LISN
LABORATOIRE INTERDISCIPLINAIRE
DES SCIENCES DU NUMÉRIQUE

université
PARIS-SACLAY

cnrs

Artificial Intelligence

Cours7 - Frequent Itemset Mining

L3 - Informatique

Nadjib Lazaar
Ing - Phd - HDR - Professor - Paris-Saclay University - LISN - LaHDAK
lazaar@lisn.fr https://perso.lisn.upsaclay.fr/lazaar/
14/03/2025

Introduction à la fouille de données

Définitions

- **Qu'est-ce que la découverte de connaissances (KDD) ?**
  - La découverte de connaissances dans les bases de données tourne autour de l'investigation et de la création de connaissances, des processus, des algorithmes et des mécanismes permettant d'extraire des connaissances potentielles à partir des collections de données.

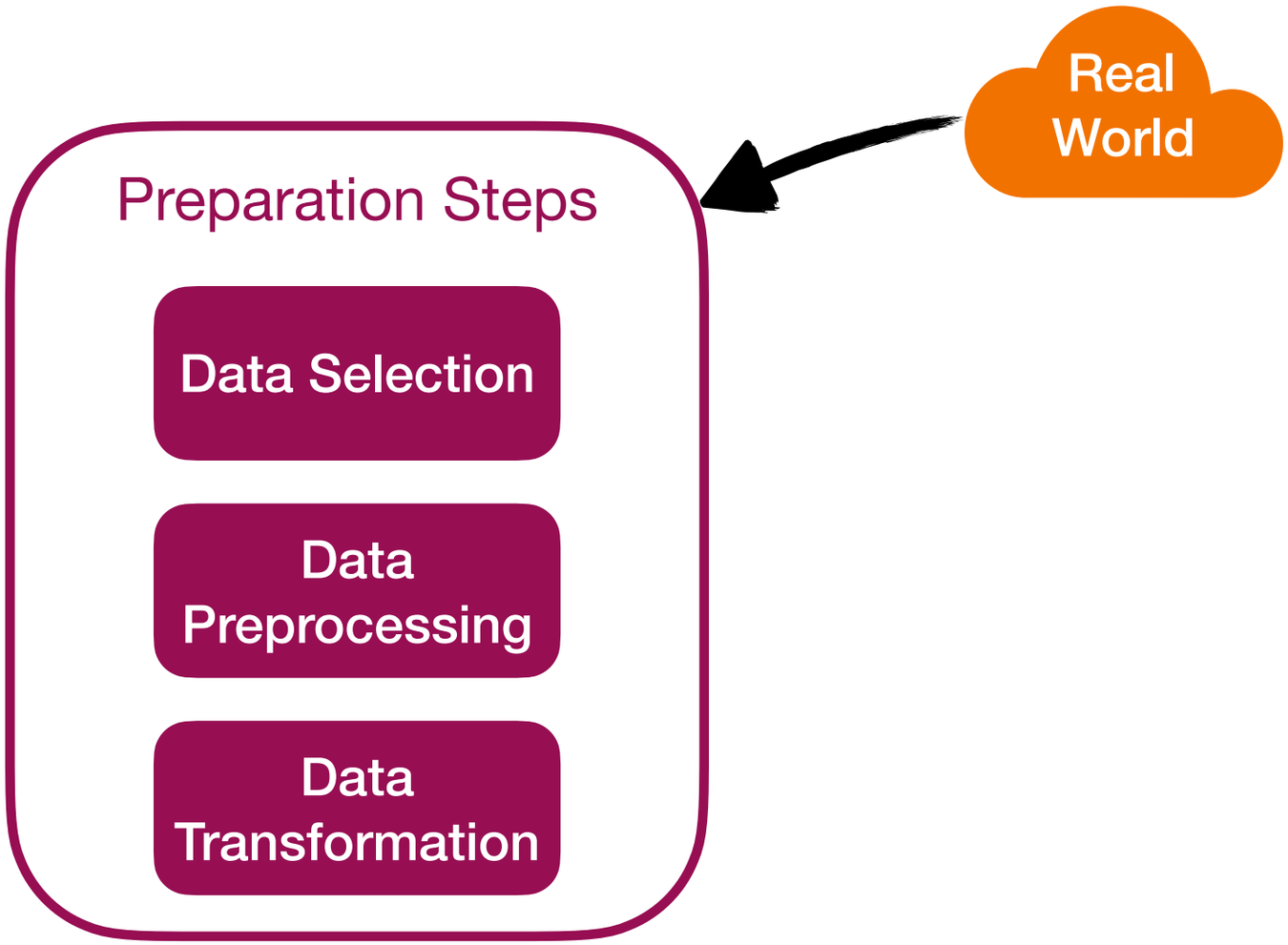
- **Qu'est-ce que la fouille de données (Data Mining) ?**
  - Processus d'extraction de modèles utiles et de connaissances à partir de grandes quantités de données.
  - Applications : segmentation de marché, détection de fraude, analyse des comportements des consommateurs, etc.

Introduction à la fouille de données

Définitions

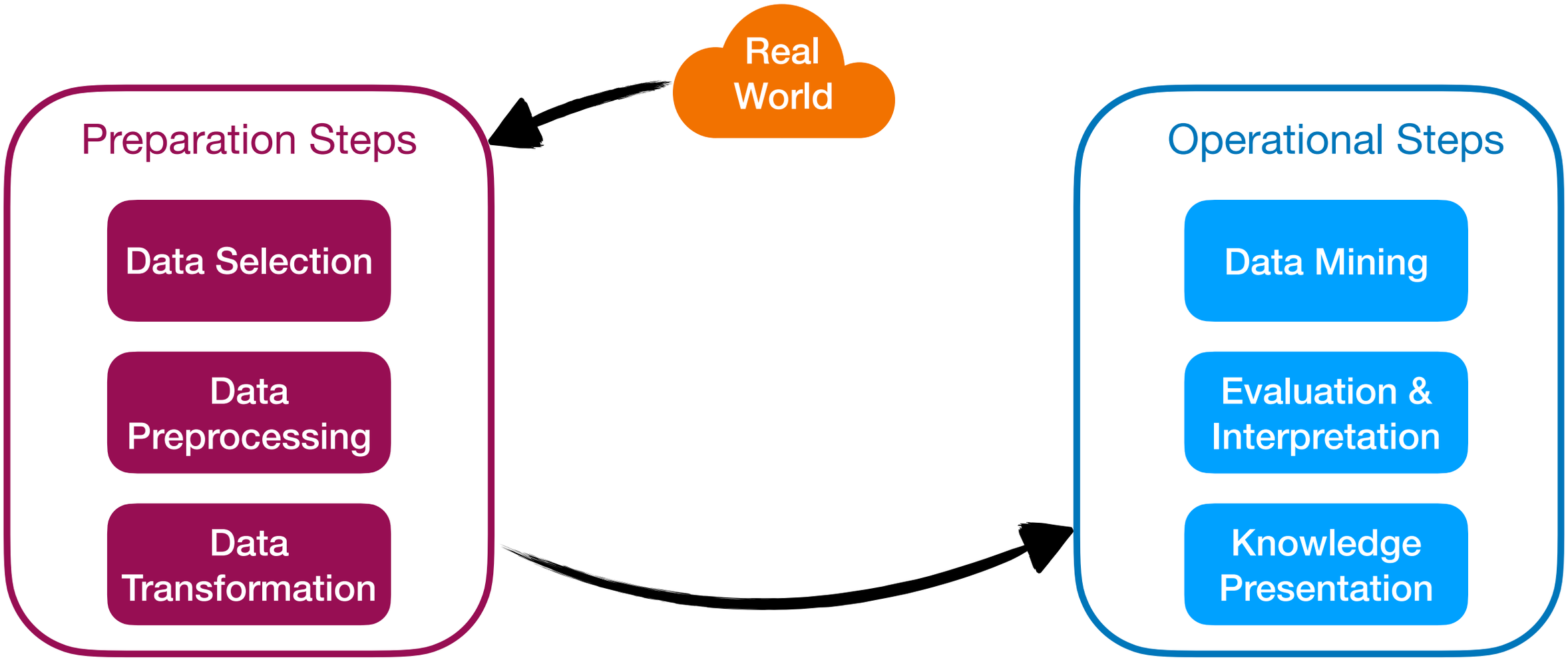
Introduction à la fouille de données

Définitions



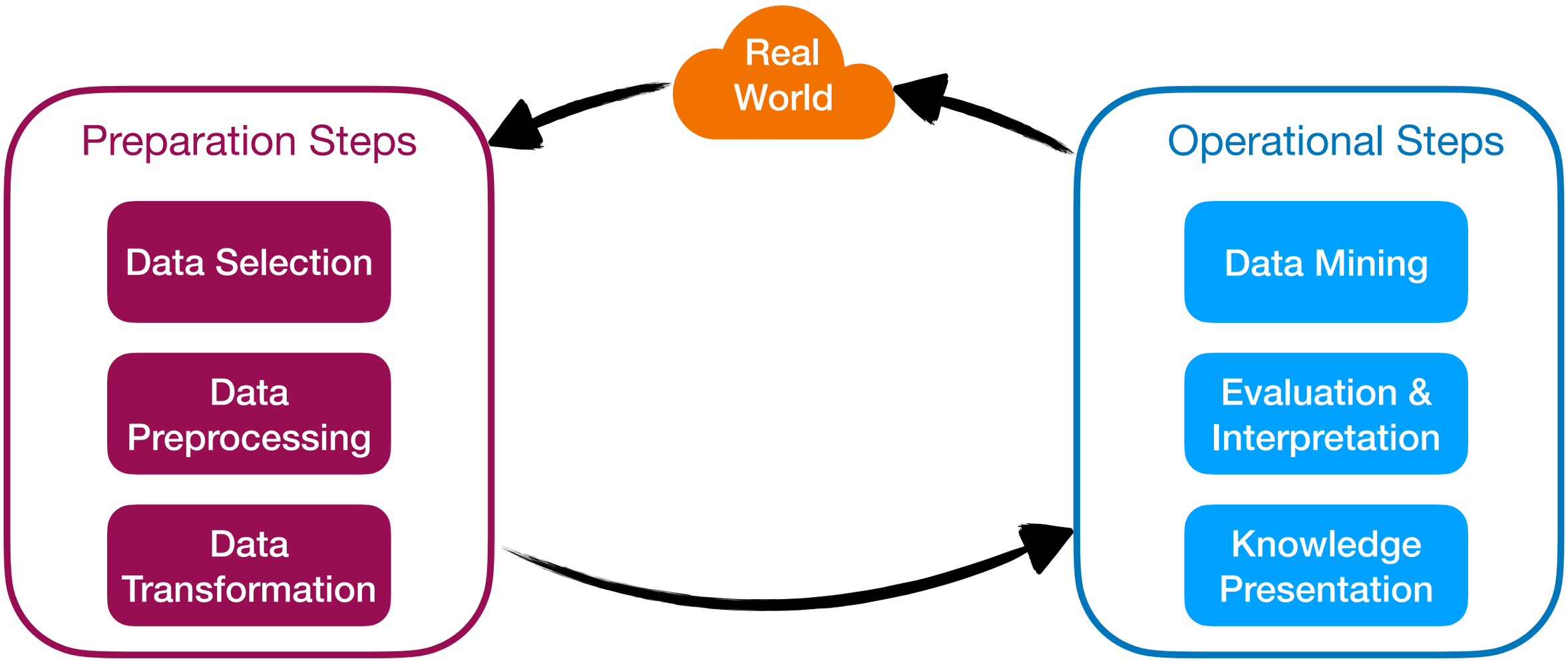
Introduction à la fouille de données

Définitions



Introduction à la fouille de données

Définitions



Types de fouille et leurs techniques

Fouille descriptive : Résumer et explorer les caractéristiques des données

- **Exemple :**
  - **Analyse des ventes dans un magasin** : Identifier les produits les plus populaires sur une période donnée (par exemple, top 10 des produits les plus vendus)
  - **Analyse démographique des utilisateurs** : Explorer des données pour comprendre les caractéristiques des utilisateurs d'une plateforme, comme l'âge, le sexe et la localisation.
- **Techniques utilisées :**
  - **Statistiques descriptives :** Moyenne, médiane, variance, écart-type, etc
  - **Analyse en composantes principales (PCA) :** Réduction de dimensionnalité pour identifier les principales variables qui expliquent les données
  - **Analyse de clusters (Clustering) :** Groupement des données similaires en clusters. Par exemple, le k-means ou DBSCAN.
  - **Visualisation des données :** Utilisation de graphiques et de diagrammes (diagrammes de dispersion, histogrammes, etc.) pour examiner des relations et des distributions de données.
  - **Histogrammes et Boxplots :** Pour explorer la distribution des données.

Types de fouille et leurs techniques

Fouille prédictive : Utiliser les données pour prédire des résultats futurs

- **Exemple :**
  - **Prévision des ventes** : Utiliser l'historique des ventes pour prédire les ventes futures pendant la période de fêtes
  - **Prévision du churn (attrition des clients)** : Prévoir les clients susceptibles de quitter un service ou une plateforme en fonction de leur comportement passé (ex. : abonnements à un service de streaming)
- **Techniques utilisées :**
  - **Régression : Régression linéaire** : Prédire une variable continue en fonction d'autres variables (par exemple, prédire les ventes en fonction de la publicité); **Régression logistique** : Utilisée pour les variables catégorielles (ex. : prédire si un client va acheter ou non un produit)
  - **Arbres de décision** : Modélisation des décisions sous forme d'un arbre binaire pour prédire des résultats en fonction de conditions
  - **Forêts aléatoires (Random Forest)** : Ensemble d'arbres de décision pour améliorer la précision de la prédiction
  - **Réseaux de neurones** : Utilisés pour des prédictions complexes, par exemple pour la classification ou la régression dans des ensembles de données volumineux
  - **SVM (Support Vector Machines)** : Pour classer des données dans des catégories prédictives
  - **K plus proches voisins (K-NN)** : Prédire une catégorie ou une valeur en fonction des voisins les plus proches dans l'espace des caractéristiques.

Types de fouille et leurs techniques

Fouille de motifs : Identifier des motifs récurrents dans un ensemble de données

- **Exemple :**
  - **Analyse des paniers d'achat (règles d'association)** : Identifier des motifs d'achats fréquents, comme "Les clients qui achètent des ordinateurs portables achètent souvent aussi une souris et un sac."
  - **Fouille de motifs ensemblistes** : Trouver des groupes d'articles achetés ensemble fréquemment, comme un ensemble de produits (par exemple, un kit de fitness incluant des haltères, des bandes de résistance et un tapis de yoga)
- **Techniques utilisées :**
  - **Motifs Ensemblistes** : Algorithme Apriori, Algorithme FP-growth, Eclat, LCM, ect.
  - **Fouille de motifs séquentiels :** Algorithme SPADE, Algorithme GSP (Generalized Sequential Pattern).

Applications pratiques de la fouille de motifs

Bioinformatique et analyse génomique

- **Identification de motifs dans les séquences ADN et protéines**
  - Détection de séquences fréquentes dans l'ADN qui peuvent être associées à certaines maladies génétiques
  - **Exemple :** Identification de mutations spécifiques dans des génomes de patients atteints de maladies rares
- **Analyse des réseaux d'expression génique**
  - Identification de groupes de gènes co-exprimés qui pourraient être impliqués dans le même processus biologique
  - Utilisation de techniques de fouille de motifs pour prédire les interactions entre protéines.

Applications pratiques de la fouille de motifs

Santé et médecine prédictive

- **Analyse des dossiers médicaux et des traitements**
  - Identification des combinaisons de médicaments fréquemment prescrites ensemble
  - Analyse des données hospitalières pour détecter des tendances dans l'apparition de maladies (exemple : liens entre certains facteurs environnementaux et les maladies chroniques)
- **Diagnostic assisté par IA**
  - Découverte de motifs récurrents dans les données d'imagerie médicale (IRM, radiographies) pour détecter des anomalies
  - **Exemple :** Utilisation de motifs ensemblistes pour identifier des caractéristiques communes aux patients atteints d'un cancer spécifique

Applications pratiques de la fouille de motifs

Détection de fraude et cybersécurité

- **Fraude bancaire et assurance**
  - Identification de transactions frauduleuses en détectant des modèles anormaux dans les données bancaires
  - **Exemple :** Un client effectuant des transactions inhabituelles dans plusieurs pays en peu de temps
- **Détection d'attaques en cybersécurité**
  - Analyse des journaux de connexion pour détecter des accès non autorisés basés sur des schémas inhabituels
  - **Exemple :** Identification de motifs de connexions suspectes à des heures irrégulières sur un réseau d'entreprise

Applications pratiques de la fouille de motifs

Commerce et marketing

- **Analyse du comportement des consommateurs**
  - Identification de groupes de produits souvent achetés ensemble pour optimiser les recommandations
  - **Exemple :** Un client qui achète un smartphone achète souvent une coque et un chargeur dans la même commande
- **Segmentation des clients**
  - Regroupement des clients en fonction de leurs habitudes d'achat pour personnaliser les offres marketing
  - **Exemple :** Détection de motifs de consommation pour proposer des promotions adaptées.

Applications pratiques de la fouille de motifs

Commerce et marketing

- **Analyse du comportement des consommateurs**
  - Identification de groupes de produits souvent achetés ensemble pour optimiser les recommandations
  - **Exemple :** Un client qui achète un smartphone achète souvent une coque et un chargeur dans la même commande
- **Segmentation des clients**
  - Regroupement des clients en fonction de leurs habitudes d'achat pour personnaliser les offres marketing
  - **Exemple :** Détection de motifs de consommation pour proposer des promotions adaptées.

Fouille de Motifs

Introduction

- **Qu’est-ce qu’un motif ?**
  - Un motif est une structure, une relation ou une règle récurrente qui apparaît fréquemment dans les données
- **Pourquoi fouiller des motifs ?**
  - Pour découvrir des tendances cachées
  - Pour comprendre les relations entre les variables
  - Pour générer des règles de décision

Fouille de Motifs Ensemblistes

Notions de base

- Items: $I = \{p_1, \ldots, p_n\}$
- Itemset, transaction: $P, T \subseteq I$
- Transactional dataset: $D = \{T_1, \ldots, T_m\}$
- Language of itemsets: $\mathscr{L}_I = 2^I$
- Cover of an itemset: $cover(P) = \{T_i \in D : \ P \subseteq T_i\}$
- Absolute Frequency: $freq(P) = |cover(P)|$
- Relative Frequency: $freq(P) = \dfrac{|cover(P)|}{|D|}$

Fouille de Motifs Ensemblistes

Définition du problème

- **Given:**
  - A set of items $I = \{p_1, \ldots, p_n\}$
  - A transactional dataset $D = \{T_1, \ldots, T_m\}$
  - A minimum support α
- **The need:**
  - The set of itemset P s.t.: $freq(P) \geq \alpha$

Fouille de Motifs Ensemblistes

Exemple (1)

- $I = \{a, b, c, d, e\}$
- $D = \{T_1, \ldots, T_{10}\}$

Fouille de Motifs Ensemblistes

Exemple (1)

- $I = \{a, b, c, d, e\}$
- $D = \{T_1, \ldots, T_{10}\}$

H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

- $I = \{a, b, c, d, e\}$
- $D = \{T_1, \ldots, T_{10}\}$

H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

V_D

a	b	c	d	e
1	2	2	1	1
3	7	3	2	3
4	9	4	4	4
5		6	6	5
6		7	8	8
8		8	10	9
10		9		10

Fouille de Motifs Ensemblistes

Exemple (1)

- $I = \{a, b, c, d, e\}$
- $D = \{T_1, \ldots, T_{10}\}$

H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

V_D

a	b	c	d	e
1	2	2	1	1
3	7	3	2	3
4	9	4	4	4
5		6	6	5
6		7	8	8
8		8	10	9
10		9		10

M_D

	a	b	c	d	e
1:	1	0	0	1	1
2:	0	1	1	1	0
3:	1	0	1	0	1
4:	1	0	1	1	1
5:	1	0	0	0	1
6:	1	0	1	1	0
7:	0	1	1	0	0
8:	1	0	1	1	1
9:	0	1	1	0	1
10:	1	0	0	1	1

Fouille de Motifs Ensemblistes

Exemple (1)

- $I = \{a, b, c, d, e\}$
- $D = \{T_1, \ldots, T_{10}\}$

$cover(bc) = ?$

$freq(bc) = ?$

H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

V_D

a	b	c	d	e
1	2	2	1	1
3	7	3	2	3
4	9	4	4	4
5		6	6	5
6		7	8	8
8		8	10	9
10		9		10

M_D

	a	b	c	d	e
1:	1	0	0	1	1
2:	0	1	1	1	0
3:	1	0	1	0	1
4:	1	0	1	1	1
5:	1	0	0	0	1
6:	1	0	1	1	0
7:	0	1	1	0	0
8:	1	0	1	1	1
9:	0	1	1	0	1
10:	1	0	0	1	1

