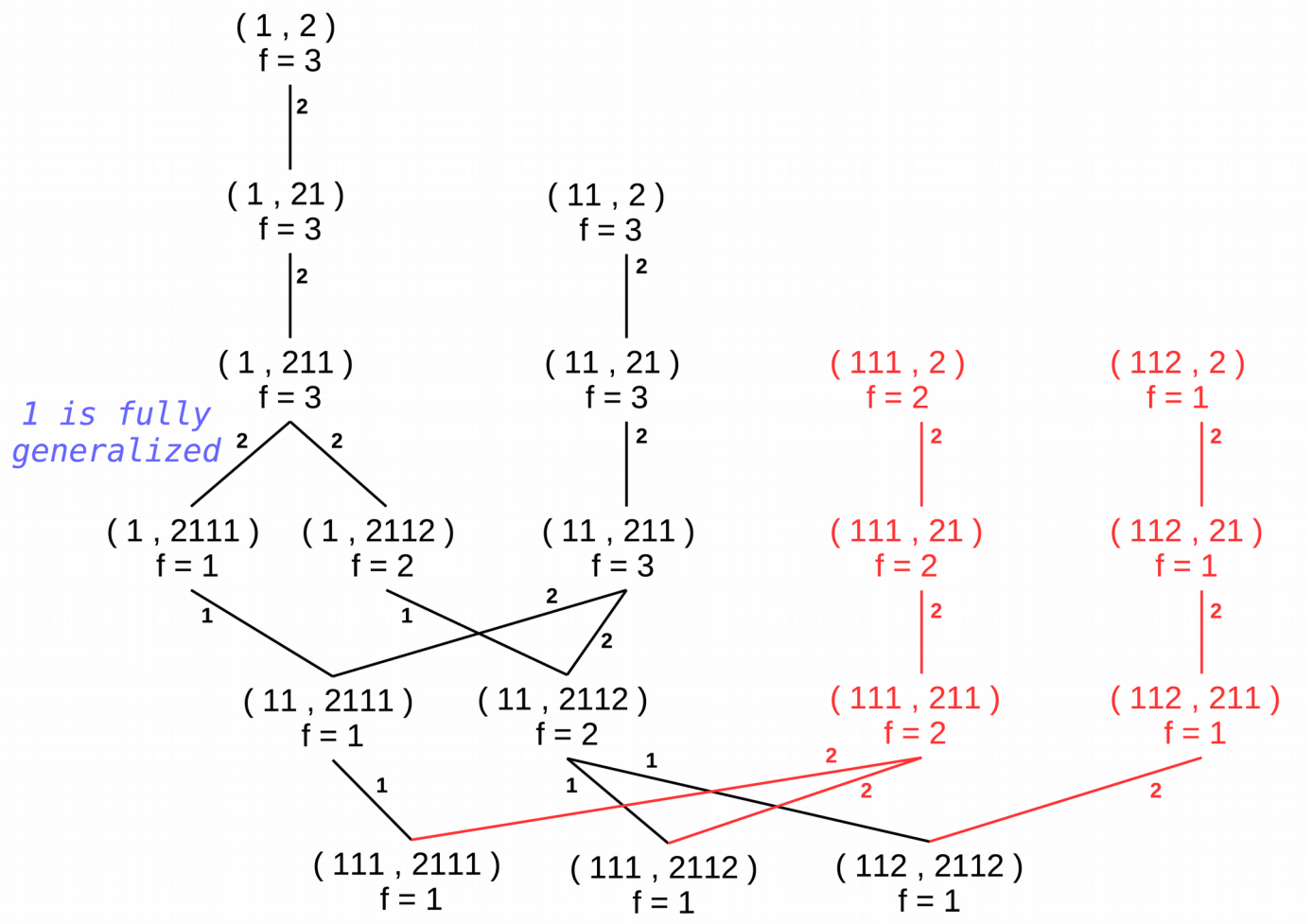
Method comparison

- Substitution method:
  - Additional subgraph isomorphism resolutions for every inserted edge
- Decomposition method:
  - Isomorphism resolutions only on top-levels
  - Frequent specializations by vector mining

