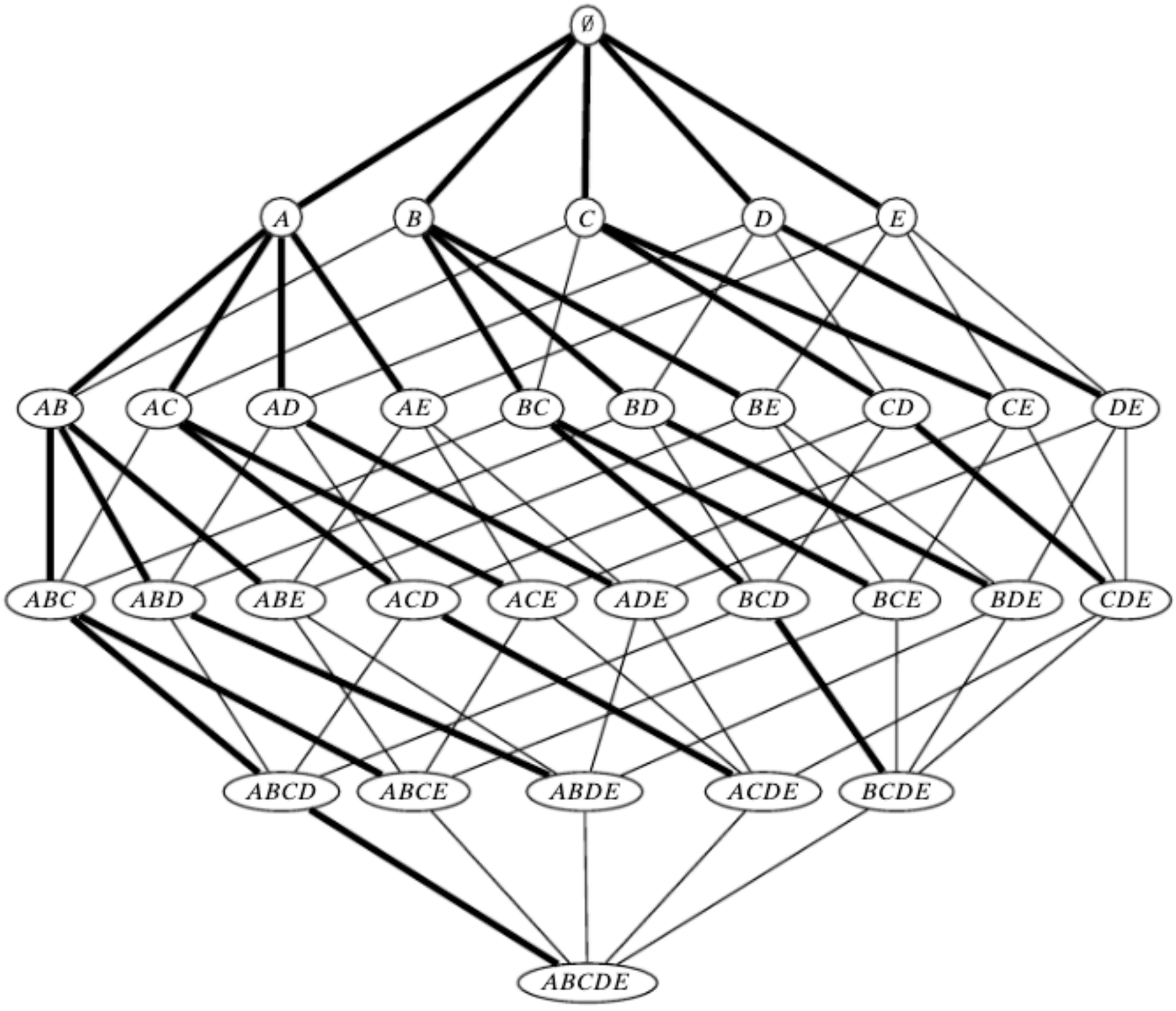
Frequent pattern mining

Chapters 4.1, 4.2, 4.3, 4.4.1, 4.4.2 (excluding 4.4.2.1), 4.4.3 (until p.110 excluding p.109) of C. Aggarwal, *Data Mining - The Textbook*

