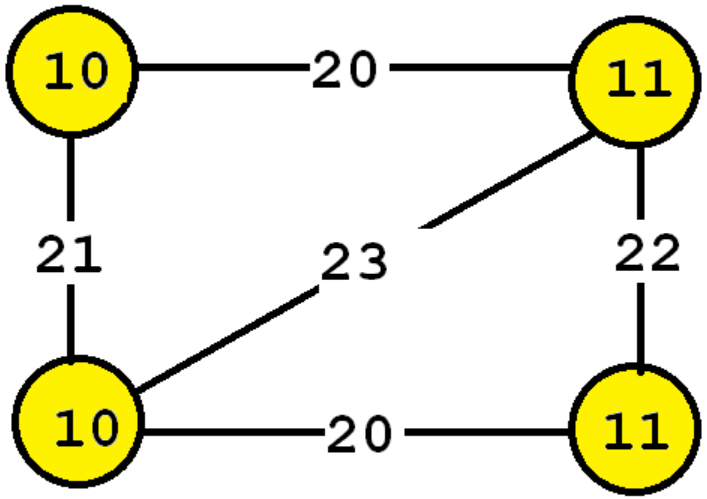
Frequent subgraph mining

Philippe Fournier-Viger
http://www.philippe-Fournier-viger.com

Source code and datasets available in the SPMF library

Frequent subgraph mining

- A **graph** (图表) is a set of vertices (顶点) and edges (边)
- **e.g.**



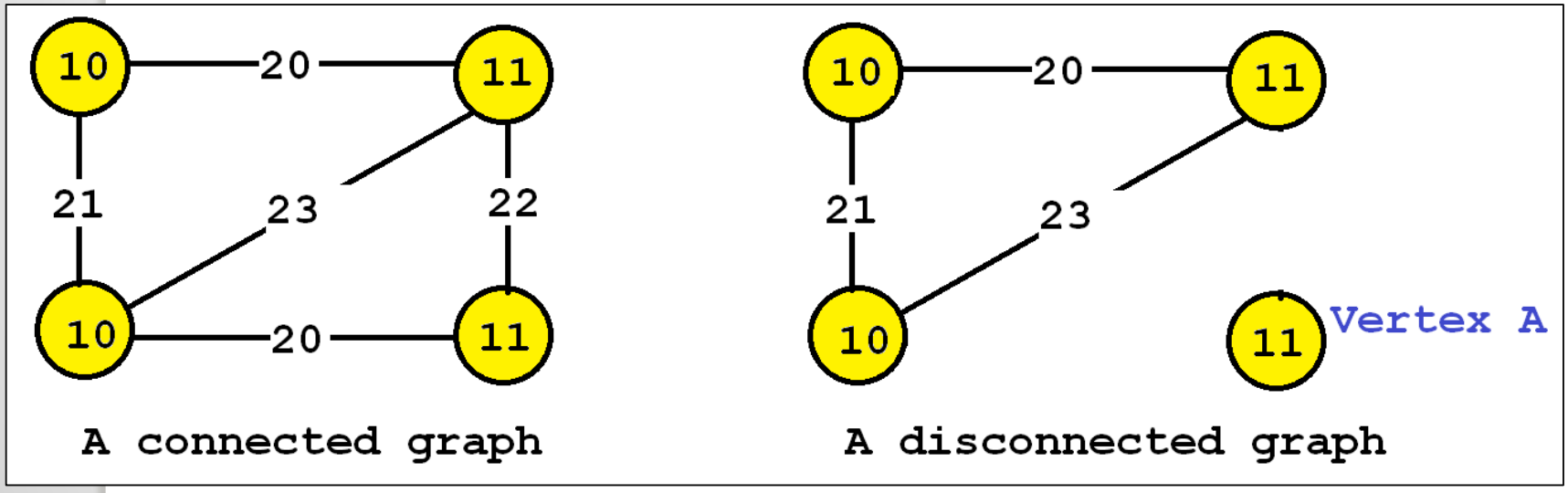
This graph has four **vertices** (in yellow color).
Each vertices has a label (10 or 11) that may not be unique.

This graph has five **edges** (black lines)
Each edge has a label (20,21,22, 23) that may not be unique.