Fouille de Motifs Ensemblistes

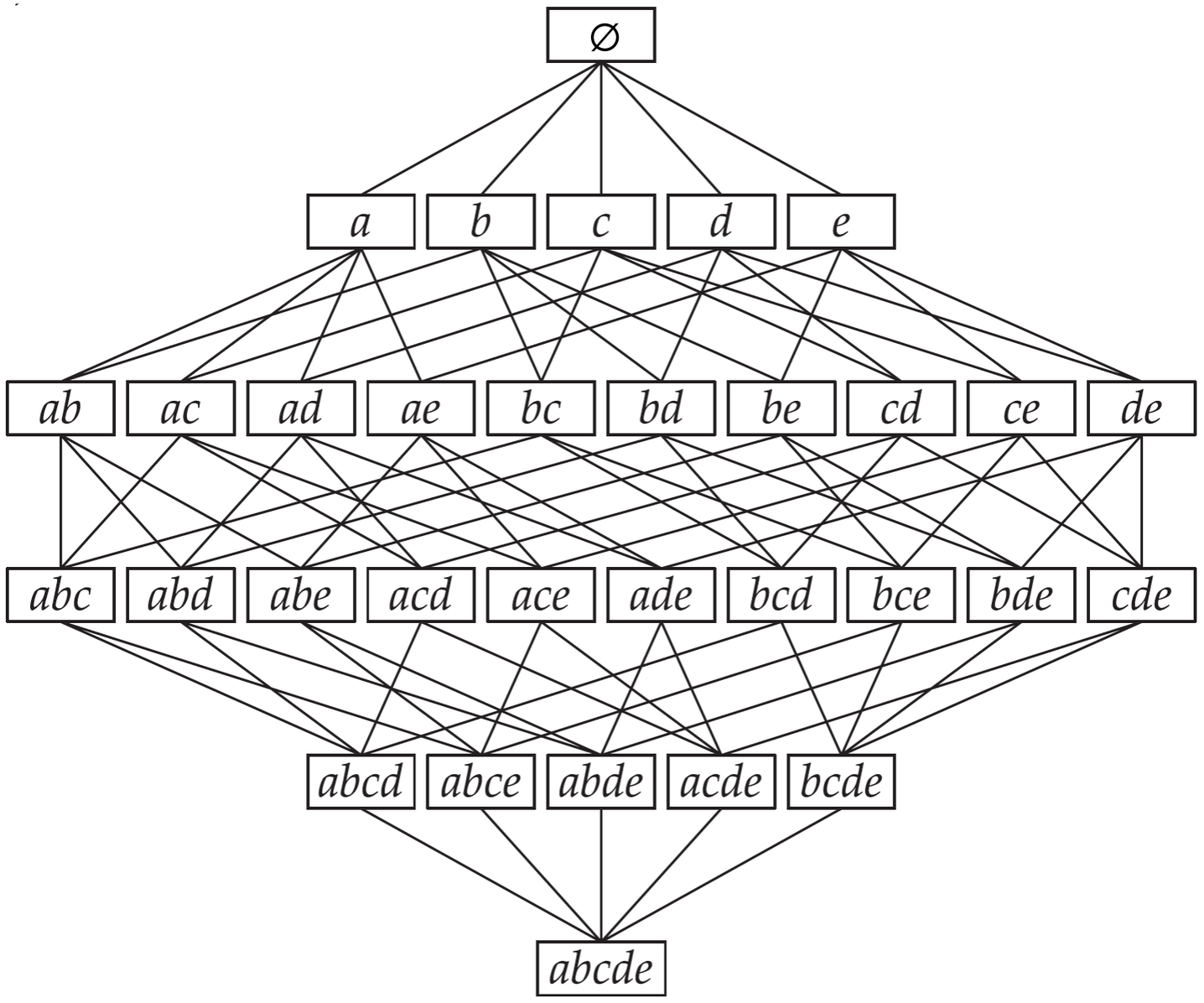
Exemple (1)

H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

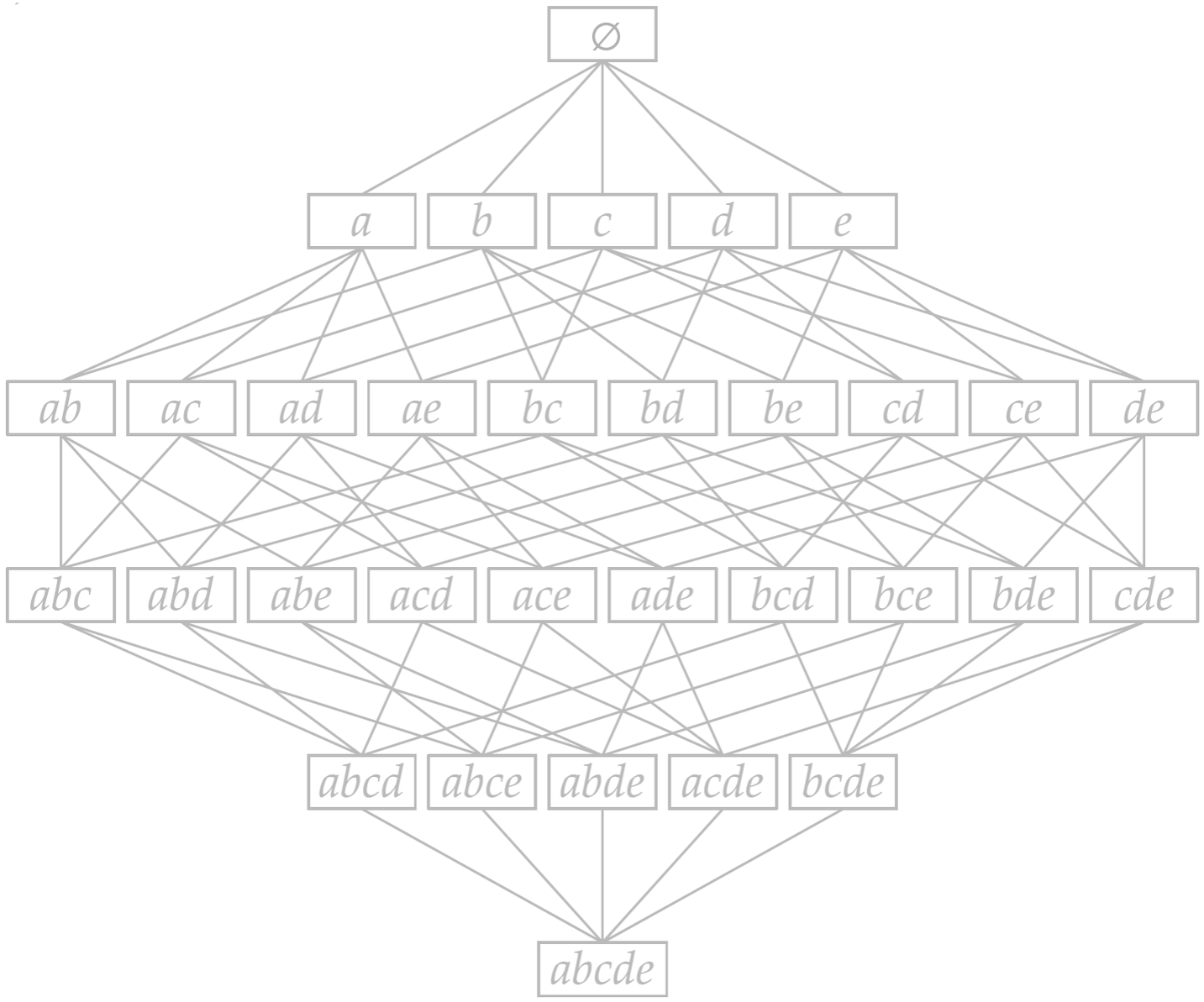


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)



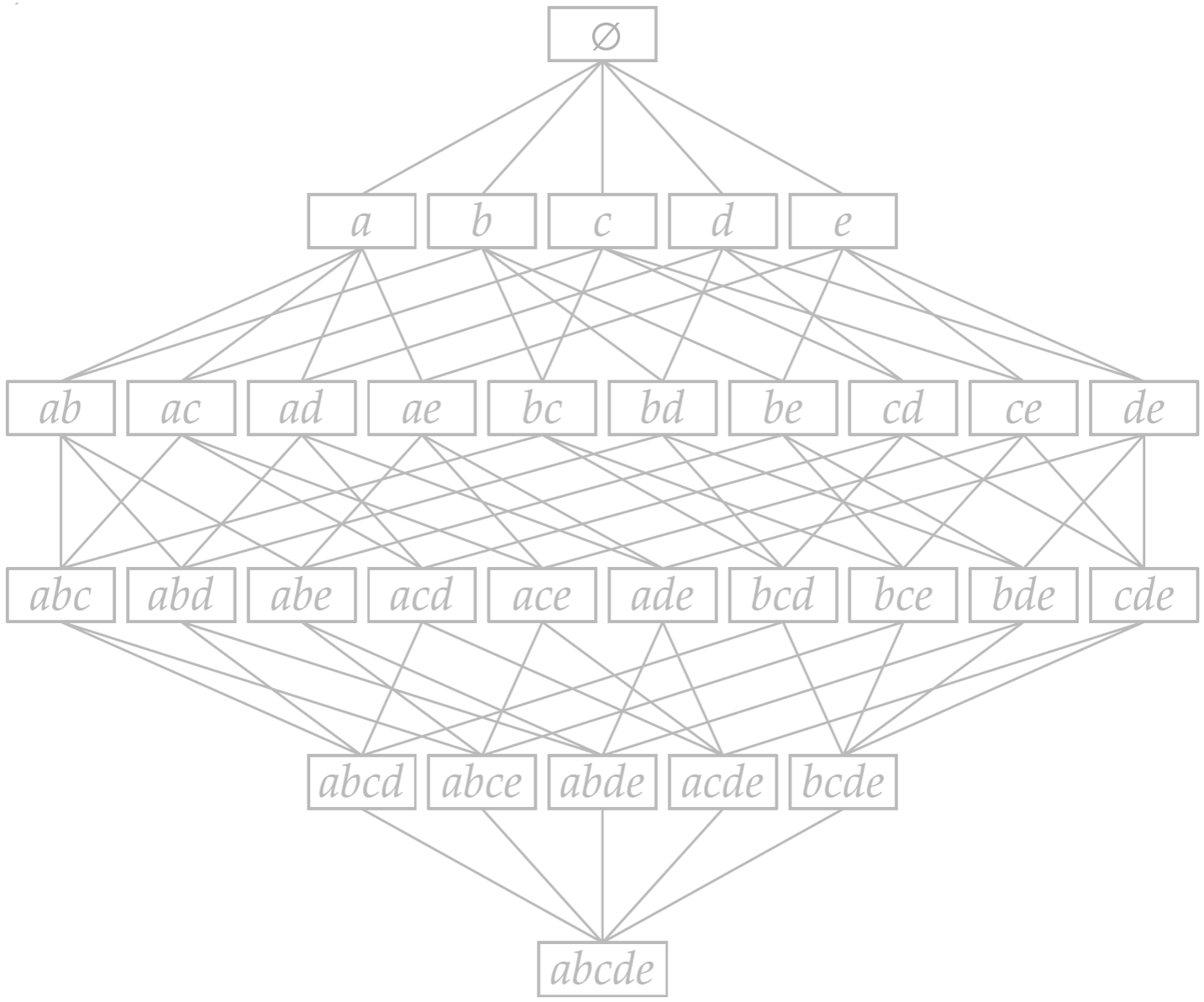
H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