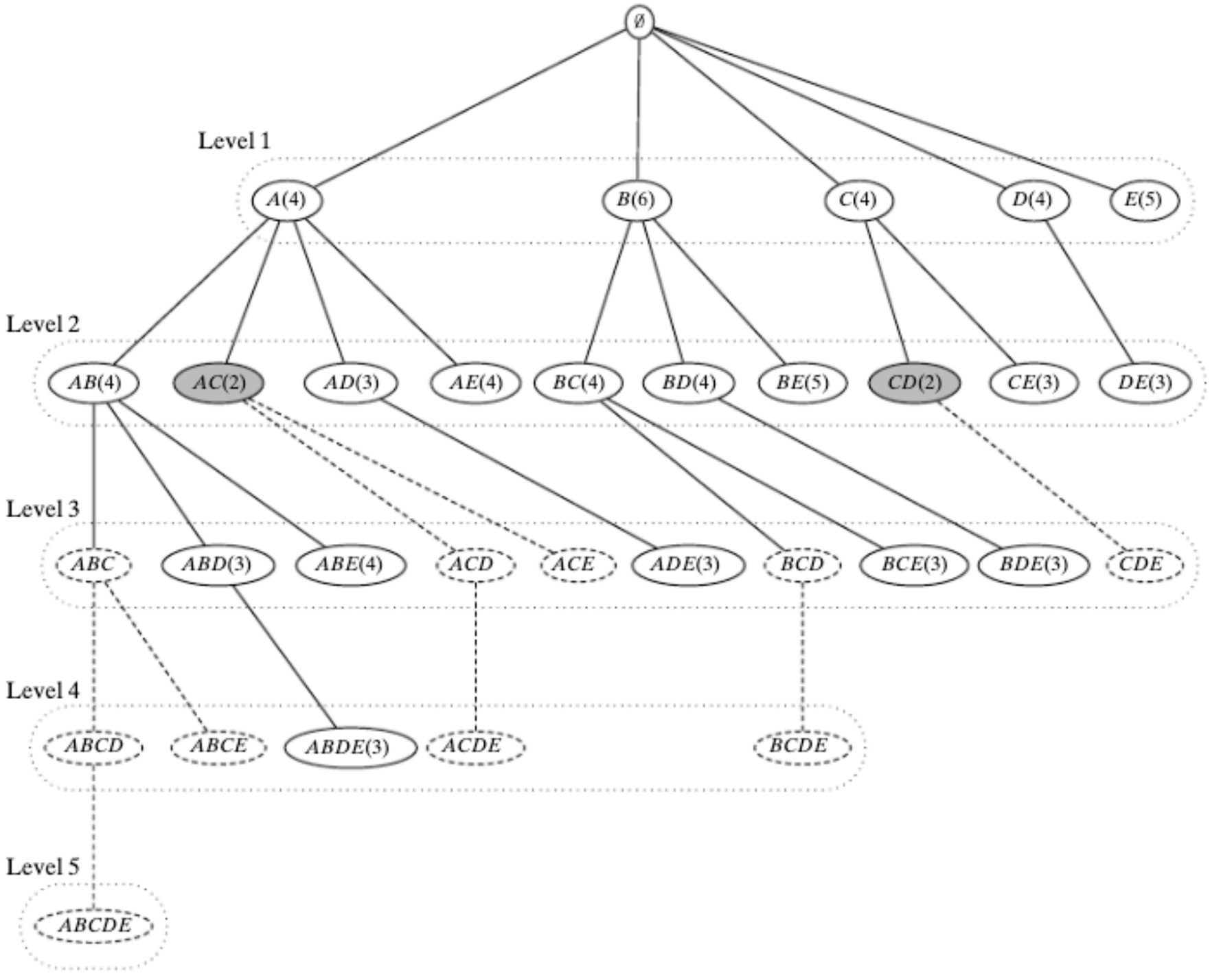
Giulia Bernardini
giulia.bernardini@units.it