Types of graphs

connected graph: by following the edges, it is possible to go from any vertex to any other vertices

disconnected graph: a graph that is not a connected graph

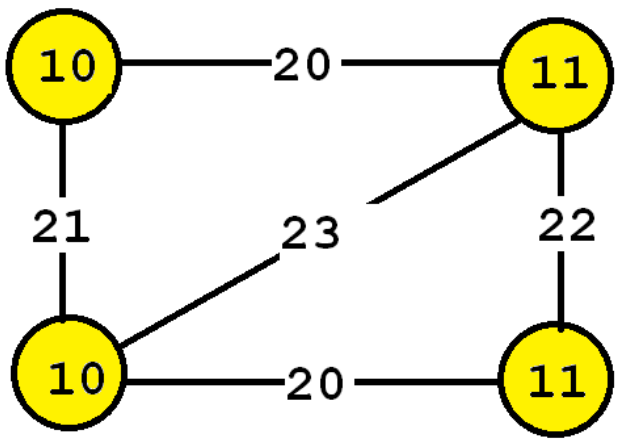


e.g. a graph where it is possible to go from any city to any other cities by following the roads.

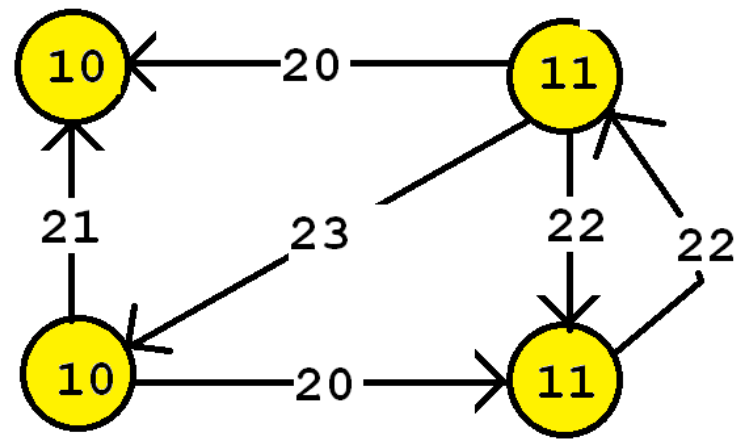
This graph is disconnected because Vertex A cannot be reached from the other vertices by following the edges

Types of graphs

undirected graph (无向图): edges are bidirectional
directed graph (有向图): edges are unidirectional



An undirected graph



A directed graph

A real-life example:

- **graphs where vertices are cities and edges are road**
- **some roads are « one-way » while others are bidirectional**

Analyzing graphs

- Many data mining tasks on graphs:
  - detecting communities, predicting friendship links, detecting influence between users, etc.
  - what is our goal?
  - what kind of data ? single graph? multiple graphs? directed graphs? etc.

- **Frequent subgraph mining**:
  - discover **interesting subgraph(s)** appearing often in a set of graphs (a graph database)