10

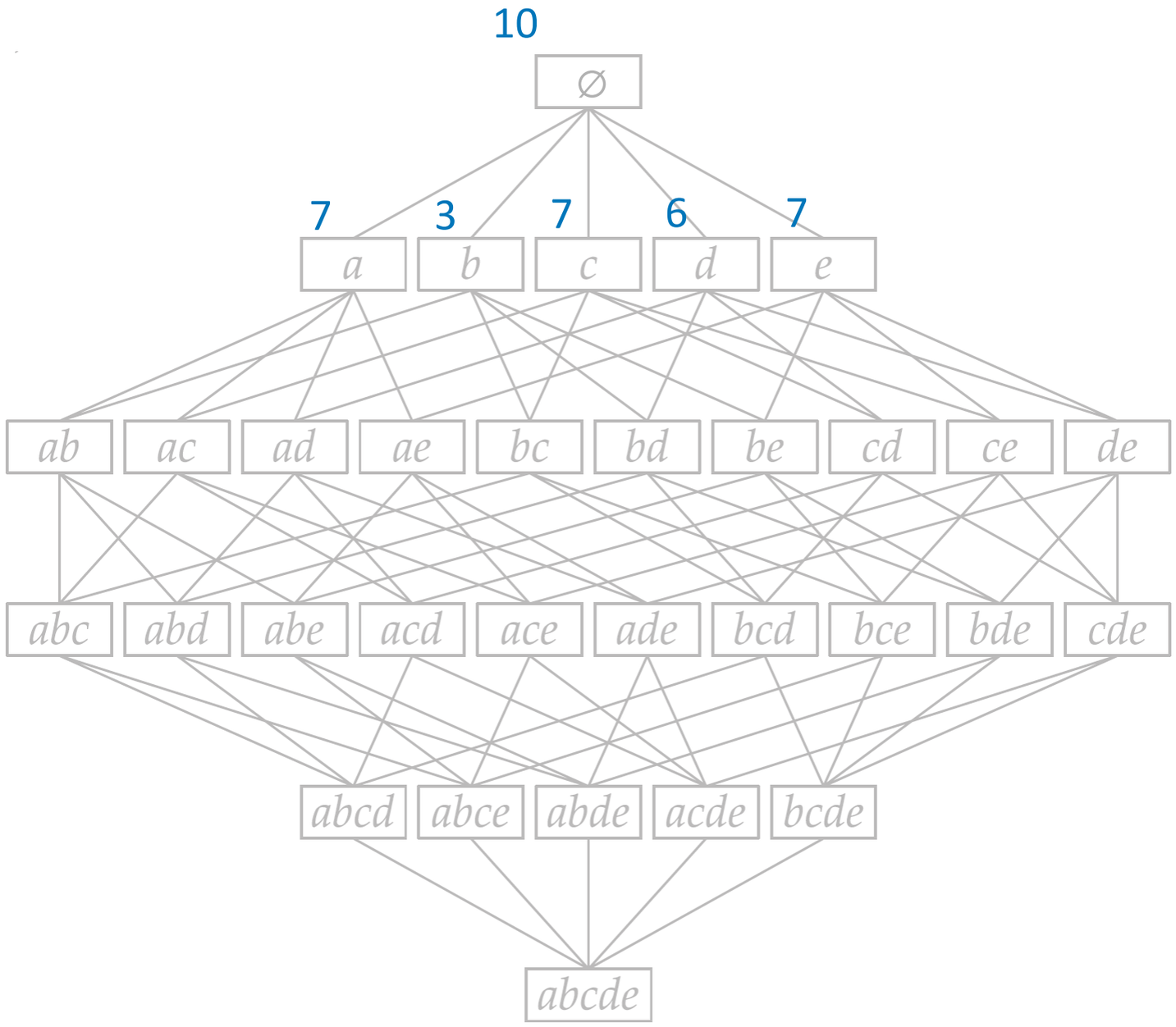


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

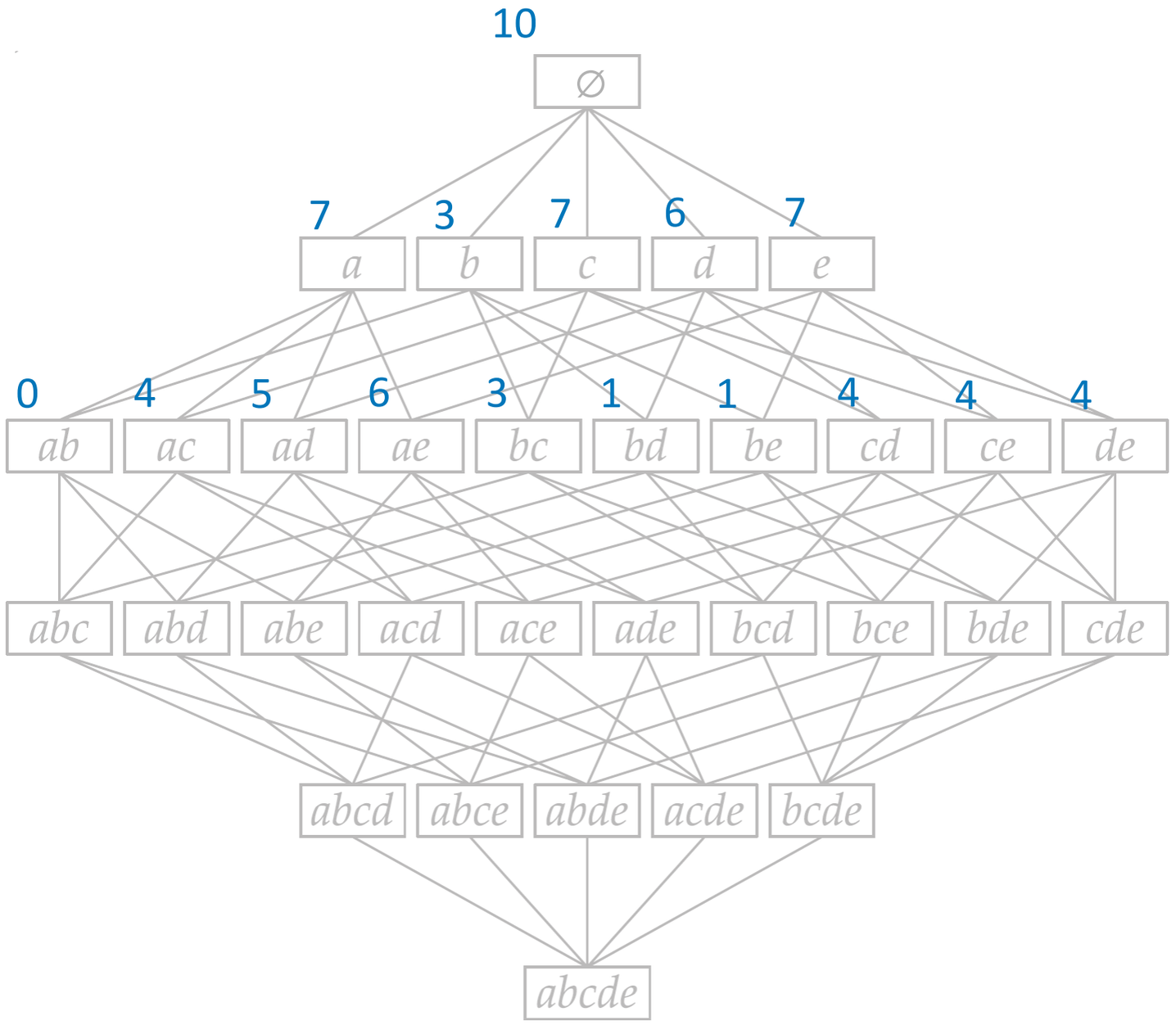


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

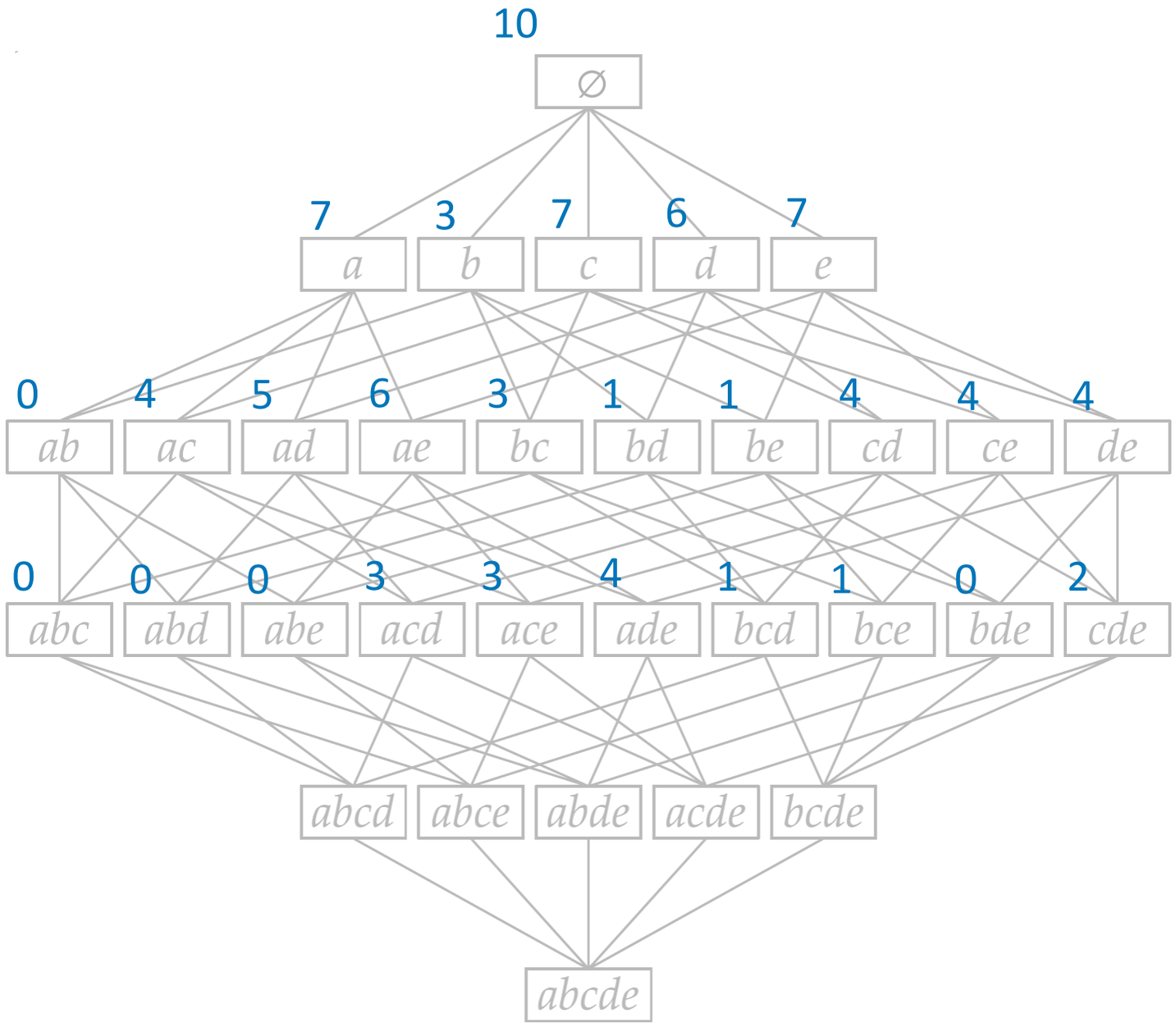


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

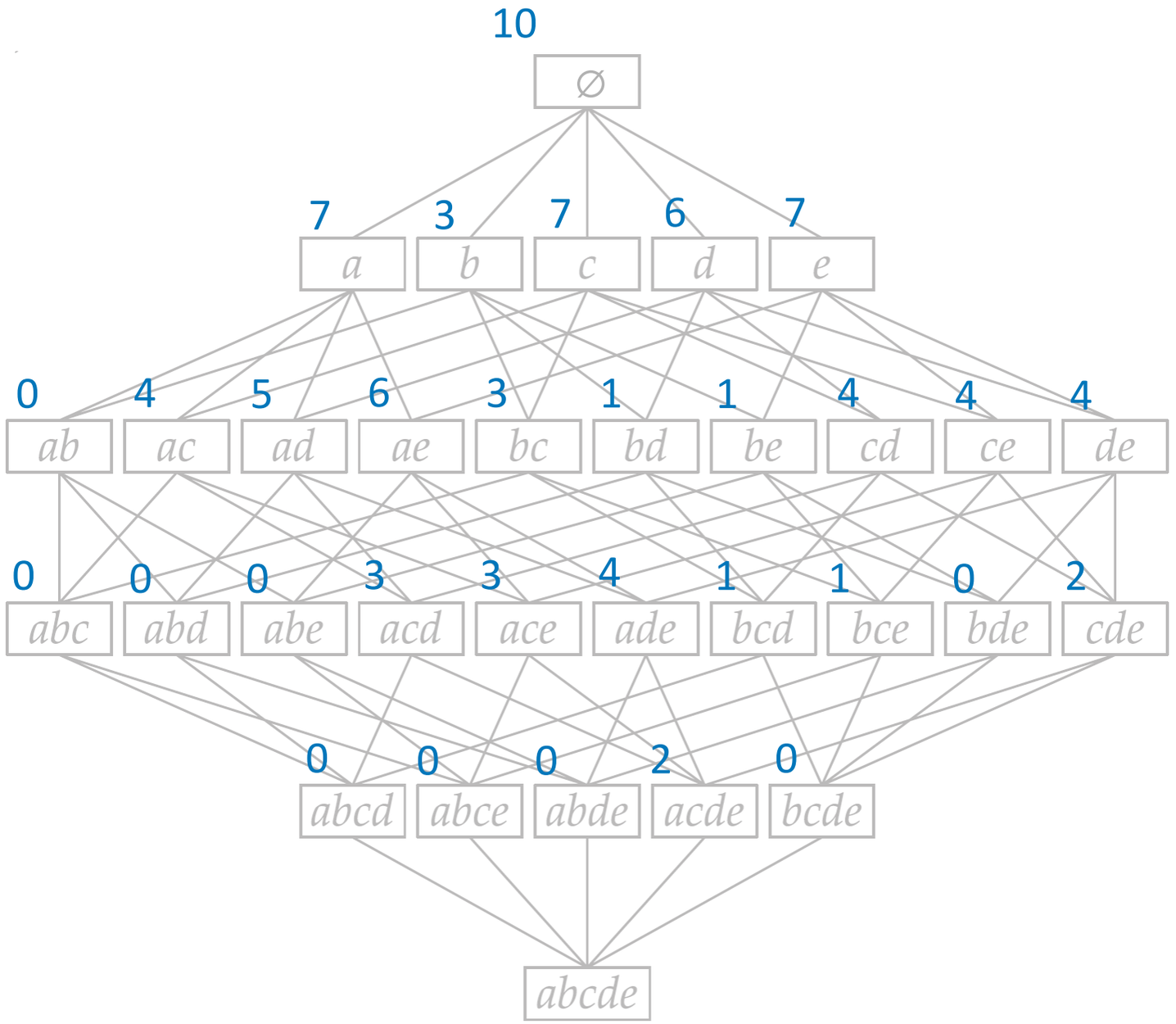


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

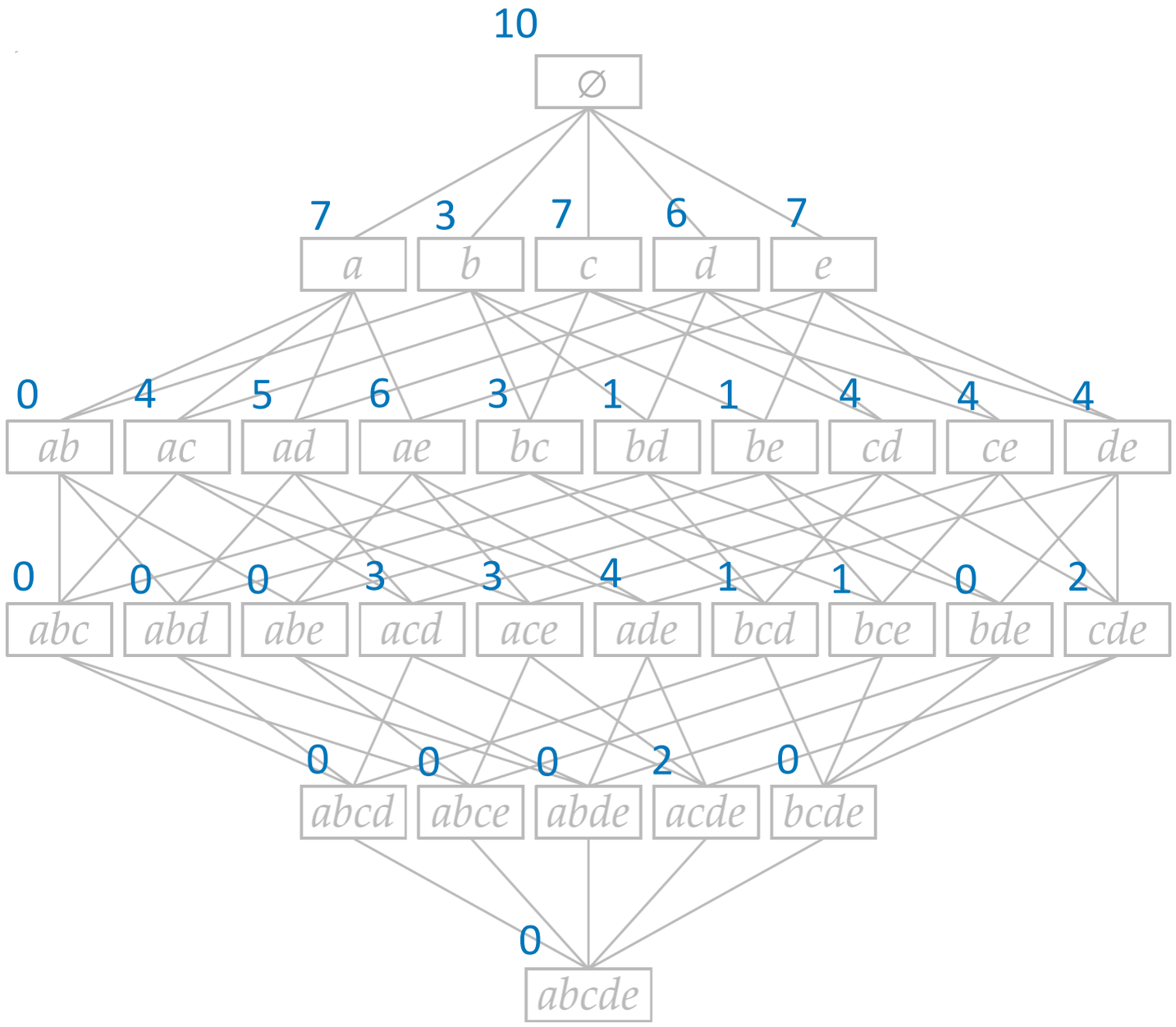


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)

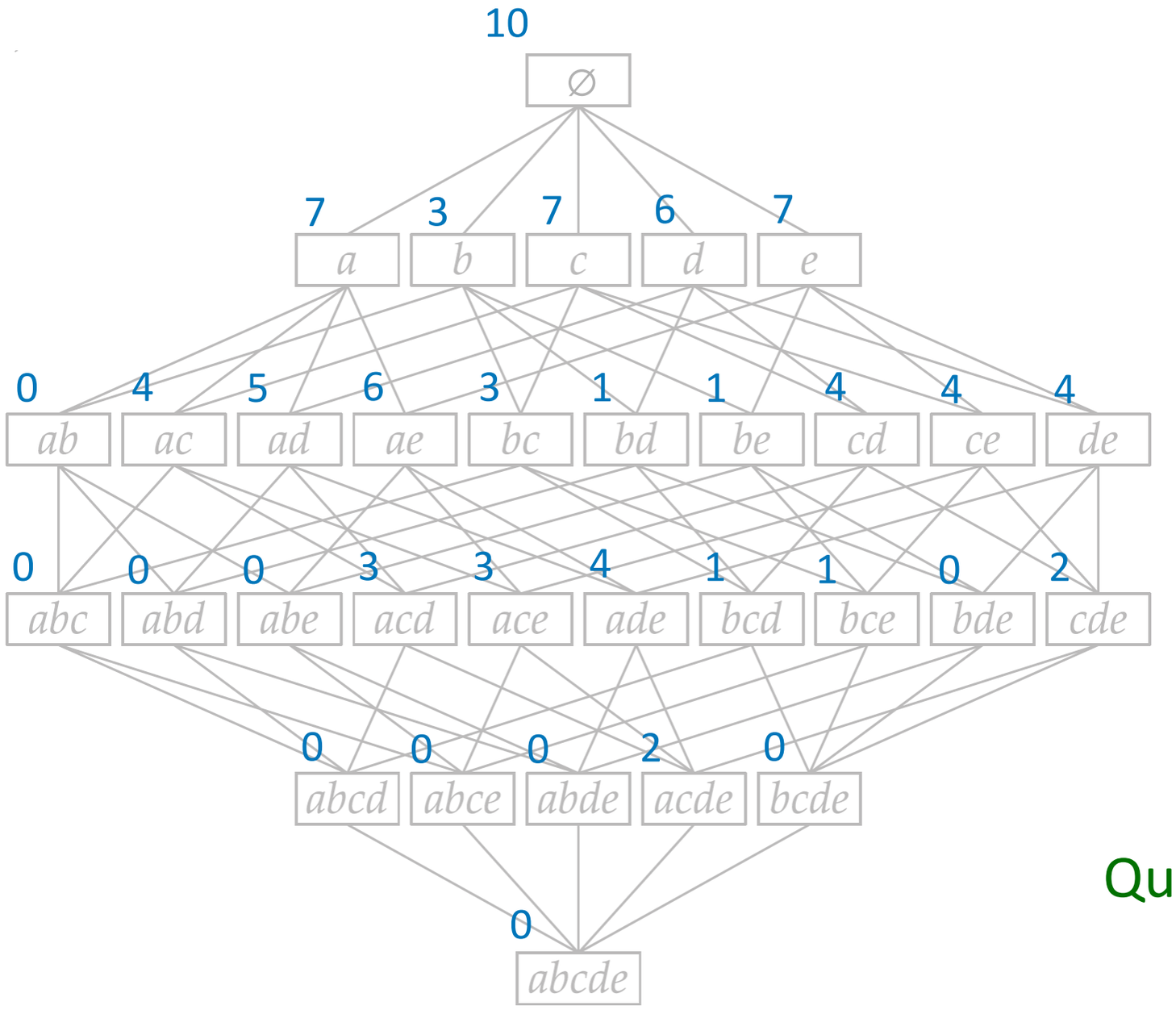


H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Fouille de Motifs Ensemblistes

Exemple (1)



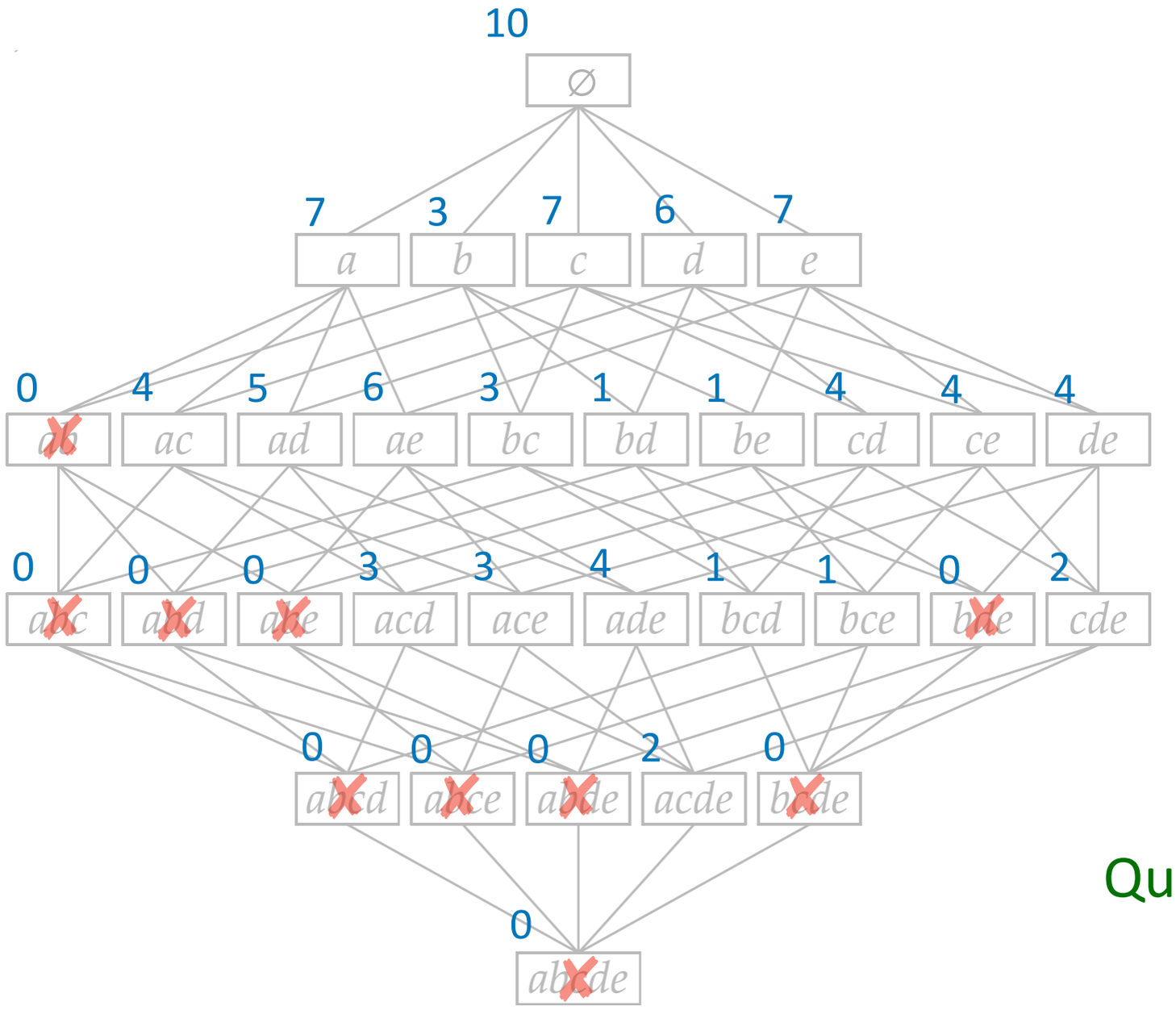
H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Query-1: Frequent itemset?

Fouille de Motifs Ensemblistes

Exemple (1)



H_D

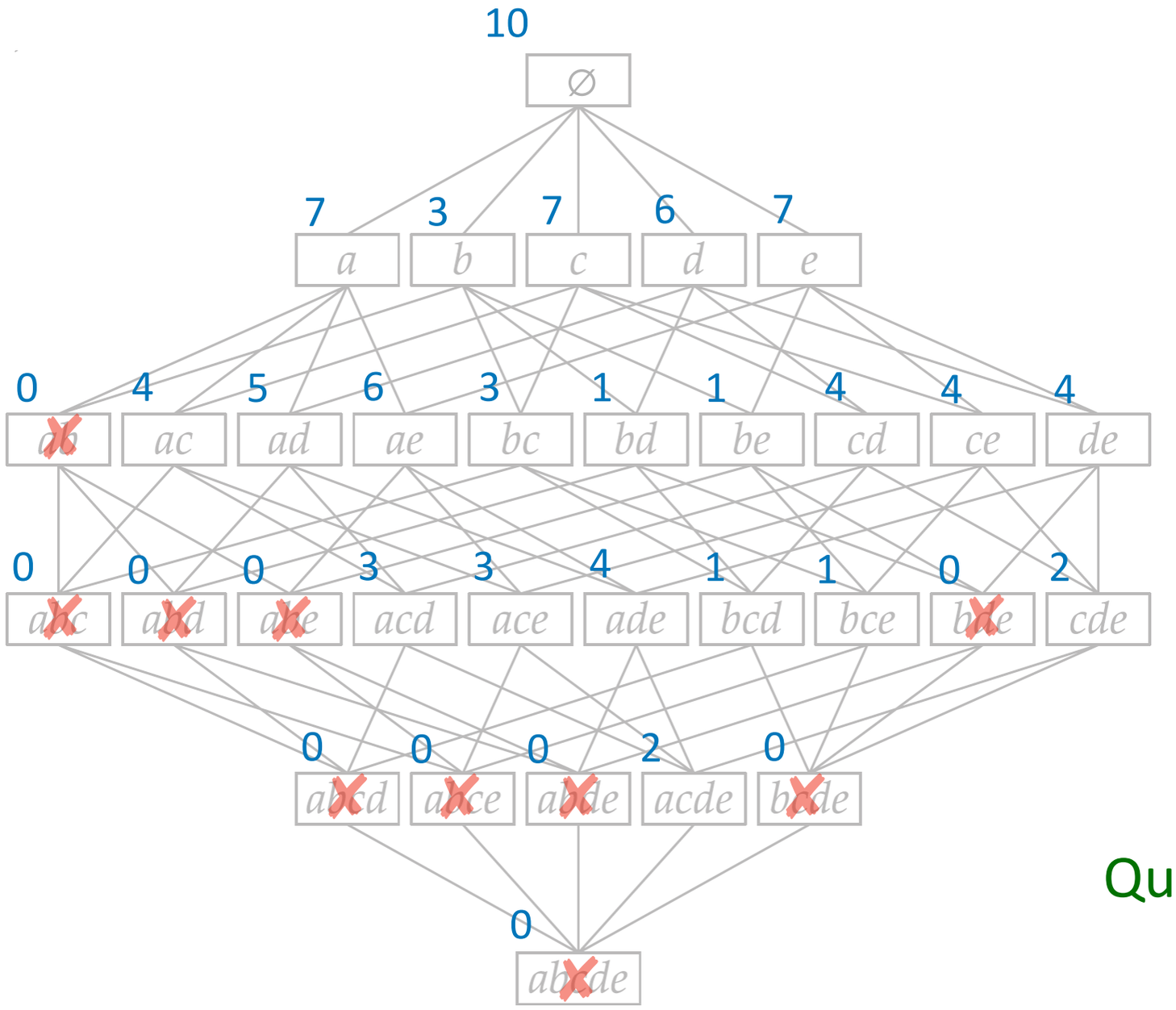
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Query-1: Frequent itemset?

Fouille de Motifs Ensemblistes

Exemple (1)

Query-2: Frequent itemset with minimum support $\alpha = 3$?