Algorithmic Data Mining
a.y. 2023/2024

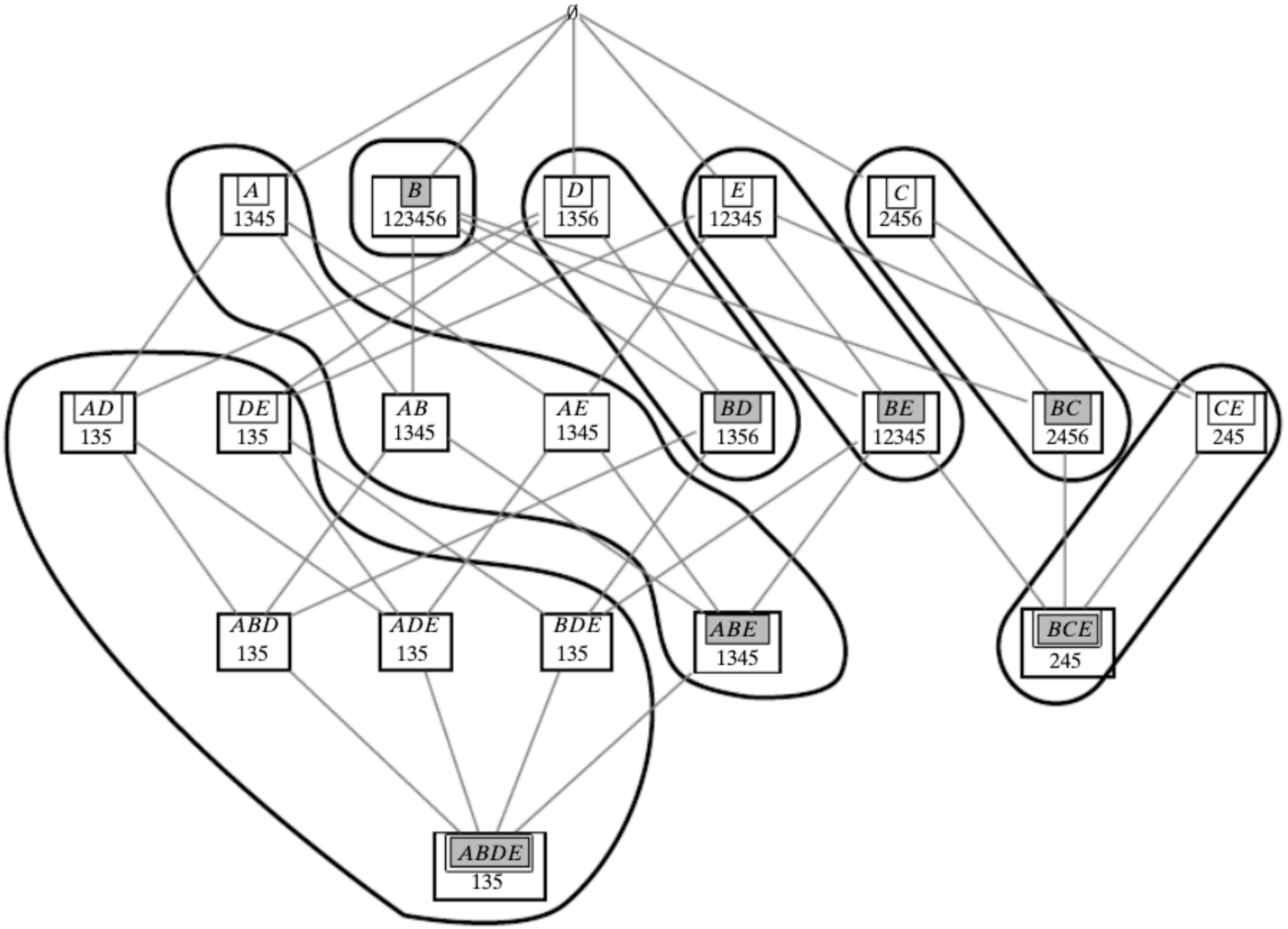
1	*ABDE*
2	*BCE*
3	*ABDE*
4	*ABCE*
5	*ABCDE*
6	*BCD*



Itemset lattice for the transaction database on the left



Enumeration tree with minsup=3. Dashed nodes are pruned itemsets. Gray nodes are infrequent itemsets.