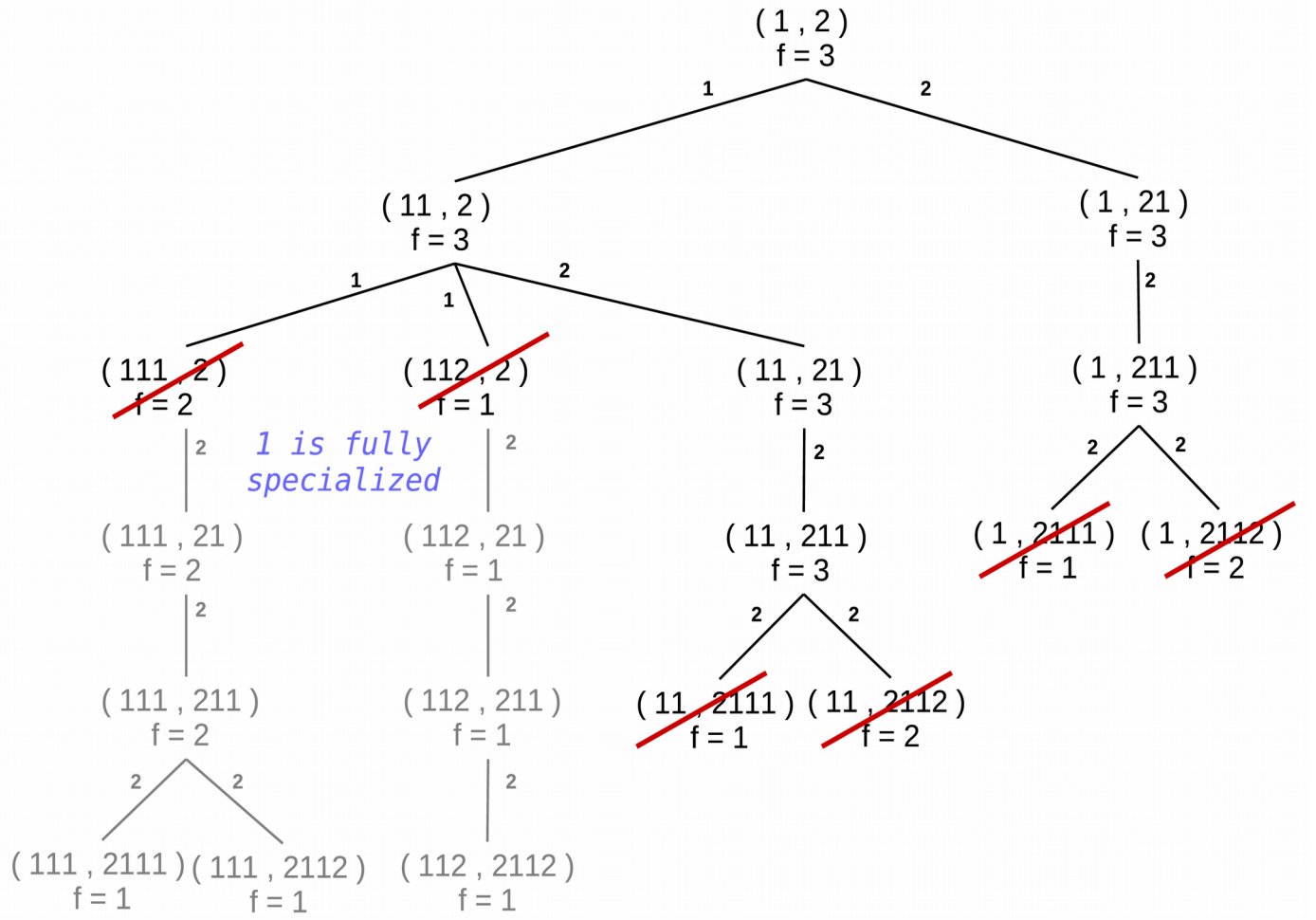
Generalized Vector Mining (GVM)



GVM: Bottom-up search



GVM: Top-down search



Result ranking

- Potentially huge number of results
- Interesting patterns should be presented first
- Significant patterns are more interesting
- Order results by p-value
- Fast analytical method [2]

[2] G. Micale, R. Giugno, A. Ferro, M. Mongioví, D. Shasha, and A. Pul virenti.
Fast analytical methods for finding signicant colored graph motifs.
To appear on Data Mining and Knowledge Discovery, 2017.