H_D

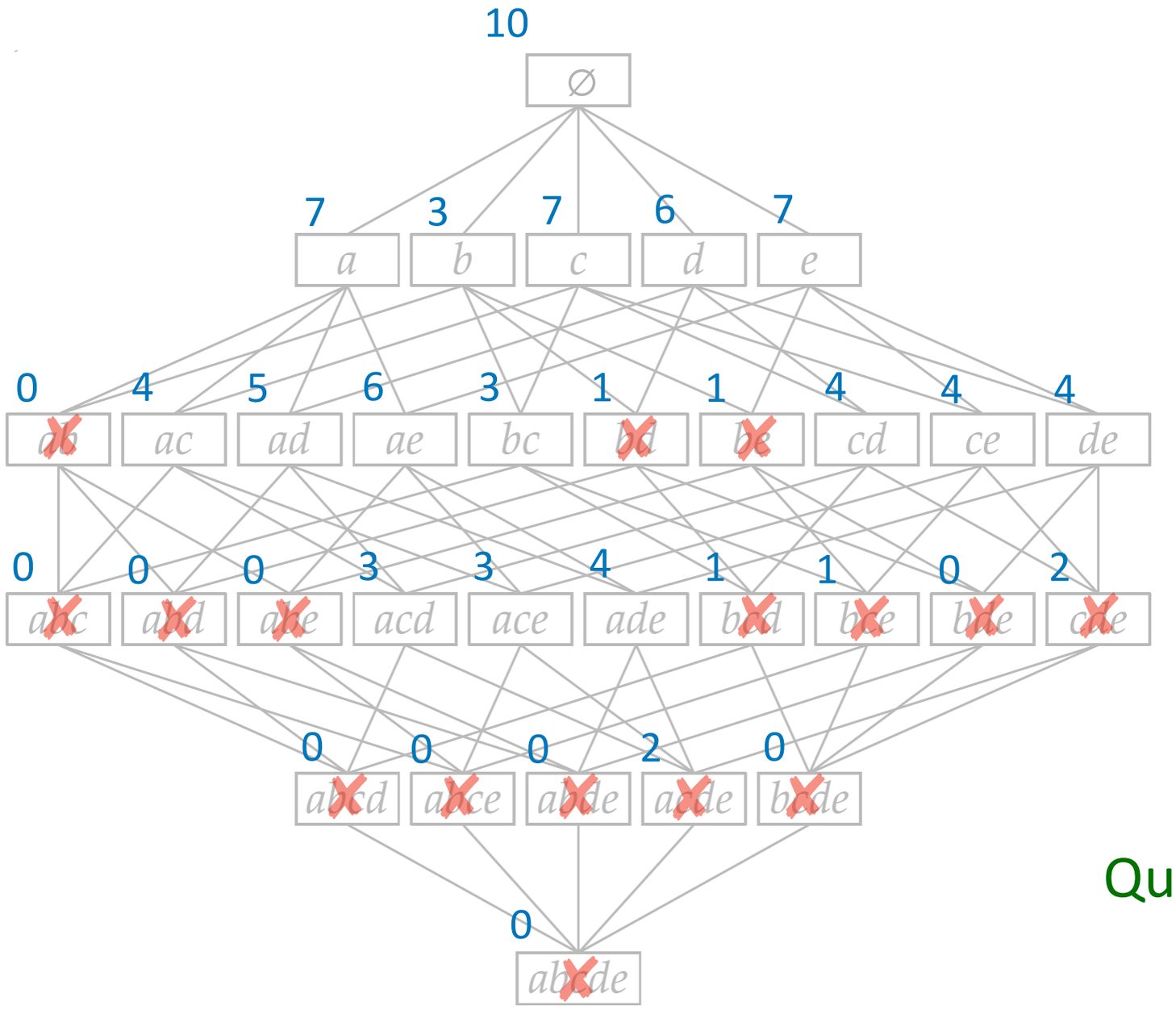
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Query-1: Frequent itemset?

Fouille de Motifs Ensemblistes

Exemple (1)

Query-2: Frequent itemset with minimum support $\alpha = 3$?



H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Query-1: Frequent itemset?

Recherche de Motifs

Recherche naive

- A naïve search that consists of enumerating and testing the frequency of itemset candidates in a given dataset is usually infeasible

Recherche de Motifs

Recherche naive

- A naïve search that consists of enumerating and testing the frequency of itemset candidates in a given dataset is usually infeasible

Number of items (n)	Search space (2^n)
10	$\approx 10^3$
20	$\approx 10^6$
30	$\approx 10^9$
100	$\approx 10^{30}$
128	$\approx 10^{68}$ (atoms in the universe)
1000	$\approx 10^{301}$

Recherche de Motifs

Propriété anti-monotone

- Given a transaction database D over items I and two itemsets P and Q:

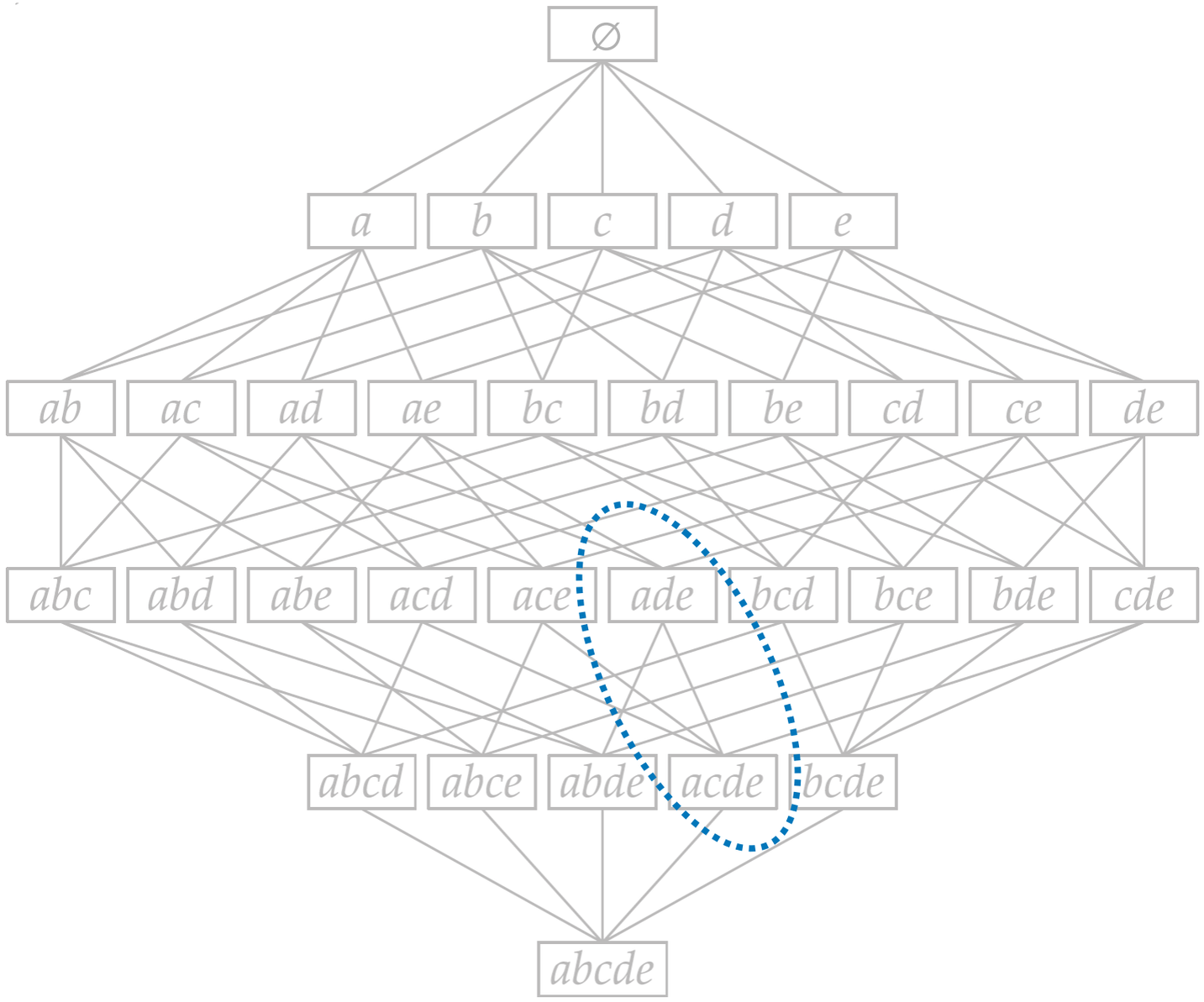
$$Q \subseteq P \Rightarrow cover(P) \subseteq cover(Q)$$

- That is,

$$Q \subseteq P \Rightarrow freq(P) \leq freq(Q)$$

Propriété Anti-Monotone

Exemple (2)



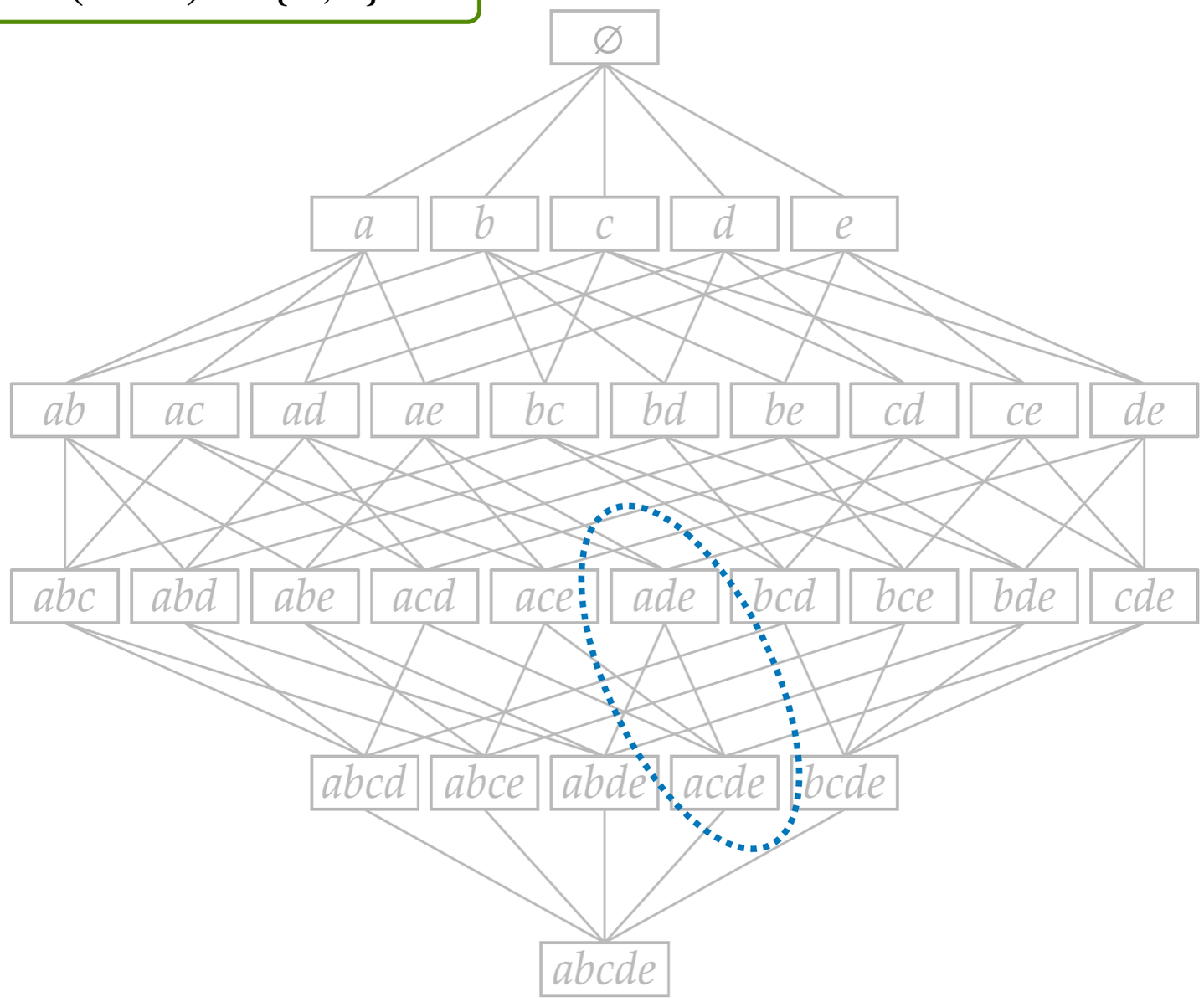
H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Anti-Monotone

Exemple (2)

$cover(ade) = \{1,4,8,10\}$
$cover(acde) = \{4,8\}$



H_D

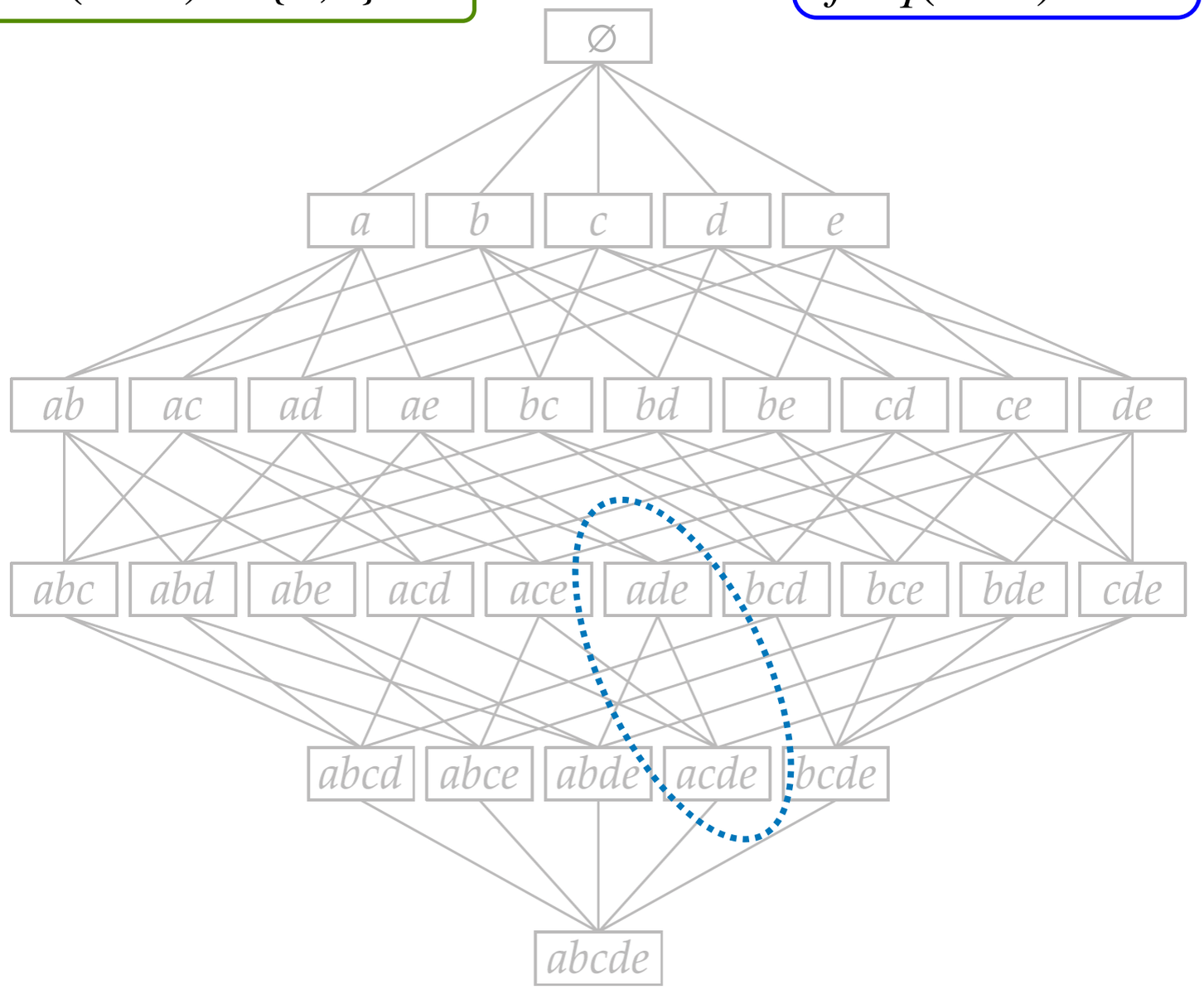
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Anti-Monotone

Exemple (2)

$cover(ade) = \{1,4,8,10\}$
$cover(acde) = \{4,8\}$

$freq(ade) = 4$
$freq(acde) = 2$



H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Recherche de Motifs

Propriété Apriori

- Given a transaction database D over items I, a minsup α and two itemsets P and Q:

$$Q \subseteq P \Rightarrow freq(P) \leq freq(Q)$$

- It follows: $Q \subseteq P \wedge freq(P) \geq \alpha \Rightarrow freq(Q) \geq \alpha$

- Contraposition: $Q \subseteq P \wedge freq(Q) < \alpha \Rightarrow freq(P) < \alpha$

Recherche de Motifs

Propriété Apriori

- Given a transaction database D over items I, a minsup α and two itemsets P and Q:

$$Q \subseteq P \Rightarrow freq(P) \leq freq(Q)$$

- It follows: $Q \subseteq P \wedge freq(P) \geq \alpha \Rightarrow freq(Q) \geq \alpha$

All subsets of a frequent itemset are frequent!

- Contraposition: $Q \subseteq P \wedge freq(Q) < \alpha \Rightarrow freq(P) < \alpha$

Recherche de Motifs

Propriété Apriori

- Given a transaction database D over items I, a minsup α and two itemsets P and Q:

$$Q \subseteq P \Rightarrow freq(P) \leq freq(Q)$$

- It follows: $Q \subseteq P \wedge freq(P) \geq \alpha \Rightarrow freq(Q) \geq \alpha$

All subsets of a frequent itemset are frequent!

- Contraposition: $Q \subseteq P \wedge freq(Q) < \alpha \Rightarrow freq(P) < \alpha$

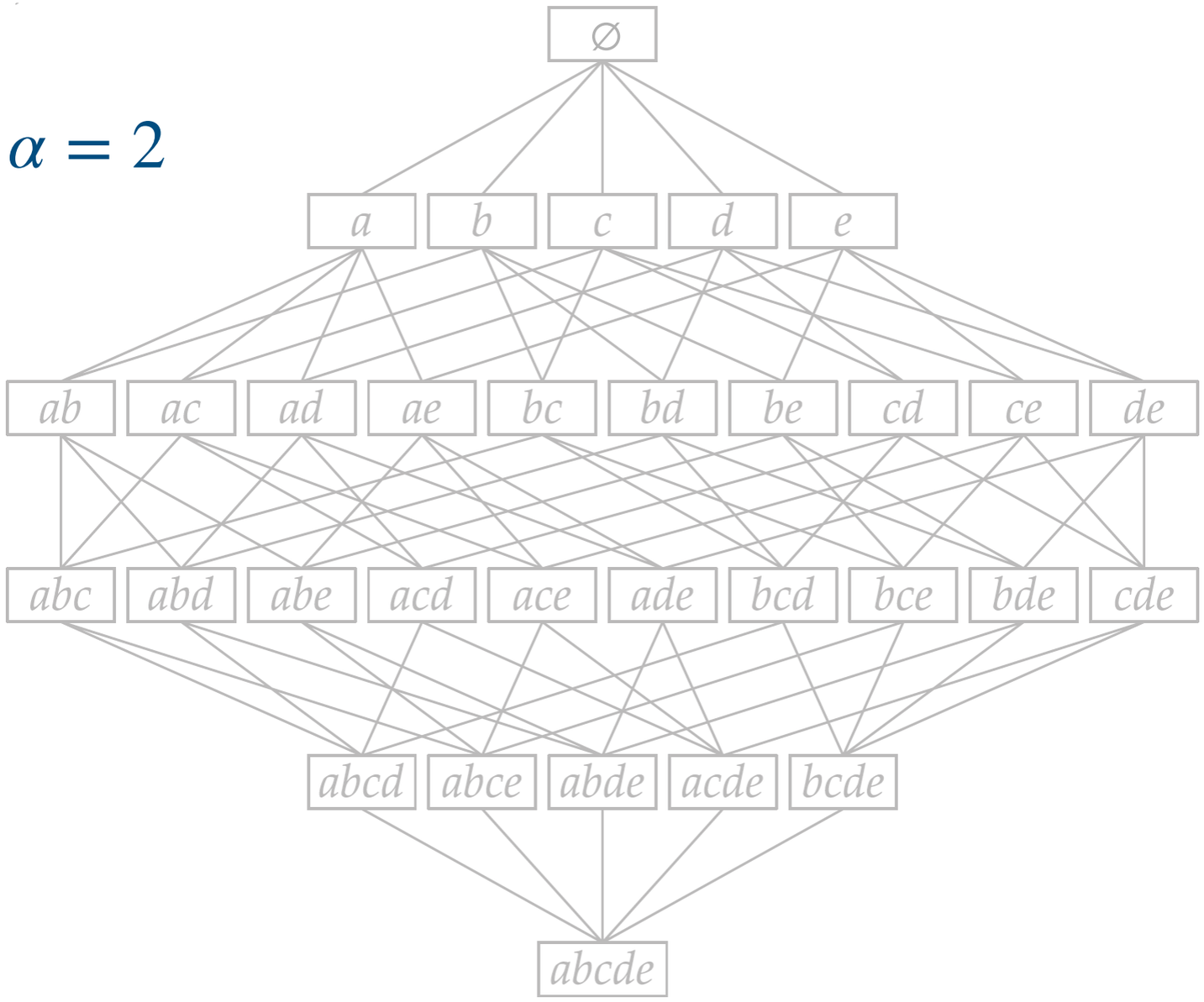
All supersets of an infrequent itemset are infrequent!