Lattice for the same dataset (different ordering), showing the support equivalent classes and the representative closed patterns (gray). Maximal patterns are double boxed.

Graph pattern mining

Chapter 11 of Mohammed J. Zaki, Wagner Meira, Jr., *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*

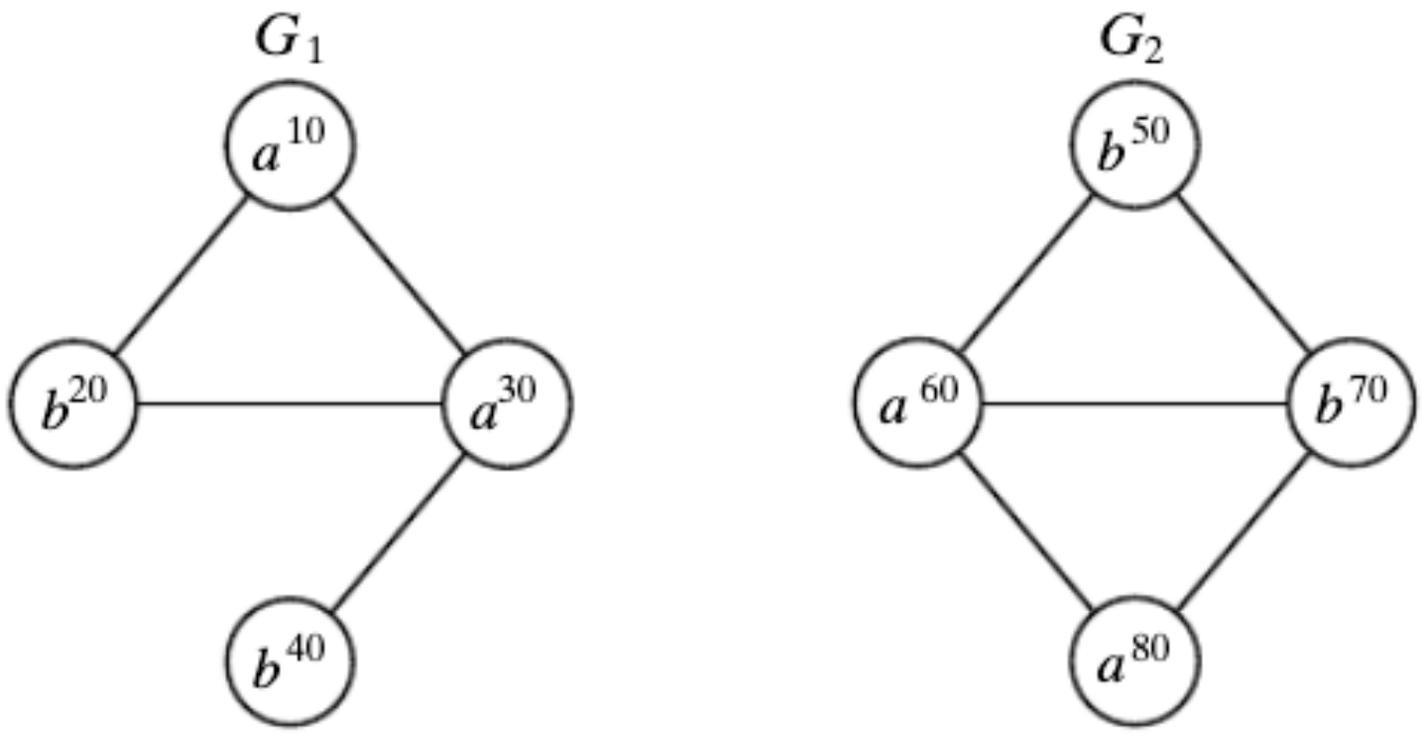
Giulia Bernardini
giulia.bernardini@units.it