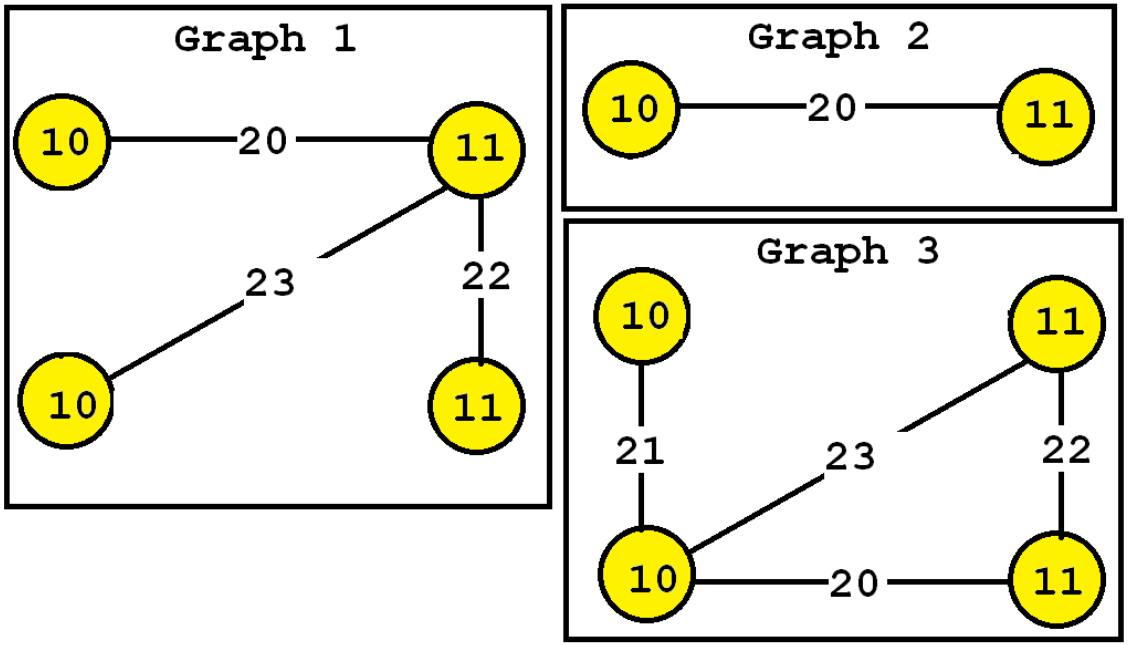
Frequent subgraph mining

Input:

- a **graph database** (a set of graphs)
- a **minimum support threshold** (***minsup***).

Example:

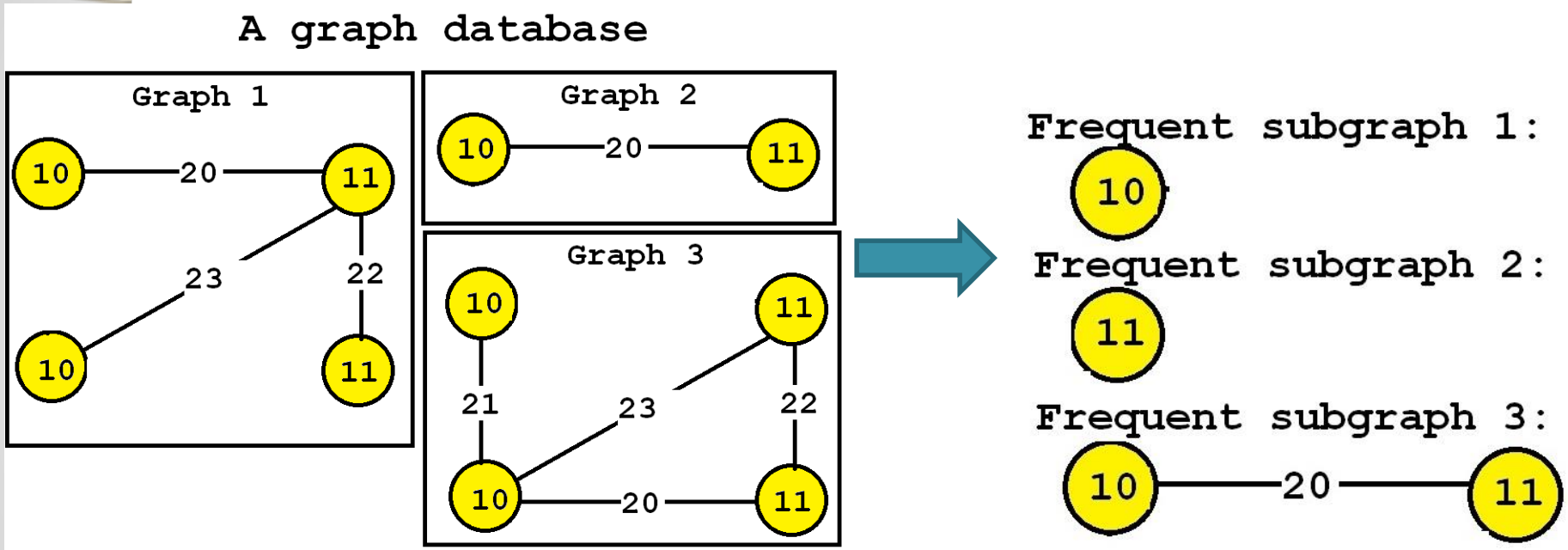
A graph database



minsup = 3

Output:

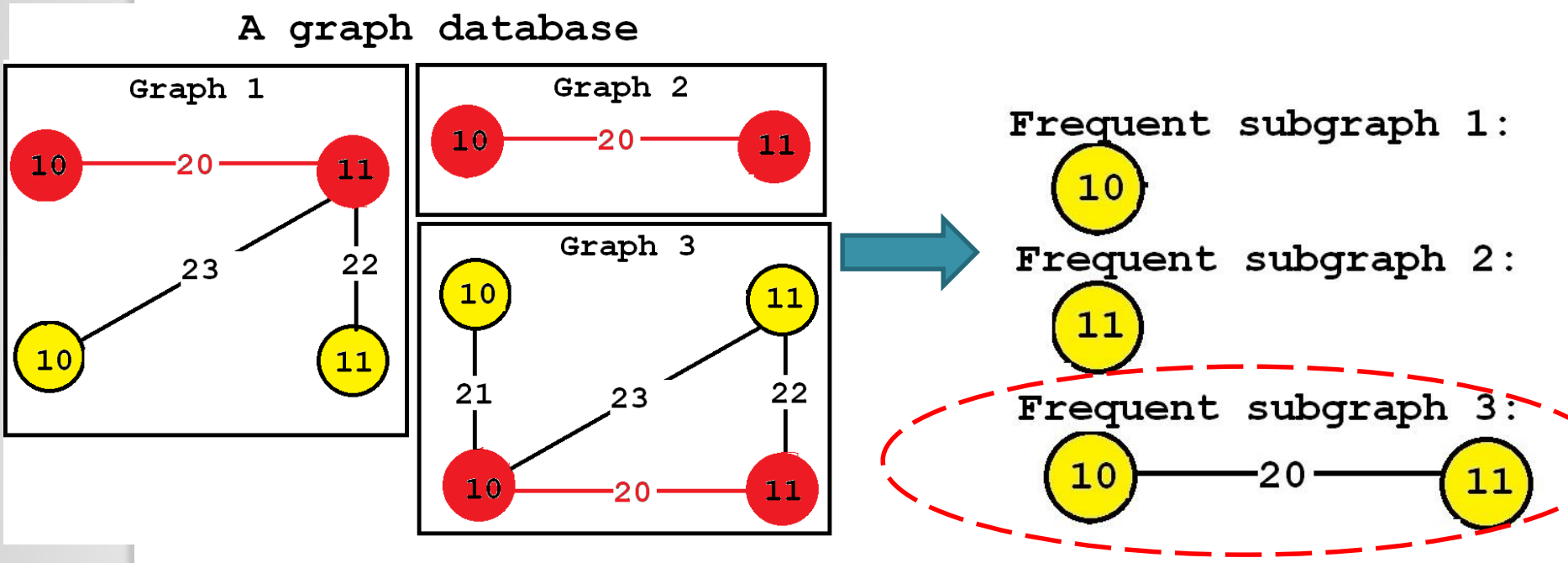
all subgraphs appearing in a least *minsup* graphs.



***minsup* = 3**

Output:

all subgraphs appearing in a least *minsup* graphs.