Propriété Apriori

Exemple (3)

All subsets of a frequent itemset are frequent!

$\alpha = 2$



H_D

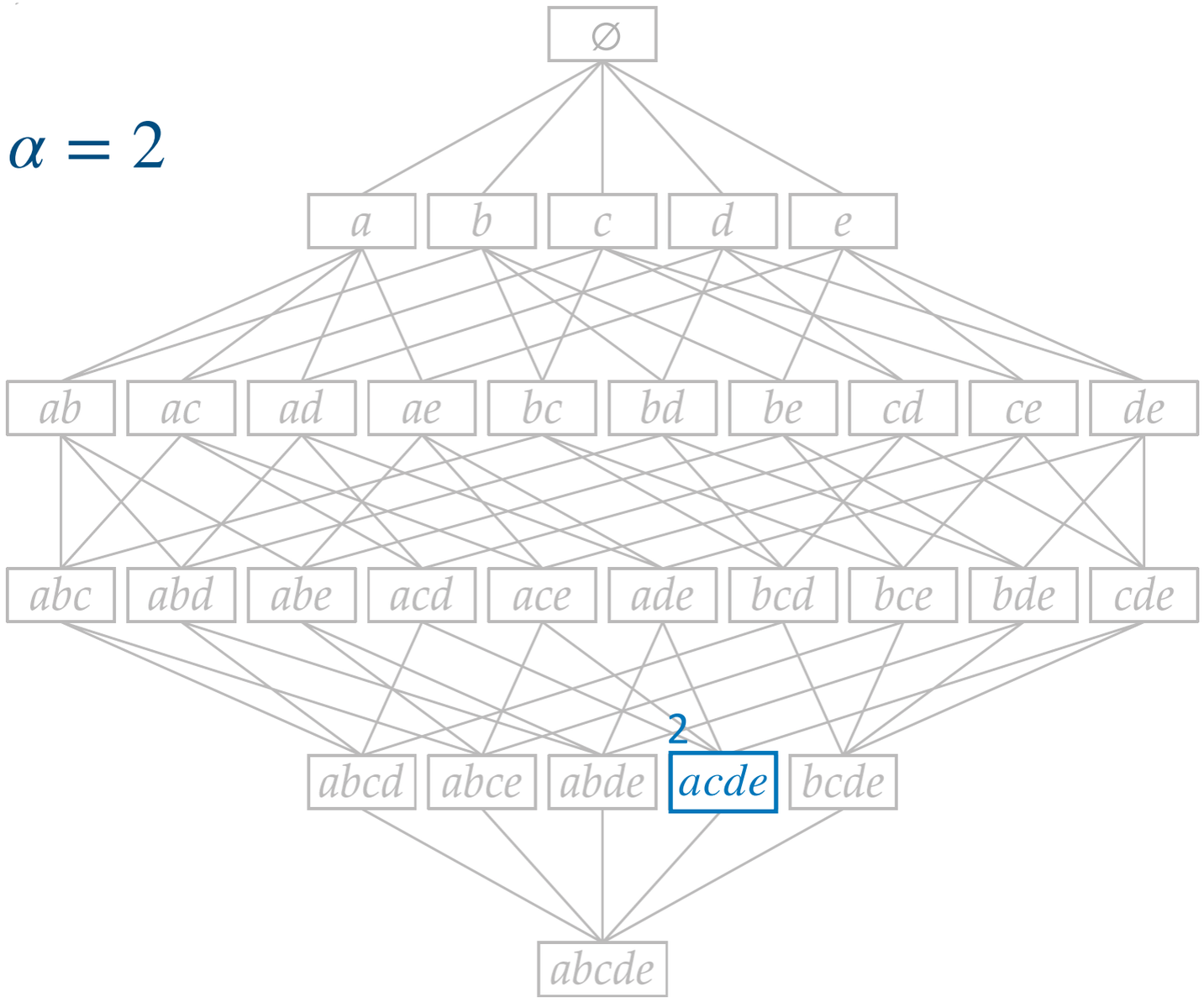
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All subsets of a frequent itemset are frequent!

$\alpha = 2$



H_D

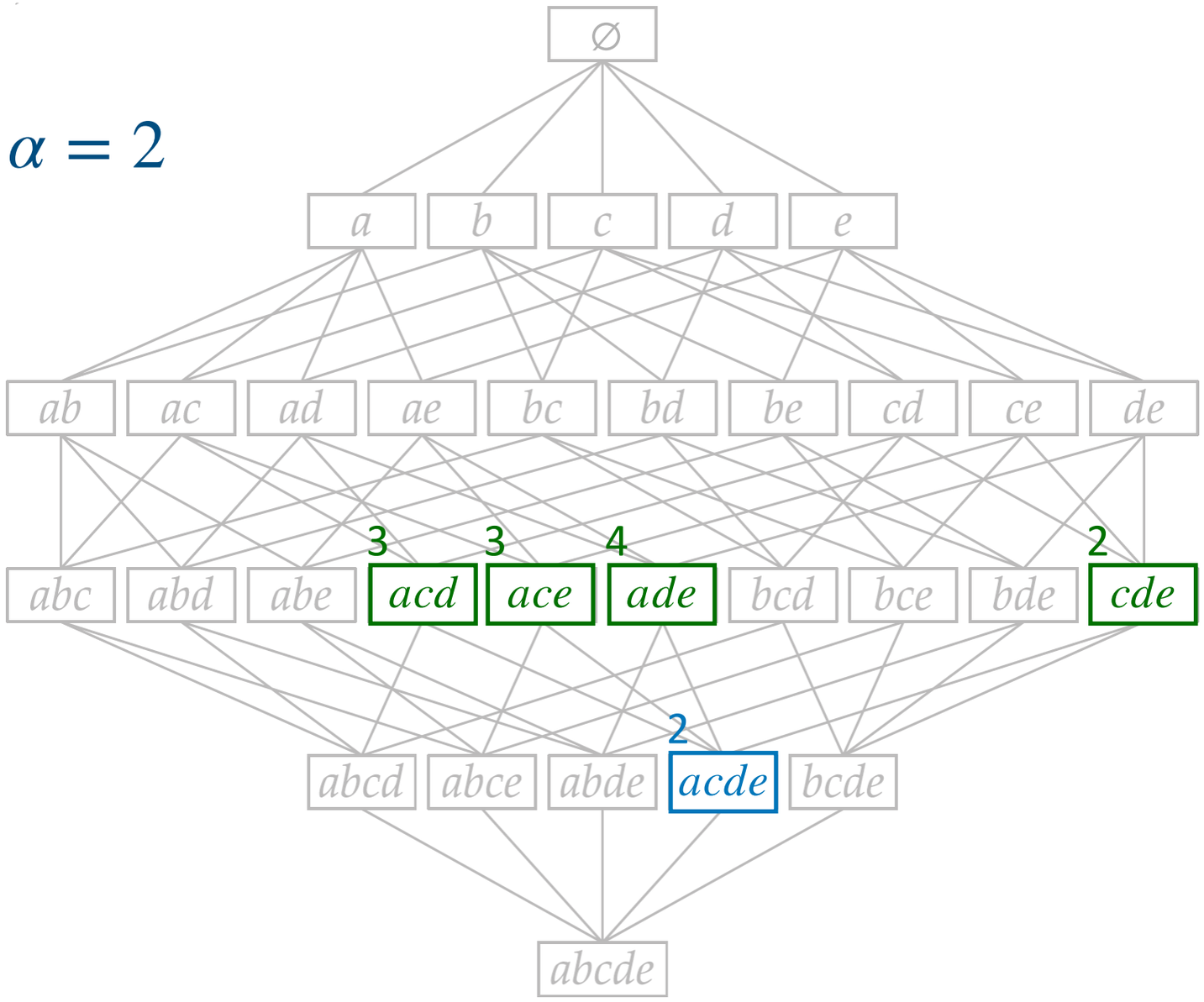
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All subsets of a frequent itemset are frequent!

$\alpha = 2$



H_D

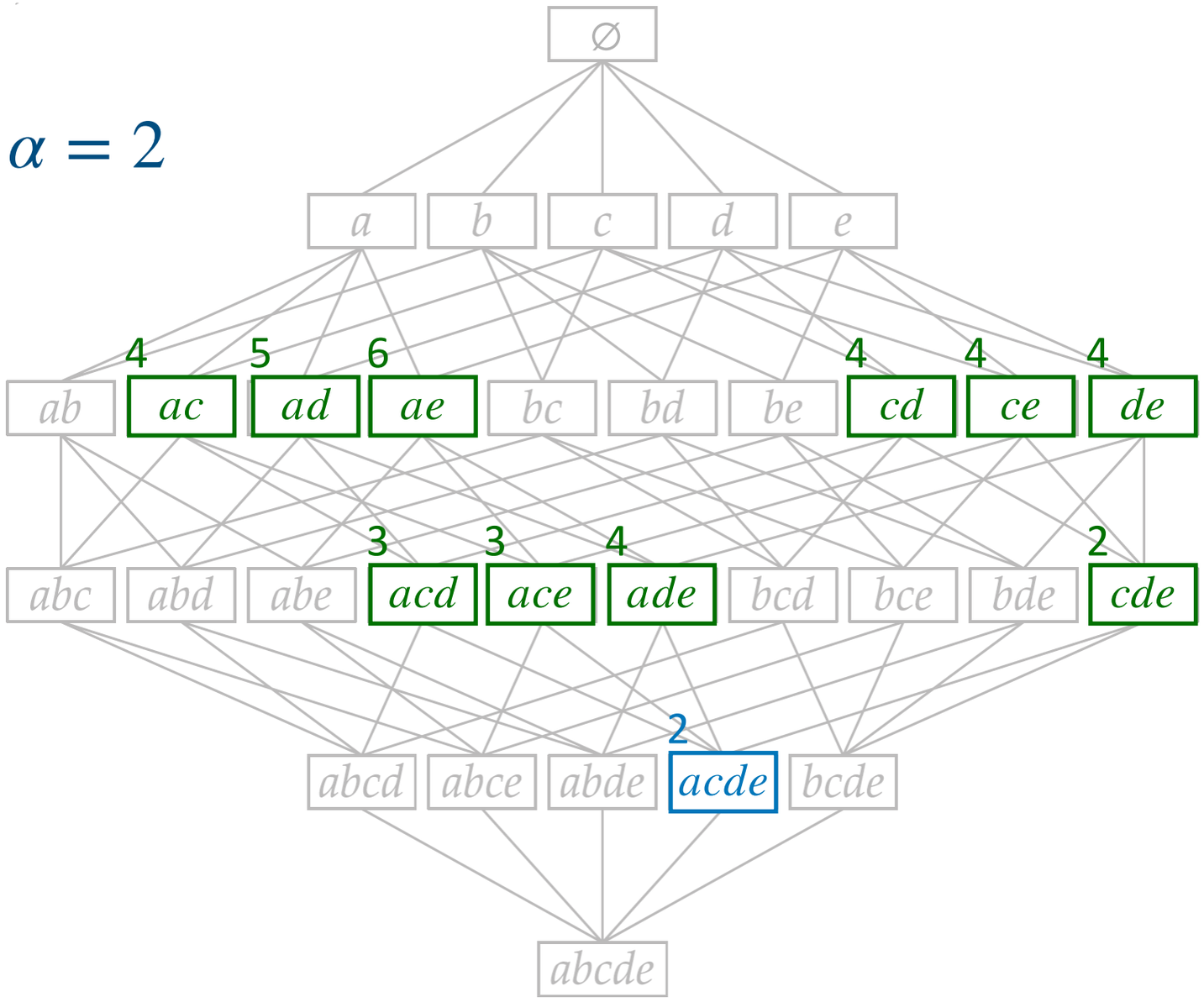
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All subsets of a frequent itemset are frequent!

$\alpha = 2$



H_D

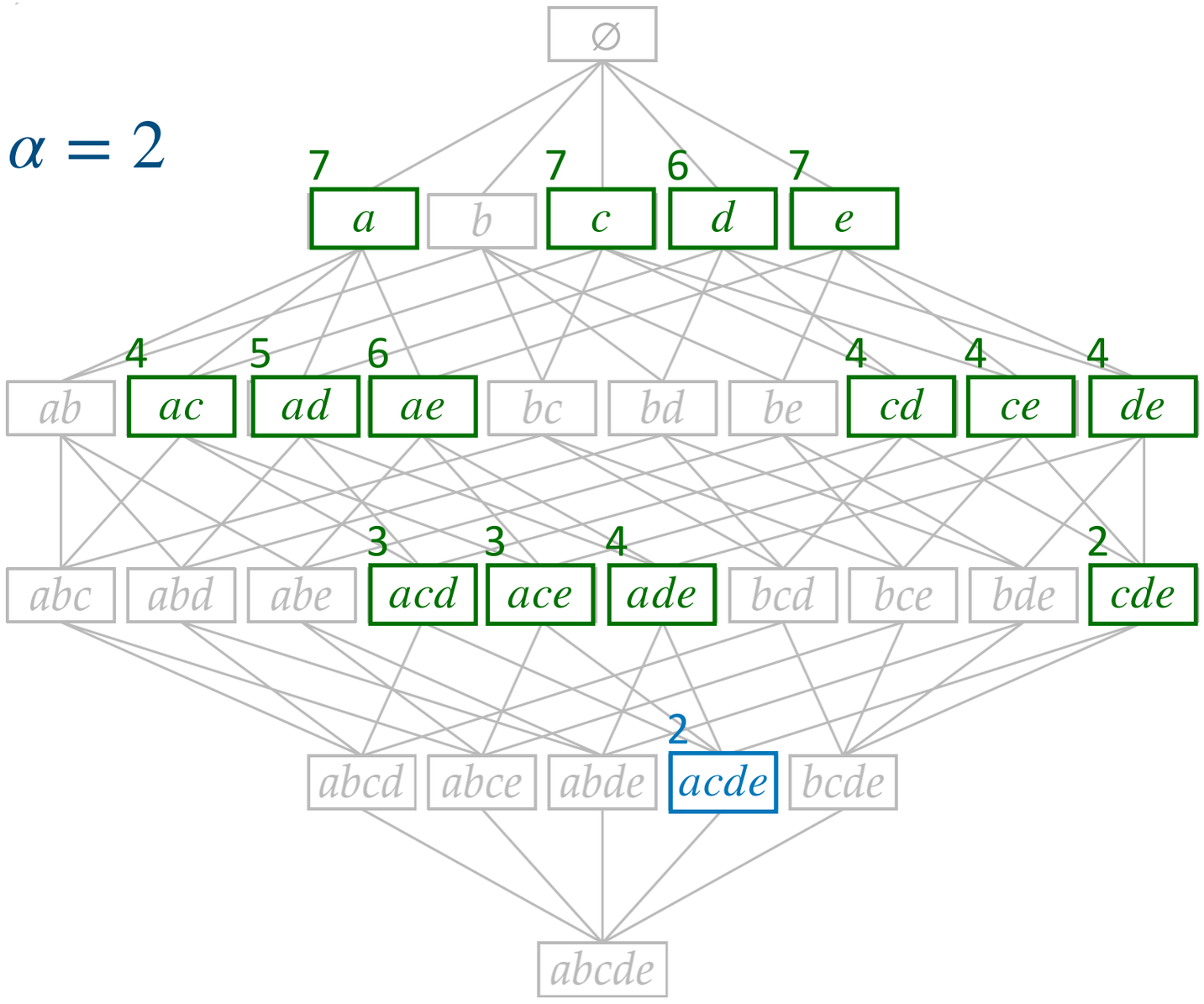
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All subsets of a frequent itemset are frequent!

$\alpha = 2$



H_D

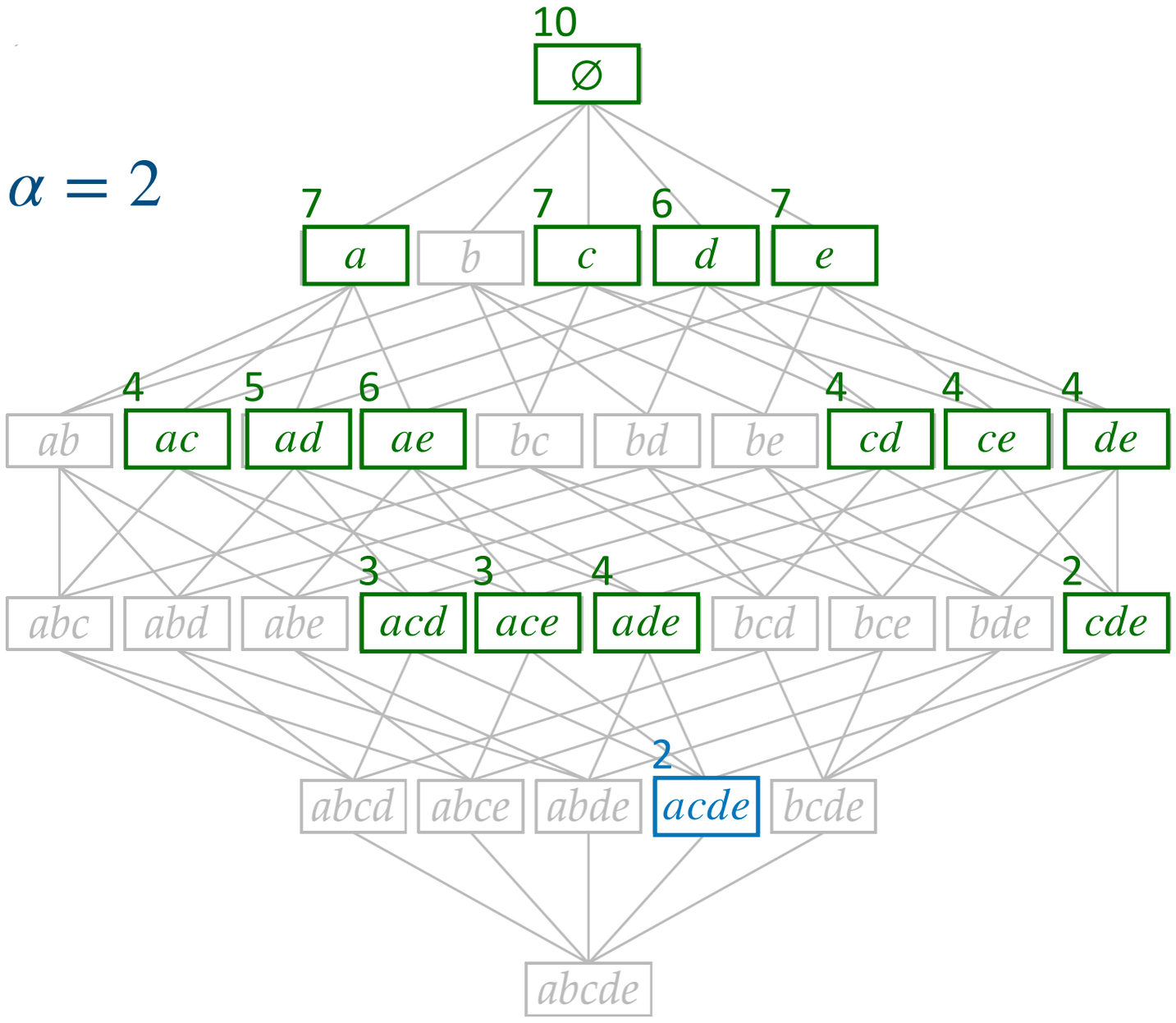
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All subsets of a frequent itemset are frequent!