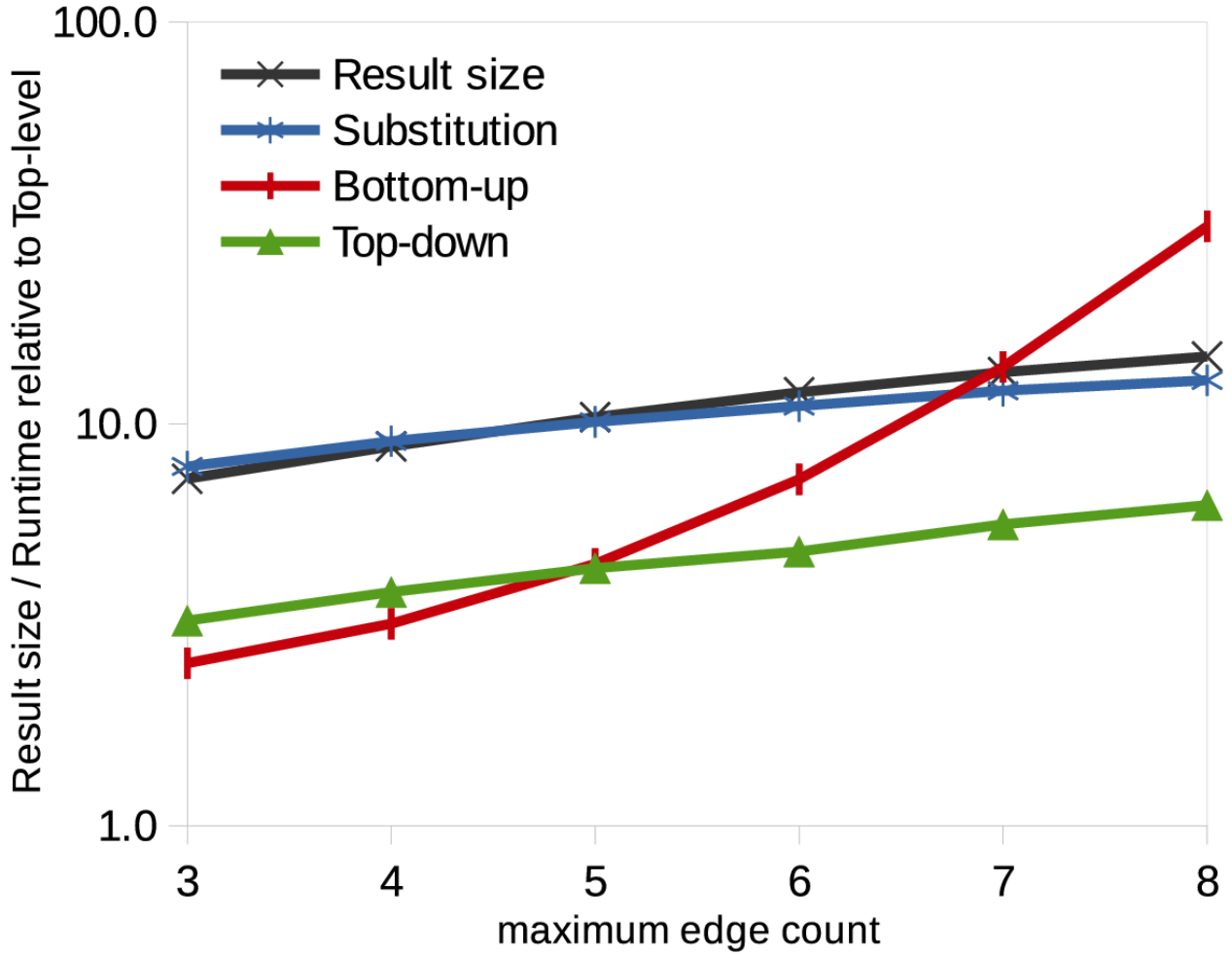
Experimental evaluation

- Graphs represent Business Process Executions
- Fully isomorphic on top-level
- $sup_{min} = 10\%$, $k_{max} = 8 \rightarrow$ 1M frequent patterns
- Generated FoodBroker [3] @ Gradoop [4]