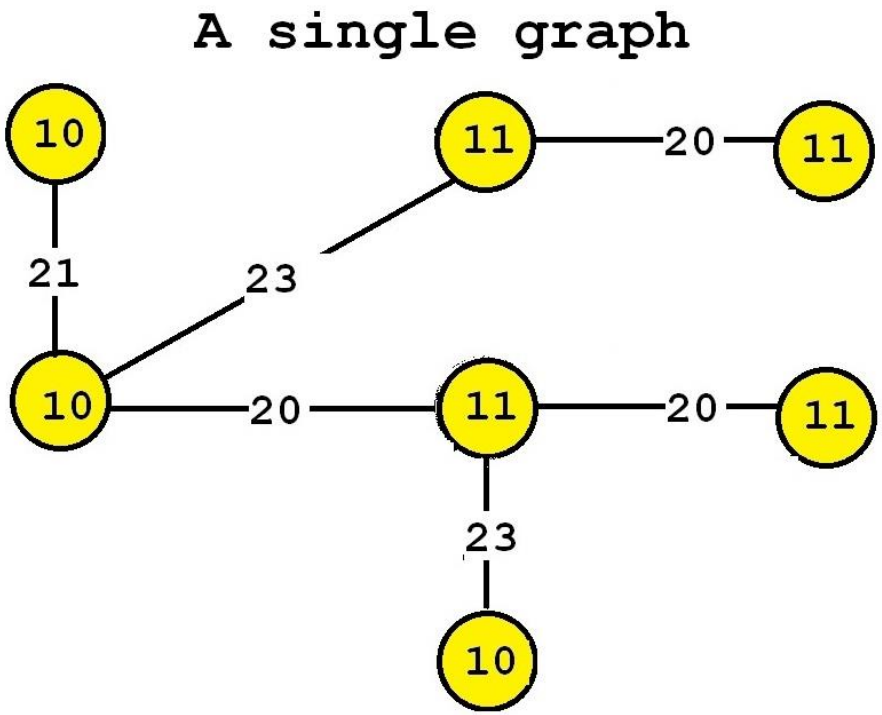


***minsup* = 3**

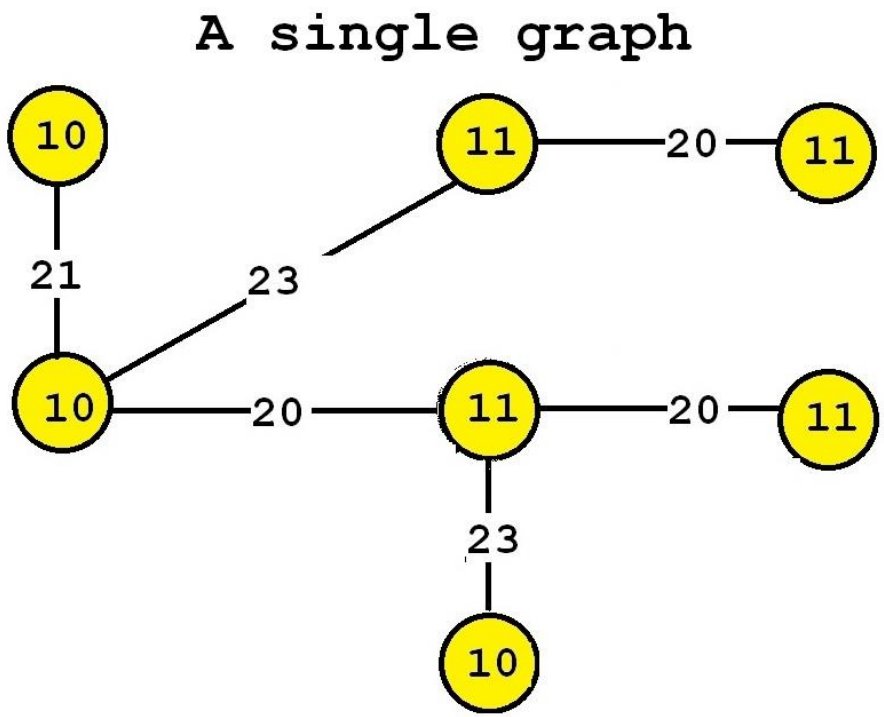
This subgraph has a support of 3

Frequent subgraph mining with a single graph

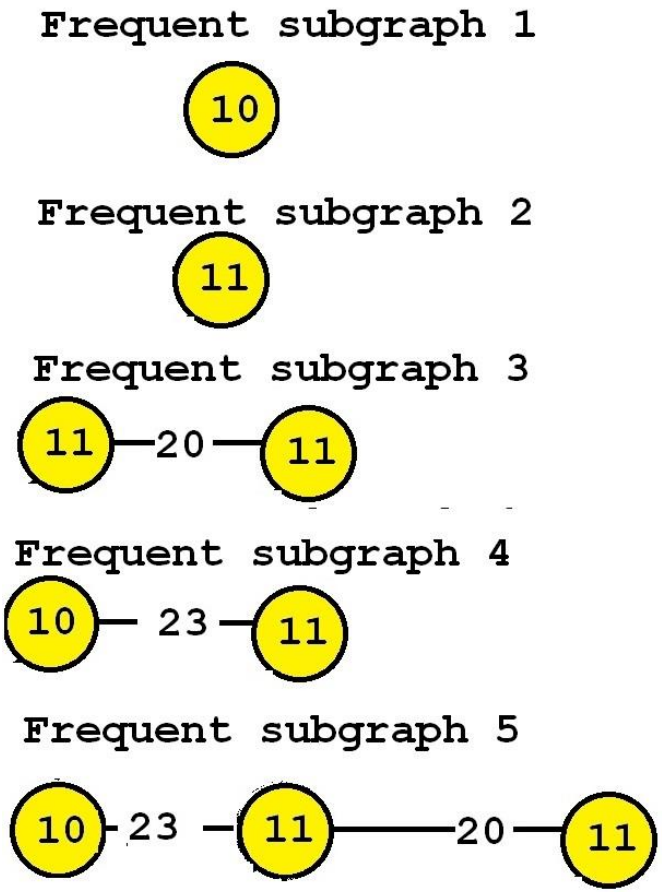
- A variation of the previous problem.
- We want to find **frequent subgraphs** in a **single large graph**.
- The **support of a subgraph** is the number of times that it appears in the single input graph