$\alpha = 2$



H_D

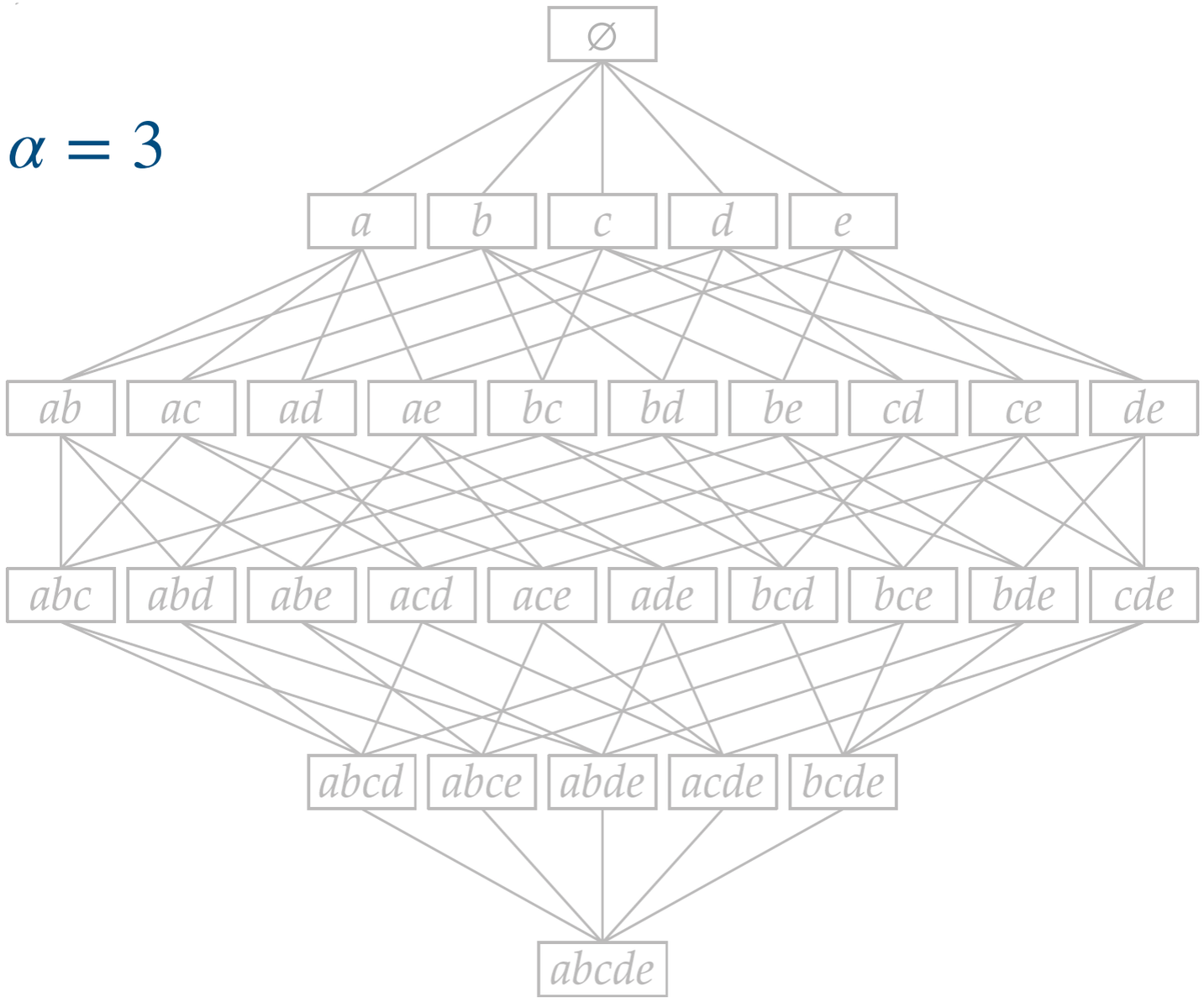
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All supersets of an infrequent itemset are infrequent!

$\alpha = 3$



H_D

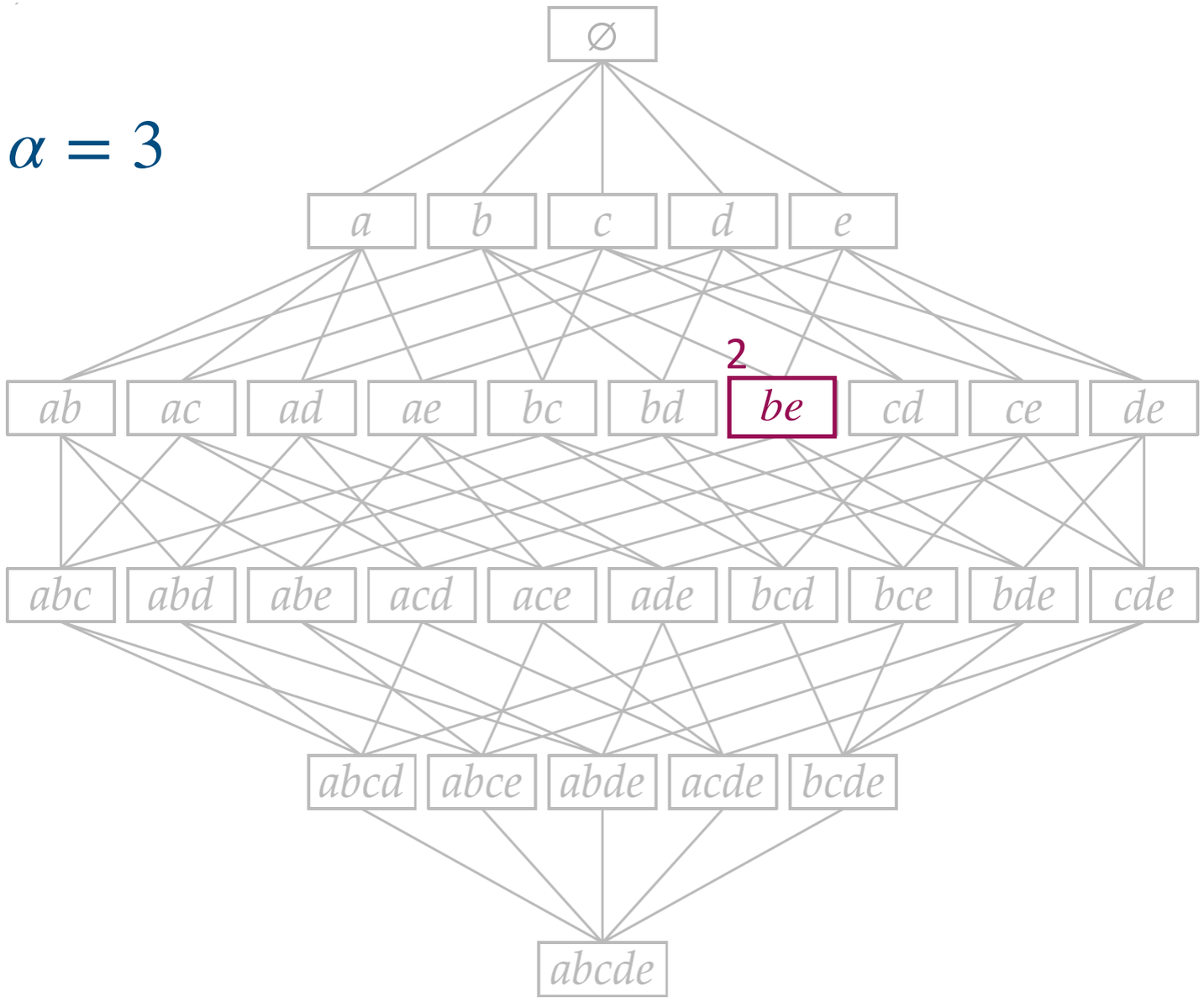
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All supersets of an infrequent itemset are infrequent!

$\alpha = 3$



H_D

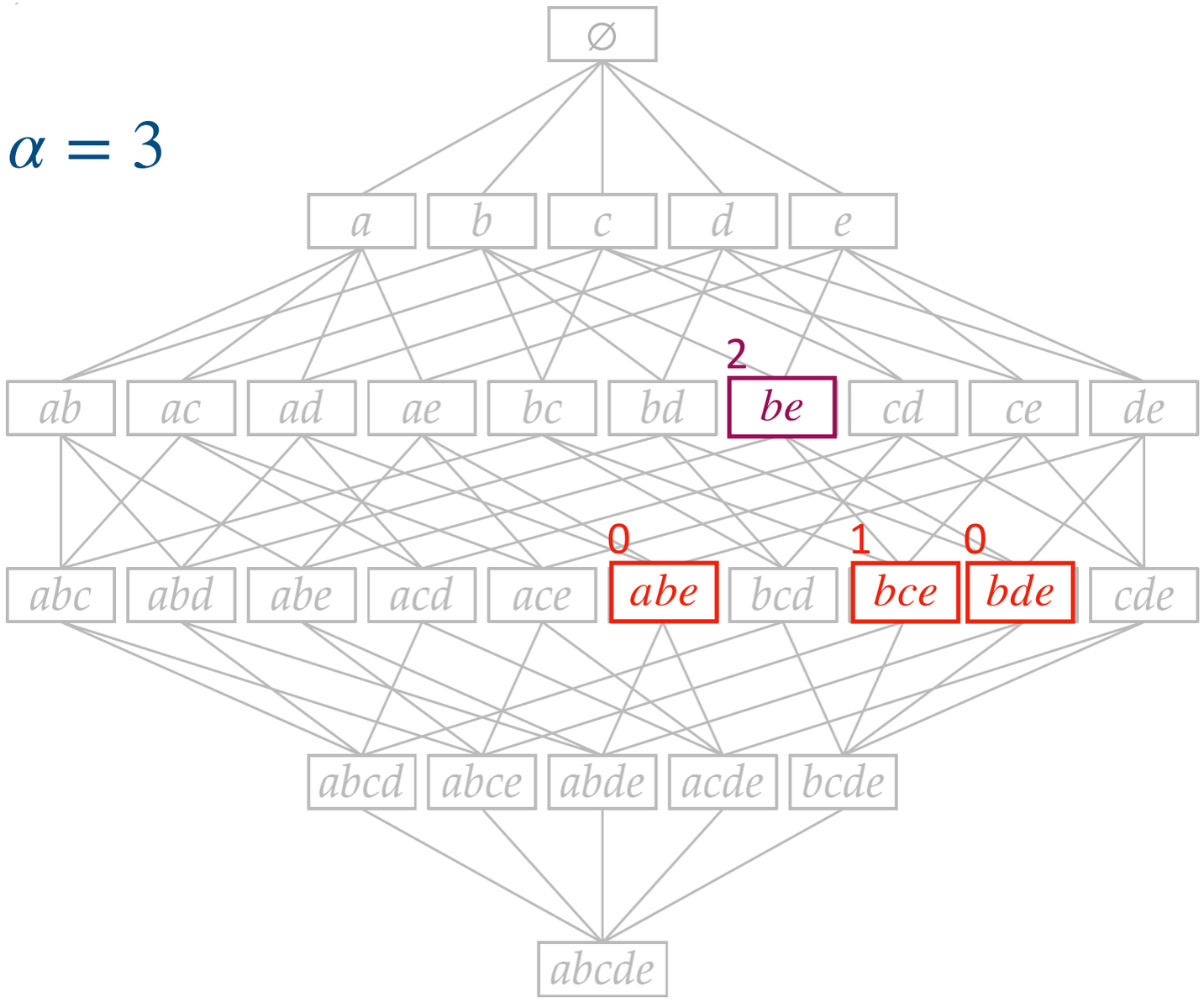
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All supersets of an infrequent itemset are infrequent!

$\alpha = 3$



H_D

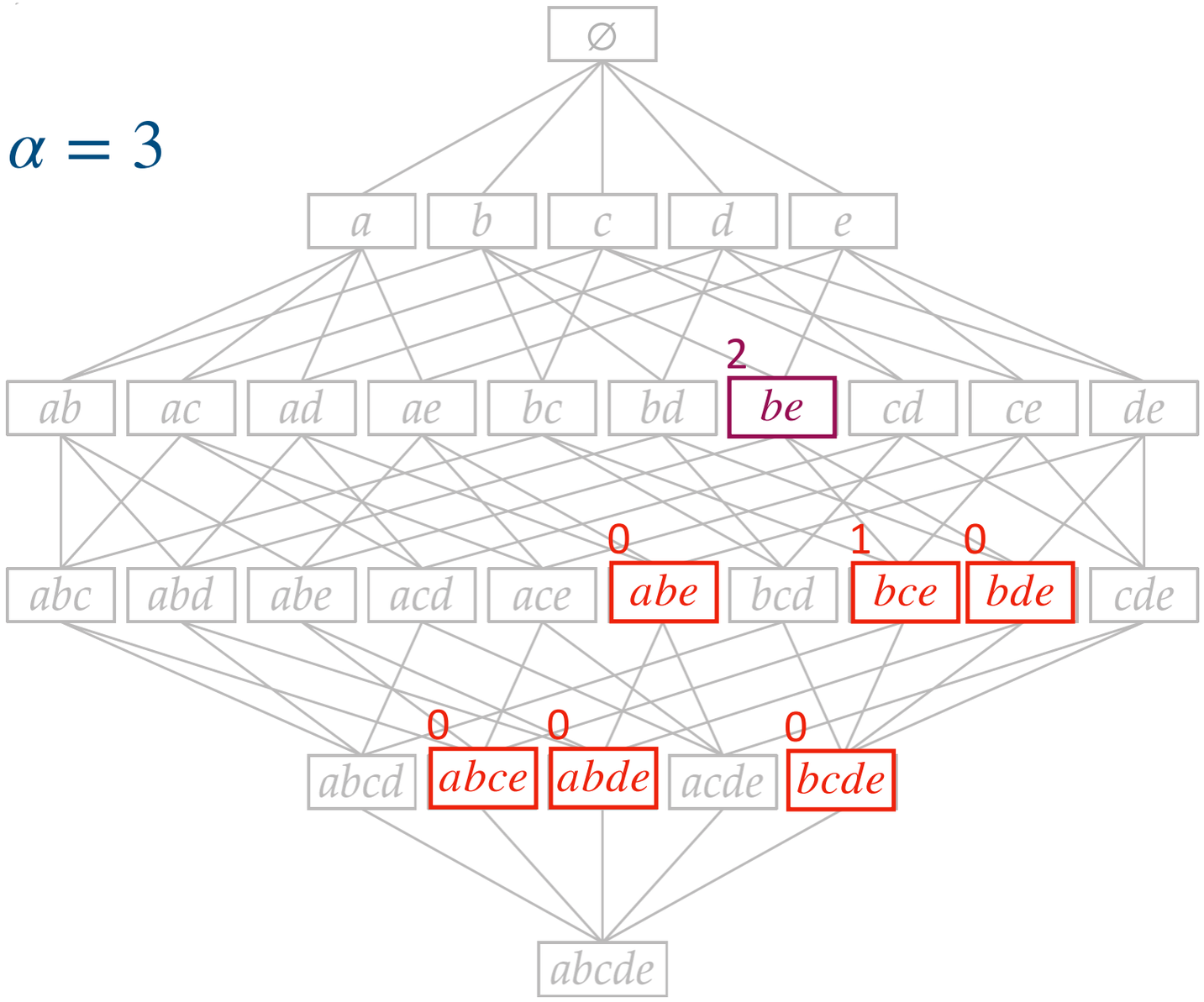
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All supersets of an infrequent itemset are infrequent!

$\alpha = 3$



H_D

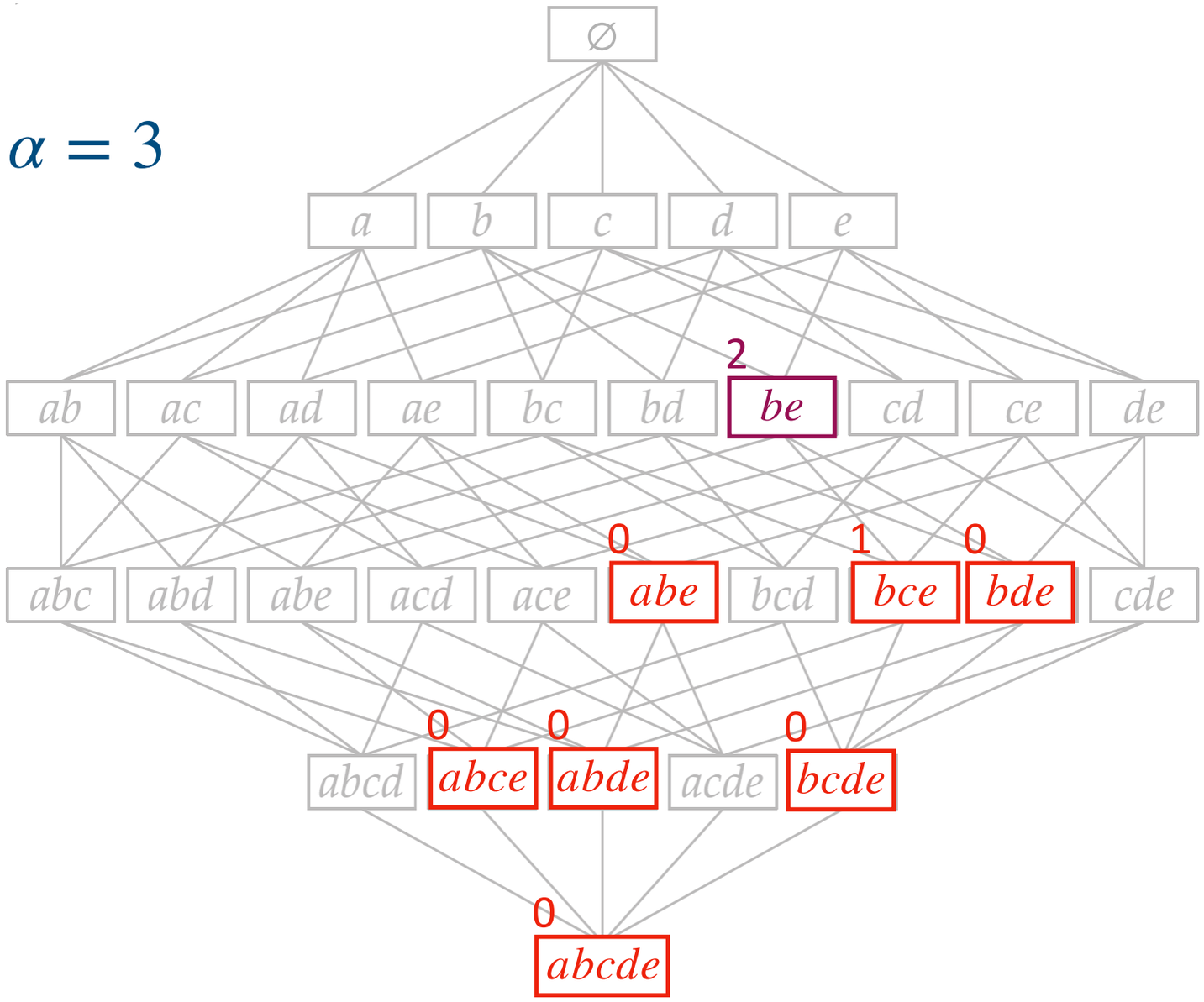
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Propriété Apriori

Exemple (3)

All supersets of an infrequent itemset are infrequent!

$\alpha = 3$



H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

Recherche de Motifs

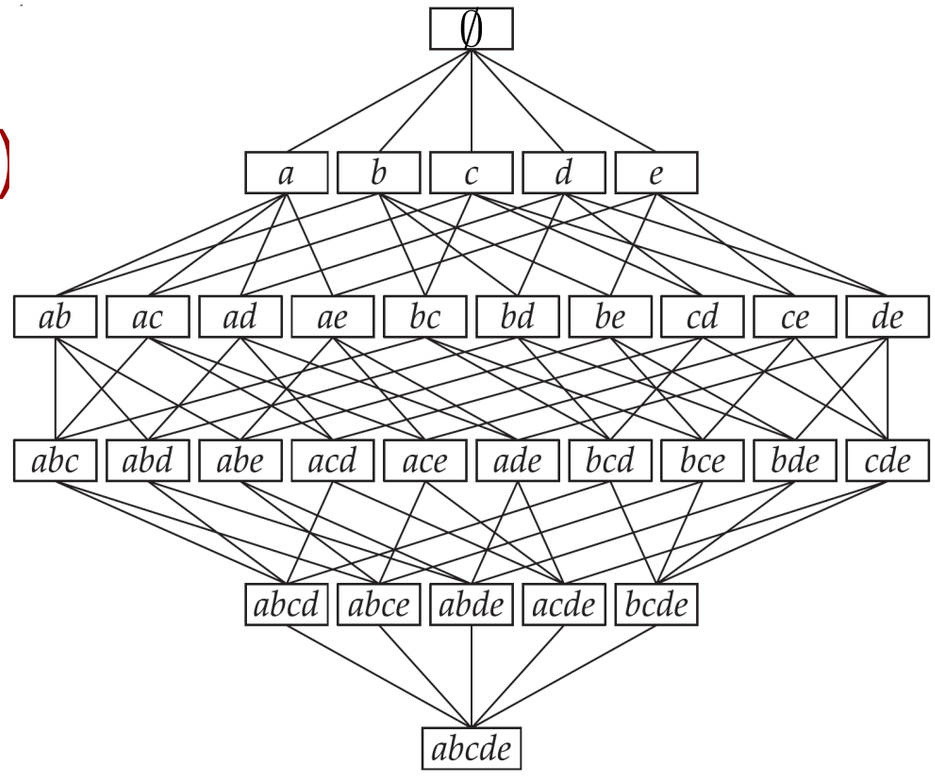
Poset $(2^I, \subseteq)$

- Test A partial order is a binary relation R over a set S :
- $\forall x, y, z \in S$
- $x \; R \; x$ (reflexivity)
- $x \; R \; y \wedge y \; R \; x \Rightarrow x = y$ (anti-symmetry)
- $x \; R \; y \wedge y \; R \; z \Rightarrow x \; R \; z$ (transitivity)
- What are S and R in Itemset Mining?