Algorithmic Data Mining
a.y. 2023/2024

Algorithm 11.2: Rightmost Path Extensions and Their Support

RIGHTMOSTPATH-EXTENSIONS $(C, \mathbf{D})$:

$R \leftarrow$ nodes on the rightmost path in C
$u_r \leftarrow$ rightmost child in C `// dfs number`
$\mathcal{E} \leftarrow \emptyset$ `// set of extensions from` C
foreach $G_i \in \mathbf{D}$, $i = 1, \ldots, n$ **do**
  if $C = \emptyset$ **then**
    `// add distinct label tuples in` G_i `as forward extensions`
    foreach *distinct* $\langle L(x), L(y), L(x,y) \rangle \in G_i$ **do**
      $f = \langle 0, 1, L(x), L(y), L(x,y) \rangle$
      Add tuple f to $\mathcal{E}$ along with graph id i
  else
    $\Phi_i = \text{SUBGRAPHISOMORPHISMS}(C, G_i)$
    foreach *isomorphism* $\phi \in \Phi_i$ **do**
      `// backward extensions from rightmost child`
      foreach $x \in N_{G_i}(\phi(u_r))$ *such that* $\exists v \leftarrow \phi^{-1}(x)$ **do**
        if $v \in R$ *and* $(u_r, v) \notin G(C)$ **then**
          $b = \langle u_r, v, L(\phi(u_r)), L(\phi(v)), L(\phi(u_r), \phi(v)) \rangle$
          Add tuple b to $\mathcal{E}$ along with graph id i
      `// forward extensions from nodes on rightmost path`
      foreach $u \in R$ **do**
        foreach $x \in N_{G_i}(\phi(u))$ *and* $\nexists \phi^{-1}(x)$ **do**
          $f = \langle u, u_r + 1, L(\phi(u)), L(x), L(\phi(u), x) \rangle$
          Add tuple f to $\mathcal{E}$ along with graph id i
`// Compute the support of each extension`
foreach *distinct extension* $s \in \mathcal{E}$ **do**
  $sup(s) =$ number of distinct graph ids that support tuple s
return *set of pairs* $\langle s, sup(s) \rangle$ *for extensions* $s \in \mathcal{E}$, *in tuple sorted order*

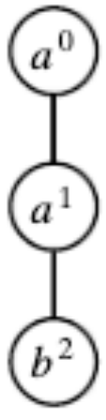


Example graph database

C

$t_1 : \langle 0, 1, a, a \rangle$
$t_2 : \langle 1, 2, a, b \rangle$

$G(C)$



(a) Code C and graph $G(C)$

Id	ϕ	Extensions
G_1	ϕ_1	$\{\langle 2,0,b,a\rangle, \langle 1,3,a,b\rangle\}$
	ϕ_2	$\{\langle 1,3,a,b\rangle, \langle 0,3,a,b\rangle\}$
	ϕ_3	$\{\langle 2,0,b,a\rangle, \langle 0,3,a,b\rangle\}$
G_2	ϕ_4	$\{\langle 2,0,b,a\rangle, \langle 2,3,b,b\rangle, \langle 0,3,a,b\rangle\}$
	ϕ_5	$\{\langle 2,3,b,b\rangle, \langle 1,3,a,b\rangle, \langle 0,3,a,b\rangle\}$
	ϕ_6	$\{\langle 2,0,b,a\rangle, \langle 2,3,b,b\rangle, \langle 1,3,a,b\rangle\}$