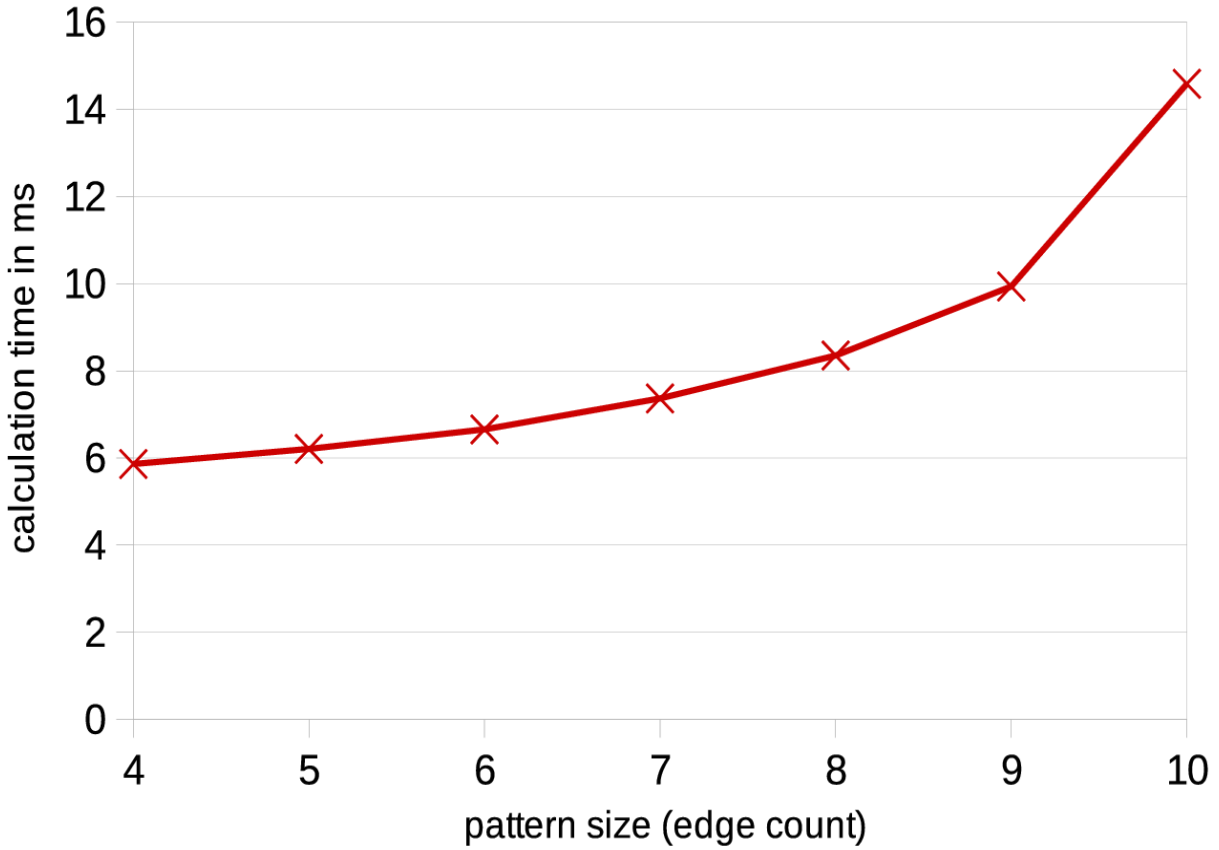
[3] A. Petermann, M. Junghanns, R. Müller, E. Rahm:
FoodBroker - Generating Synthetic Datasets for Graph-Based Business Analytics.
Int. Workshop on Big Data Benchmarking (WBDB) 2014: 145-155; Springer

[4] A. Petermann, M. Junghanns, S. Kemper, K. Gómez, N. Teichmann, E. Rahm:
Graph Mining for Complex Data Analytics.
Demo @ Int. Conf. on Data Mining (ICDM) 2016; ICDMW 2016: 1316-1319; IEEE

Experimental evaluation (mining)



Experimental evaluation (ranking)



THX

Questions?