Recherche de Motifs

Poset $(2^I, \subseteq)$

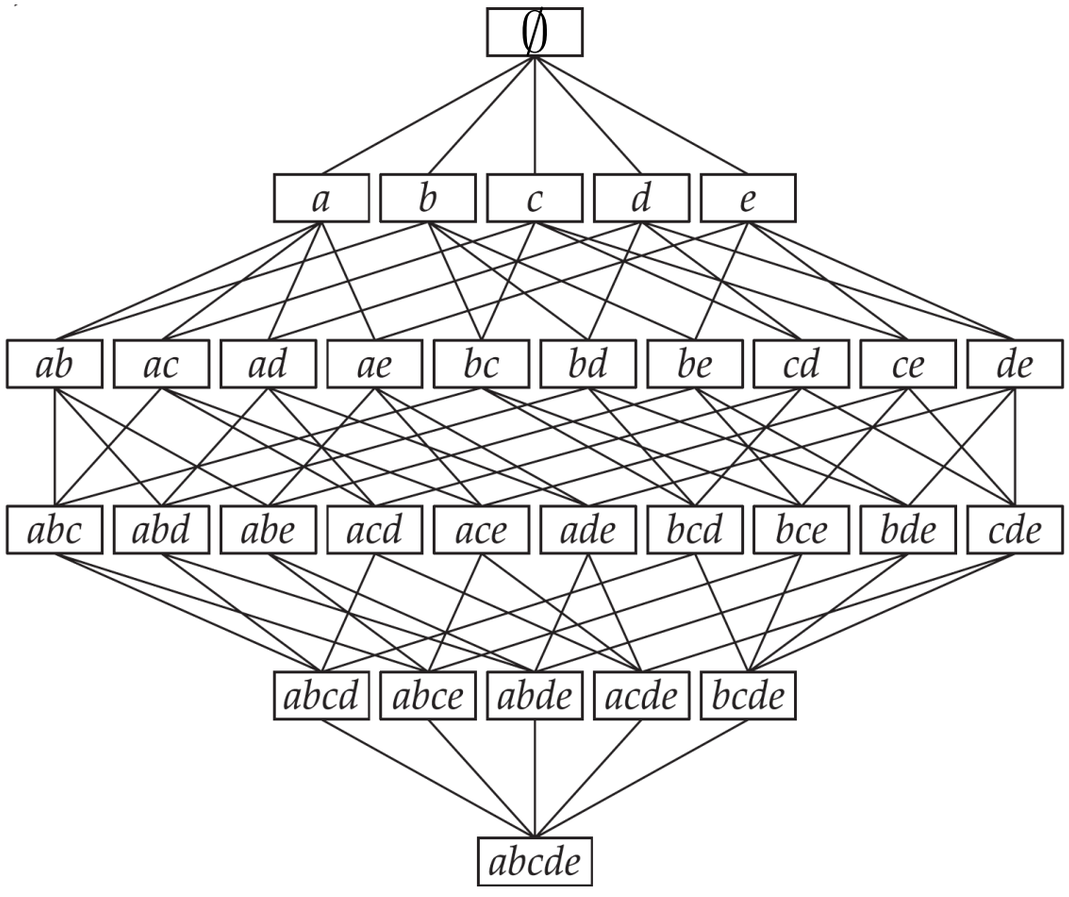
- Test A partial order is a binary relation R over a set S :
- $\forall x, y, z \in S$
- $x \; R \; x$ (reflexivity)
- $x \; R \; y \wedge y \; R \; x \Rightarrow x = y$ (anti-symmetry)
- $x \; R \; y \wedge y \; R \; z \Rightarrow x \; R \; z$ (transitivity)
- What are S and R in Itemset Mining?



Recherche de Motifs

Partially ordered sets (rappel)

- Comparable itemsets: $x \subseteq y \vee y \subseteq x$
- Incomparable itemsets: $x \not\subseteq y \wedge y \not\subseteq x$



Recherche de Motifs

Apriori Algorithm [Agrawal and Srikant 1994]

- Determine the support of the one-element item sets (i.e. singletons) and discard the infrequent items
- Form candidate itemsets with two items (both items must be frequent), determine their support, and discard the infrequent itemsets
- Form candidate item sets with three items (all contained pairs must be frequent), determine their support, and discard the infrequent itemsets
- And so on!

Recherche de Motifs

Apriori Algorithm [Agrawal and Srikant 1994]

- Determine the support of the one-element item sets (i.e. singletons) and discard the infrequent items
- Form candidate itemsets with two items (both items must be frequent), determine their support, and discard the infrequent itemsets
- Form candidate item sets with three items (all contained pairs must be frequent), determine their support, and discard the infrequent itemsets
- And so on!

Based on candidate generation and pruning

Searching for Frequent Itemsets

Apriori Algorithm [Agrawal and Srikant 1994]

Algorithm 1: Apriori Algorithm

Input: Transaction database $\mathcal{D}$, minimum support threshold α
Output: Frequent itemsets

$k \leftarrow 1$;
$L_k \leftarrow \{p_i \mid p_i \in \mathcal{I} \wedge \texttt{freq}(p_i) \geq \alpha\}$;
while $L_k \neq \emptyset$ **do**
    $C \leftarrow \texttt{aprioriGen}(L_k)$;
    $k \leftarrow k + 1$;
    $L_k \leftarrow \{c \mid c \in C \wedge \texttt{freq}(c) \geq \alpha\}$;
return $\bigcup_i L_i$;

Searching for Frequent Itemsets

Apriori Algorithm [Agrawal and Srikant 1994]

Function `aprioriGen`(L_k)**:**

$E \leftarrow \emptyset$;

for *each pair of itemsets* $P', P'' \in L_k$ *such that* $P' = \{p_{i_1}, \ldots, p_{i_{k-1}}, p_{i_k}\}$ *and* $P'' = \{p_{i_1}, \ldots, p_{i_{k-1}}, p_{i'_k}\}$ **do**

  if $p_{i_k} \neq p_{i'_k}$ **then**

    $P \leftarrow P' \cup P''$;

    if $\forall p_i \in P,\ P \setminus \{p_i\} \in L_k$ **then**

      $E \leftarrow E \cup \{P\}$;

return E ;

Représentations Condensées des Motifs : Fermés et Maximaux

Motifs Maximaux

Définition

- The set of Maximal (frequent) Itemsets:

$$M_\alpha = \{P \subset I \,|\, freq(P) \geq \alpha \wedge \forall P' \supset P : freq(P') < \alpha\}$$

- That is:

$$\forall \alpha, \forall P \in F_\alpha : (P \in M_\alpha) \vee (\exists P' \supset P : freq(P') \geq \alpha)$$

(With F_α the set of all frequent itemsets)

Motifs Maximaux

Définition

- The set of Maximal (frequent) Itemsets:

$$M_\alpha = \{P \subset I \,|\, freq(P) \geq \alpha \wedge \forall P' \supset P : freq(P') < \alpha\}$$

An itemset is considered maximal if it is frequent, and none of its proper supersets is frequent

- That is:

$$\forall \alpha, \forall P \in F_\alpha : (P \in M_\alpha) \vee (\exists P' \supset P : freq(P') \geq \alpha)$$

(With F_α the set of all frequent itemsets)

Motifs Maximaux

Définition

- Every frequent itemset has a maximal superset:

$$\forall \alpha, \forall P \in F_\alpha : (\exists P' \in M_\alpha : P \subseteq P')$$

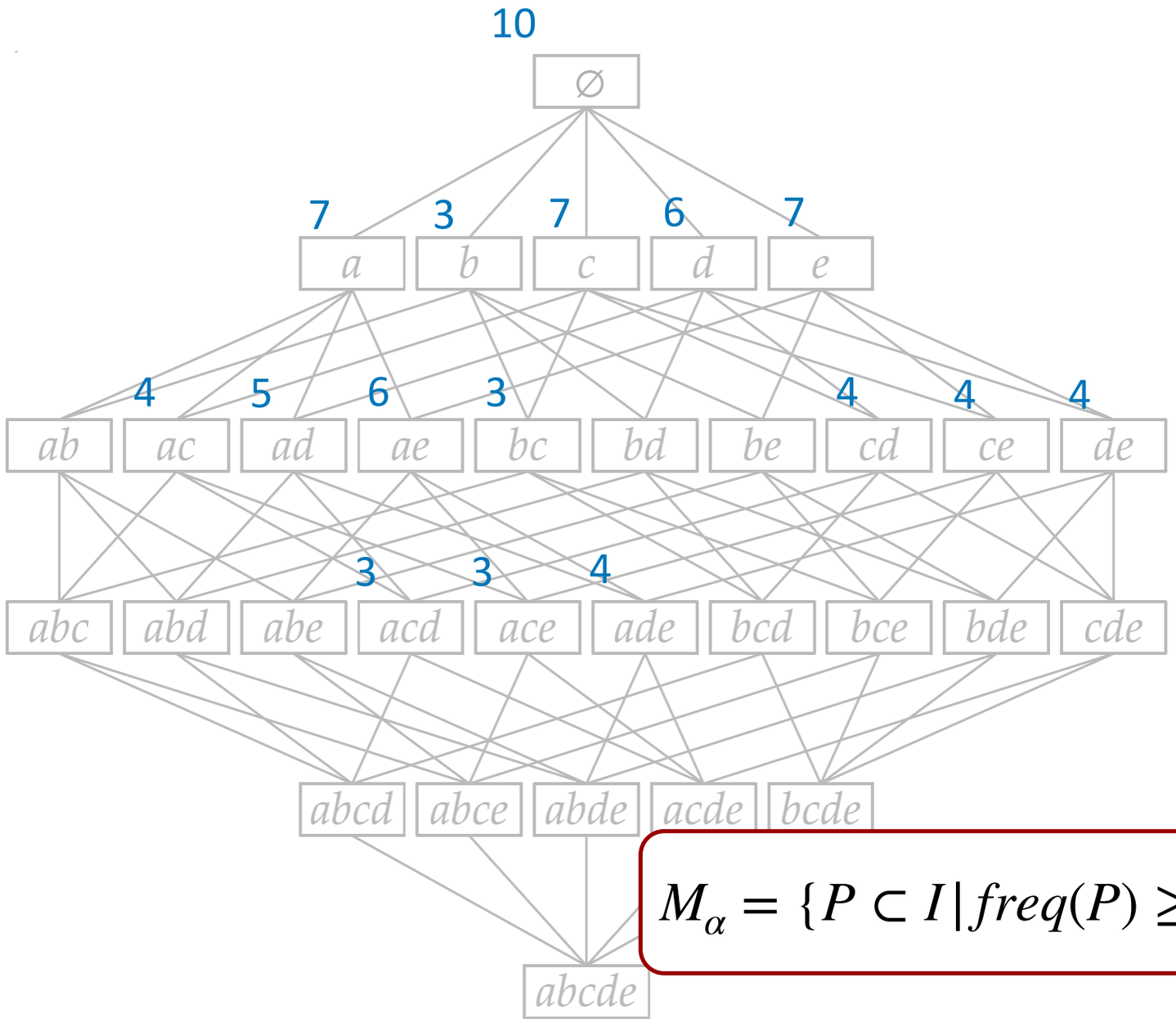
- The maximal itemsets provide a condensed representation of the frequent itemsets, where:

$$\forall \alpha : F_\alpha = \bigcup_{P \in M_\alpha} 2^P$$

Motifs Maximaux

Exemple (4)

Query-3: Maximal itemsets with $\alpha = 3$?



H_D

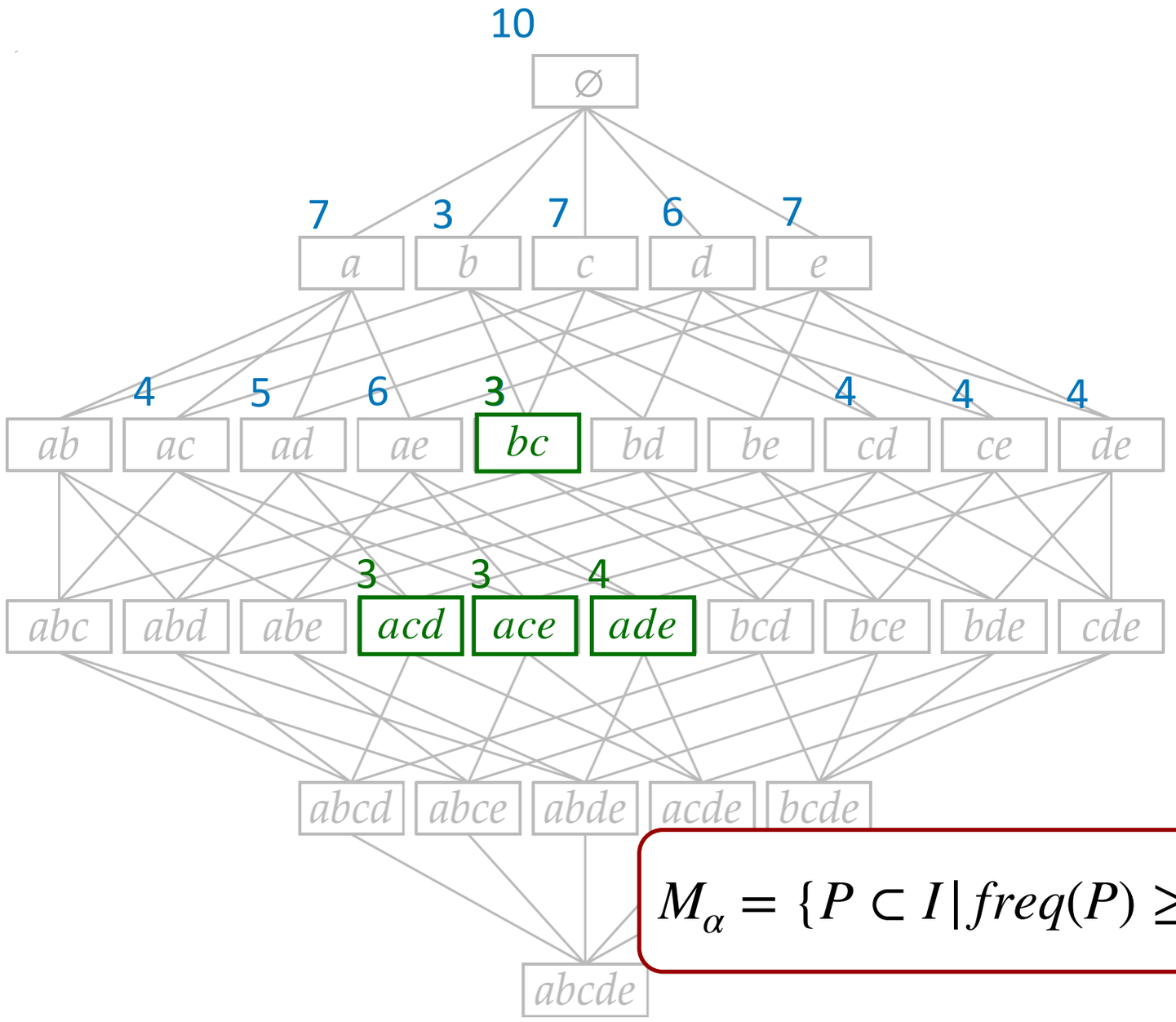
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

$$M_\alpha = \{P \subset I \,|\, freq(P) \geq \alpha \wedge \forall P' \supset P : freq(P') < \alpha\}$$

Motifs Maximaux

Exemple (4)

Query-3: Maximal itemsets with $\alpha = 3$?



H_D

1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

$$M_\alpha = \{P \subset I \,|\, freq(P) \geq \alpha \wedge \forall P' \supset P : freq(P') < \alpha\}$$

Motifs Fermés

Définition

- The set of Closed (frequent) Itemsets:

$$C_\alpha = \{P \subset I \,|\, freq(P) \geq \alpha \wedge \forall P' \supset P : freq(P') < freq(P)\}$$

An itemset is closed if it is frequent, but none of its proper supersets has the same support value

- That is:

$$\forall \alpha, \forall P \in F_\alpha : (P \in C_\alpha) \vee (\exists P' \supset P : freq(P') = freq(P))$$

Motifs Fermés

Définition

- Every frequent itemset has a closed superset:

$$\forall \alpha, \forall P \in F_\alpha : (\exists P' \in C_\alpha : P \subseteq P')$$

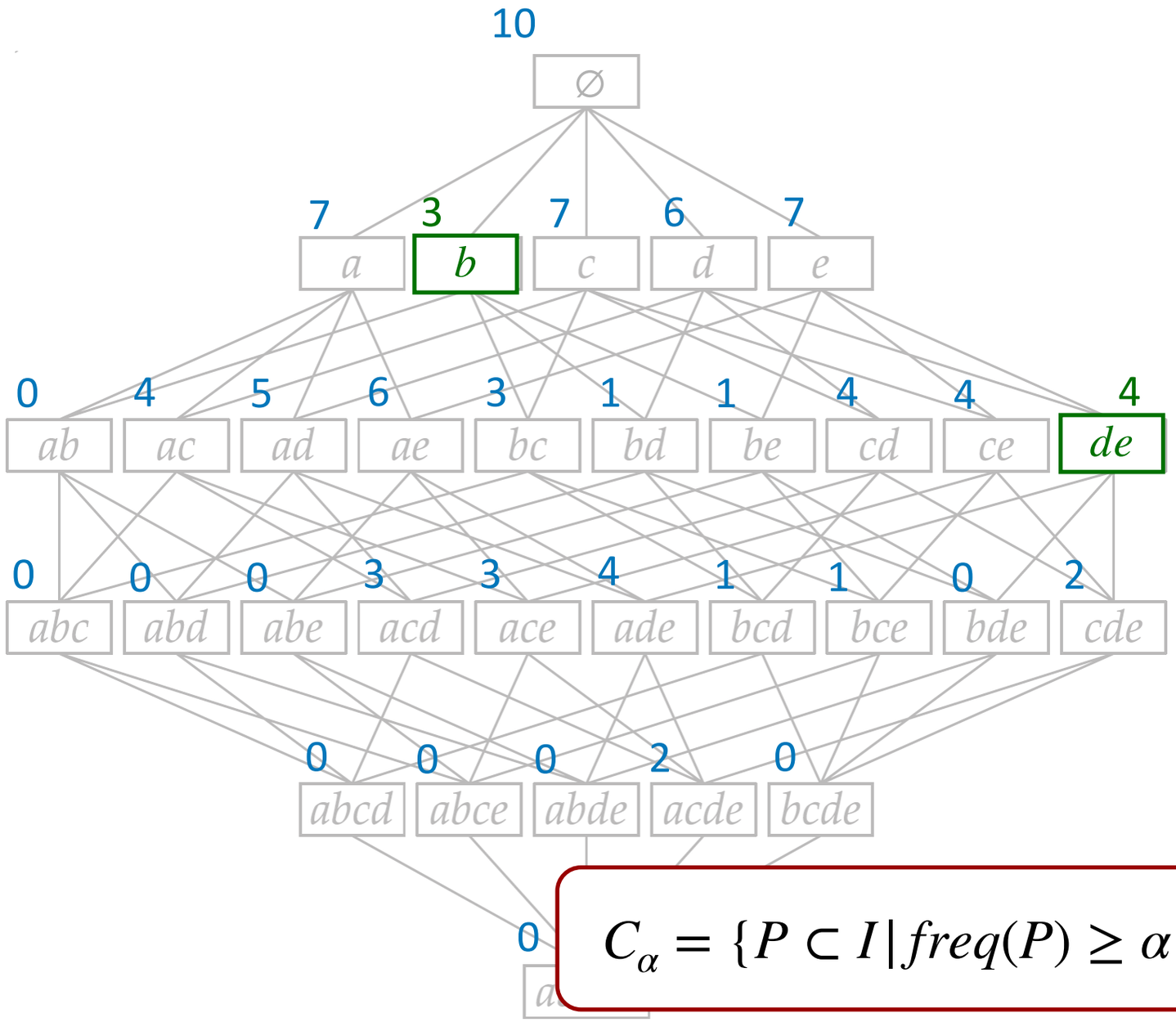
- Closed itemsets provide a condensed representation of frequent itemsets, where:

$$\forall \alpha : F_\alpha = \bigcup_{P \in C_\alpha} 2^P$$

Motifs Fermés

Exemple (5)

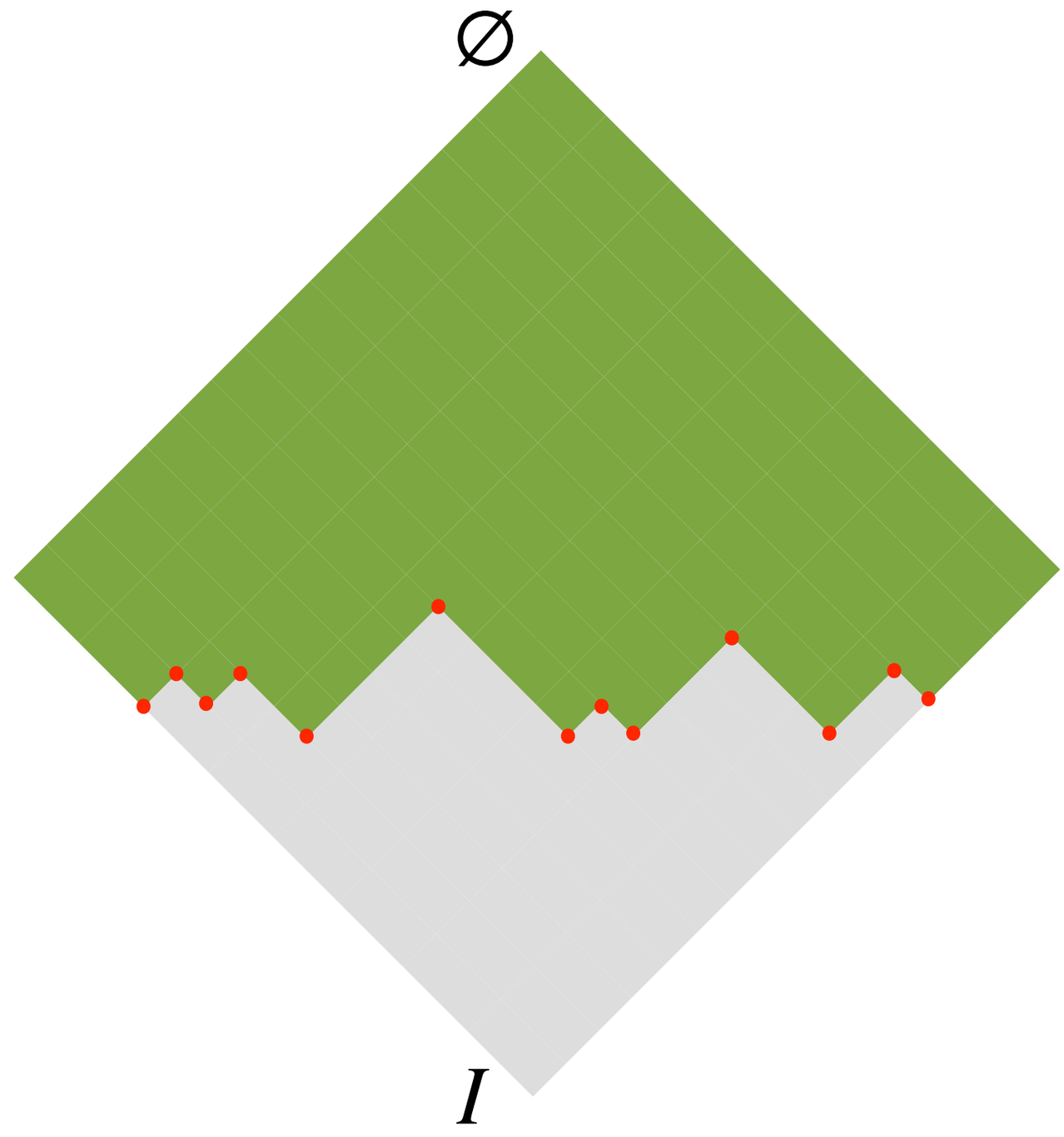
Q: Are ‘b’ and ‘de’ closed itemsets?”