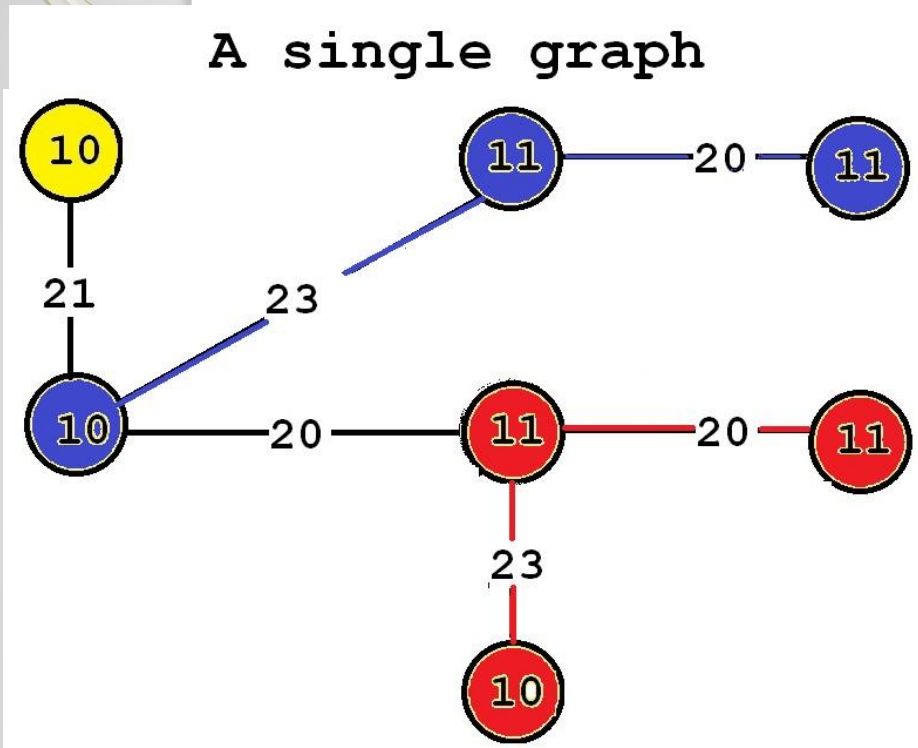
Frequent subgraph mining with a single graph