(c) Edge extensions

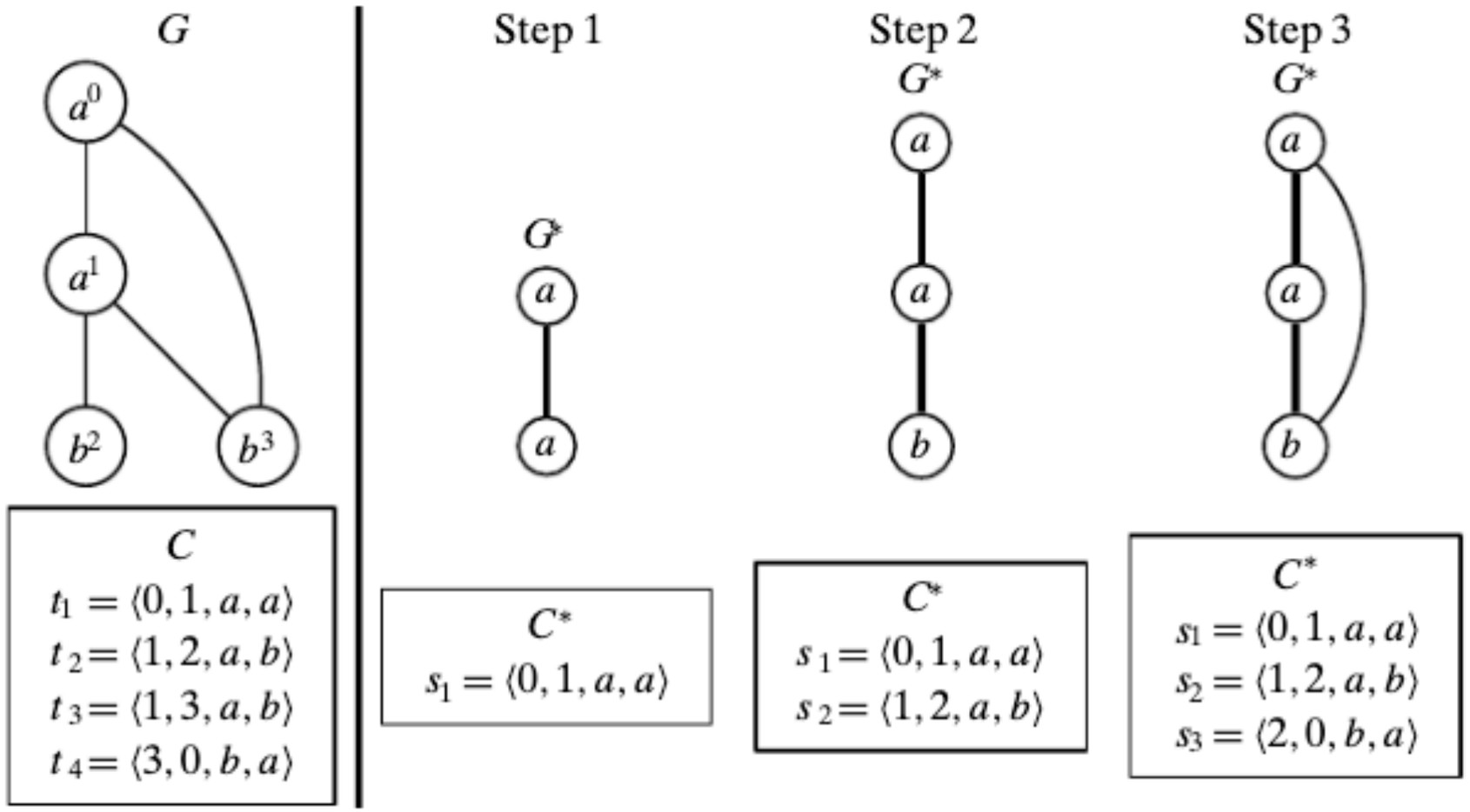
Φ	ϕ	0	1	2
Φ_1	ϕ_1	10	30	20
	ϕ_2	10	30	40
	ϕ_3	30	10	20
Φ_2	ϕ_4	60	80	70
	ϕ_5	80	60	50
	ϕ_6	80	60	70

(b) Subgraph isomorphisms

Extension	Support
$\langle 2,0,b,a\rangle$	2
$\langle 2,3,b,b\rangle$	1
$\langle 1,3,a,b\rangle$	2
$\langle 0,3,a,b\rangle$	2

(d) Extensions (sorted) and supports

Rightmost path extensions for the database in the previous slide.



Canonicality check with result FALSE.