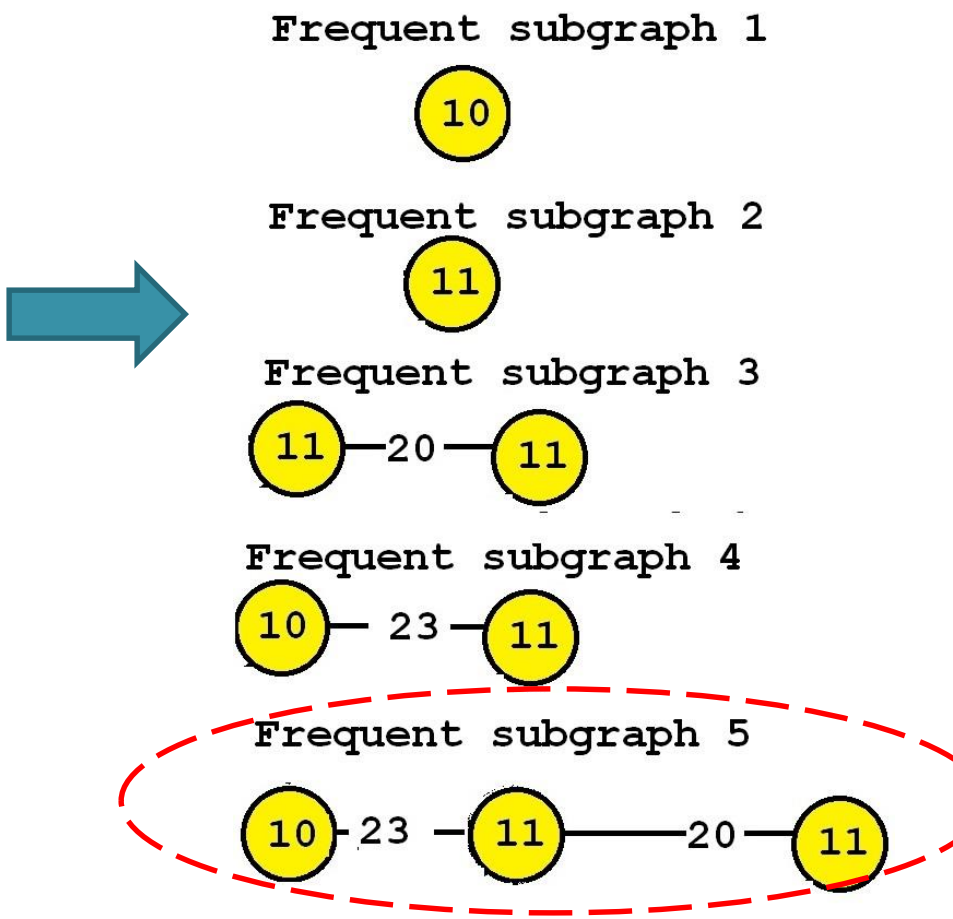
minsup = 2



Frequent subgraph mining with a single graph



minsup = 2



This subgraph has a support of 2

Algorithms for subgraph mining

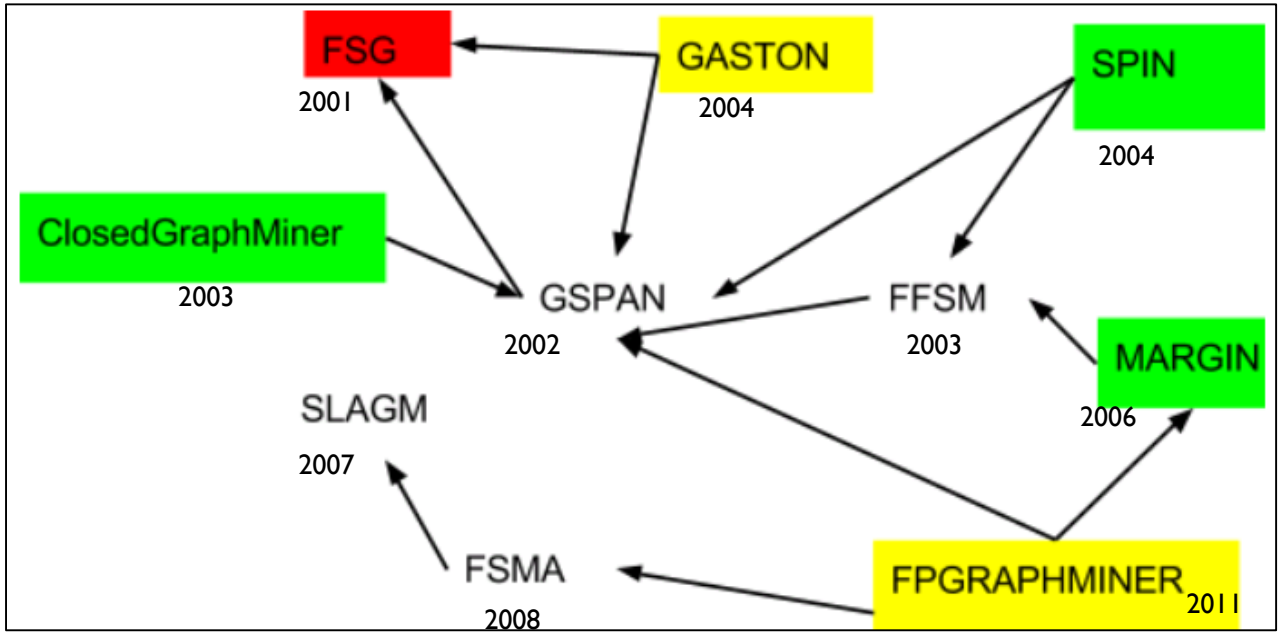
- Several algorithms:
  - FFSM, GSPAN, Gaston, etc.
- The same algorithm can usually be applied on a single graph or multiple graphs.

- Other variations:
  - finding frequent paths
  - finding frequent trees
  - finding closed/maximal subgraphs…
  - …

Performance comparison

Authors of data mining papers often do not compare their algorithms with the best ones published until now.

Frequent subgraph mining (before 2014)



Legend: arrow X → Y from an algorithm X to an algorithm Y indicates that X was shown to be a better algorithm than Y in terms of execution time by the authors of X in an aexperiment.