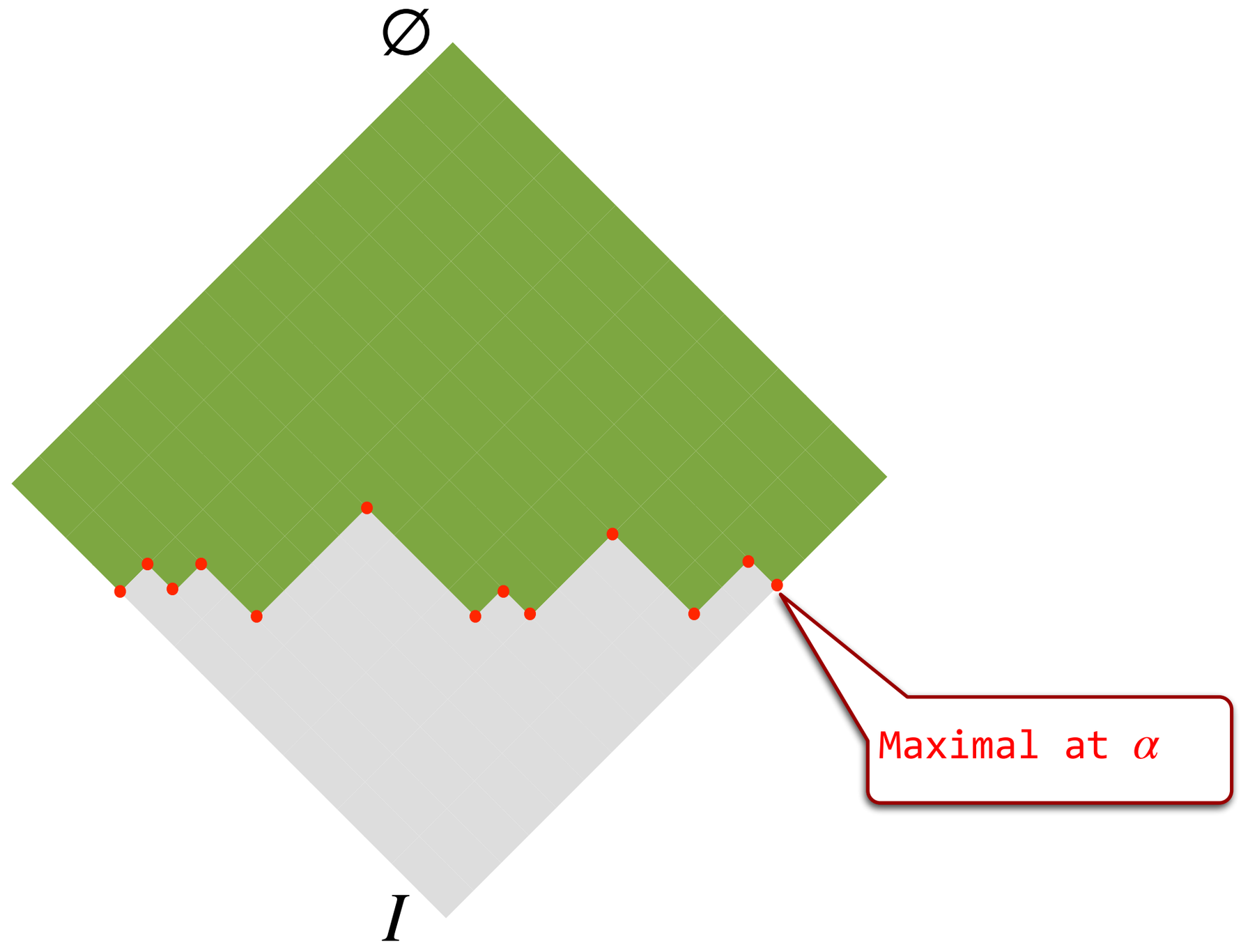


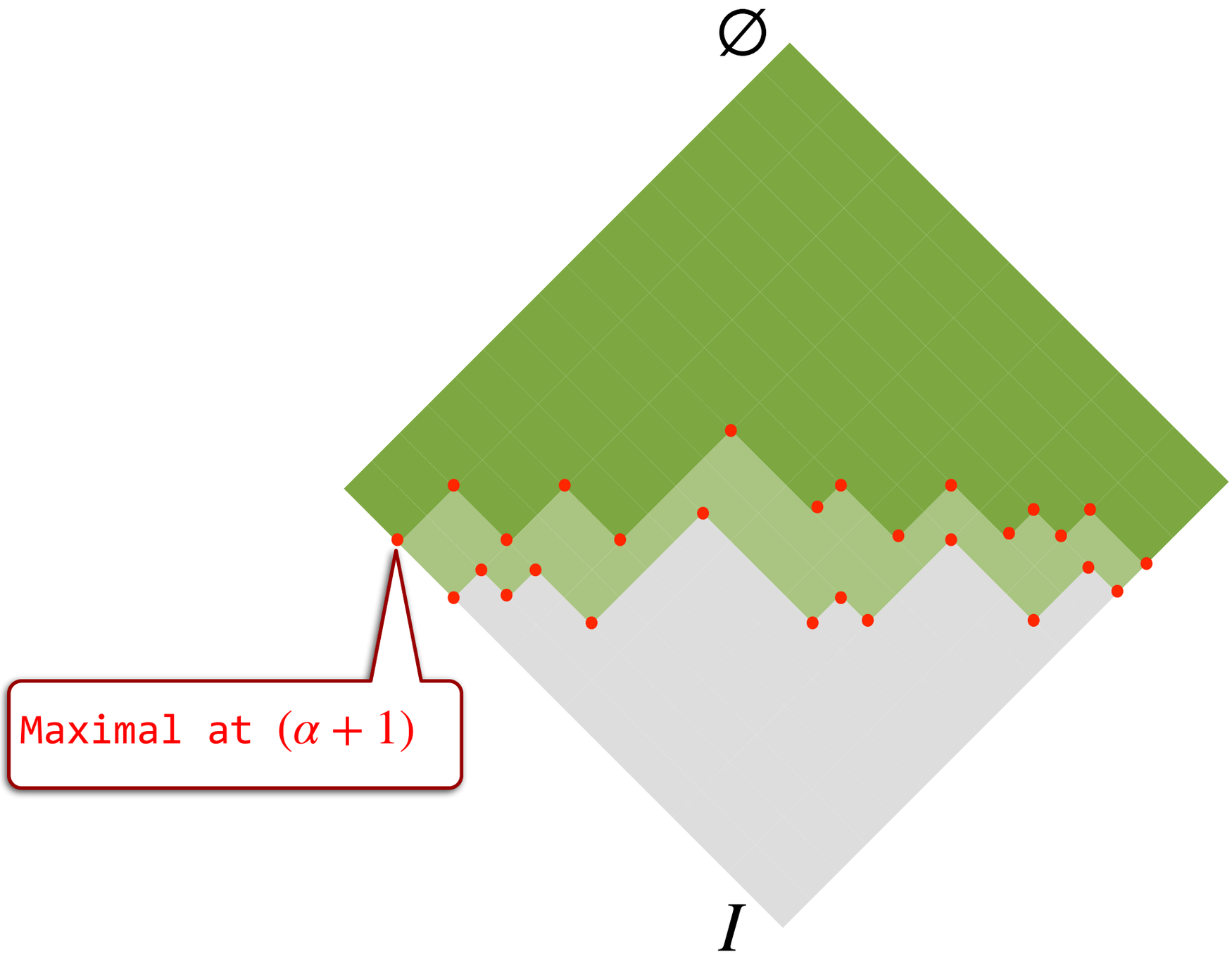
H_D

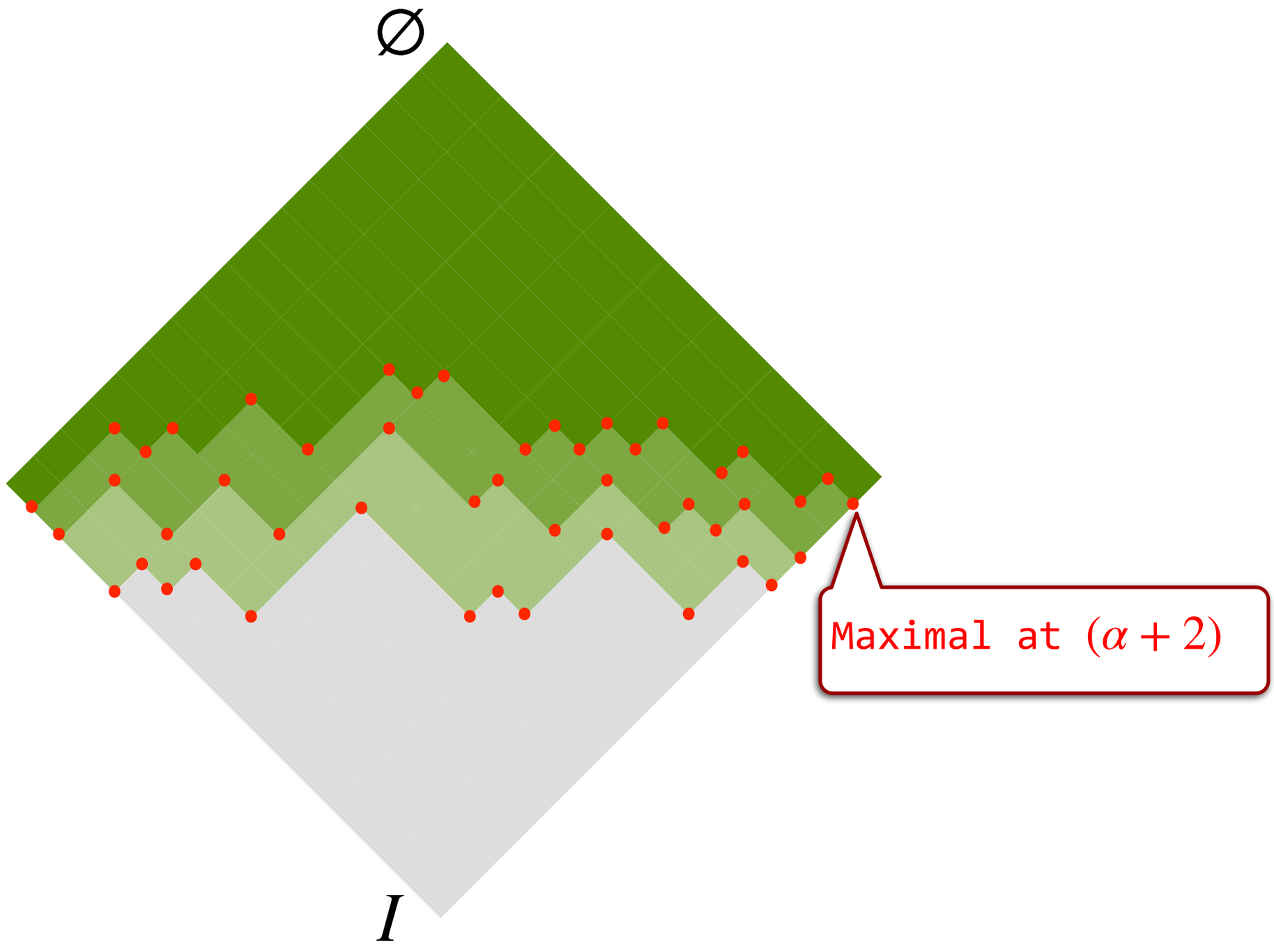
1:	a			d	e
2:		b	c	d	
3:	a		c		e
4:	a		c	d	e
5:	a				e
6:	a		c	d	
7:		b	c		
8:	a		c	d	e
9:		b	c		e
10:	a			d	E

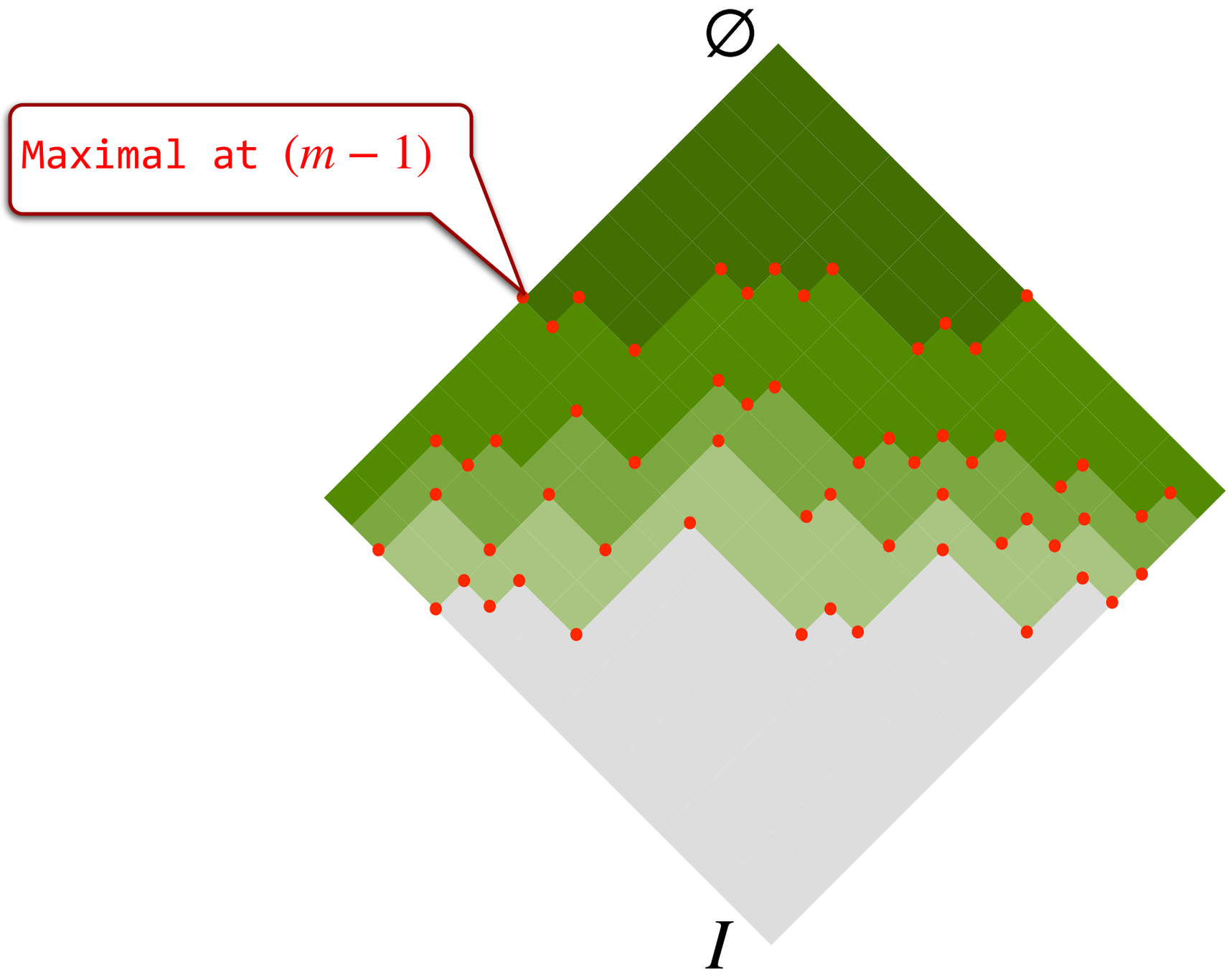
$$C_\alpha = \{P \subset I \,|\, freq(P) \geq \alpha \wedge \forall P' \supset P : freq(P') < freq(P)\}$$

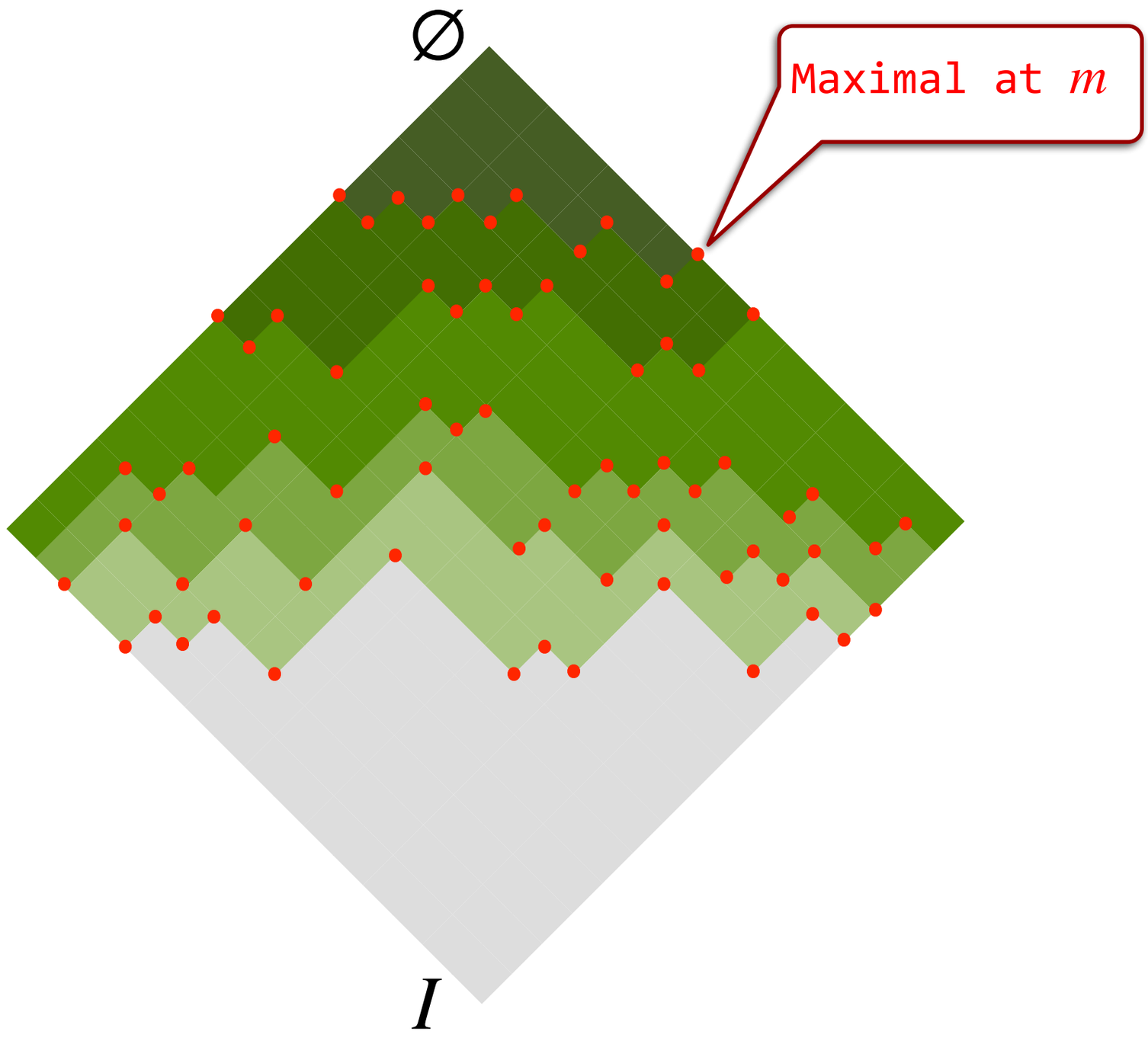

∅
I

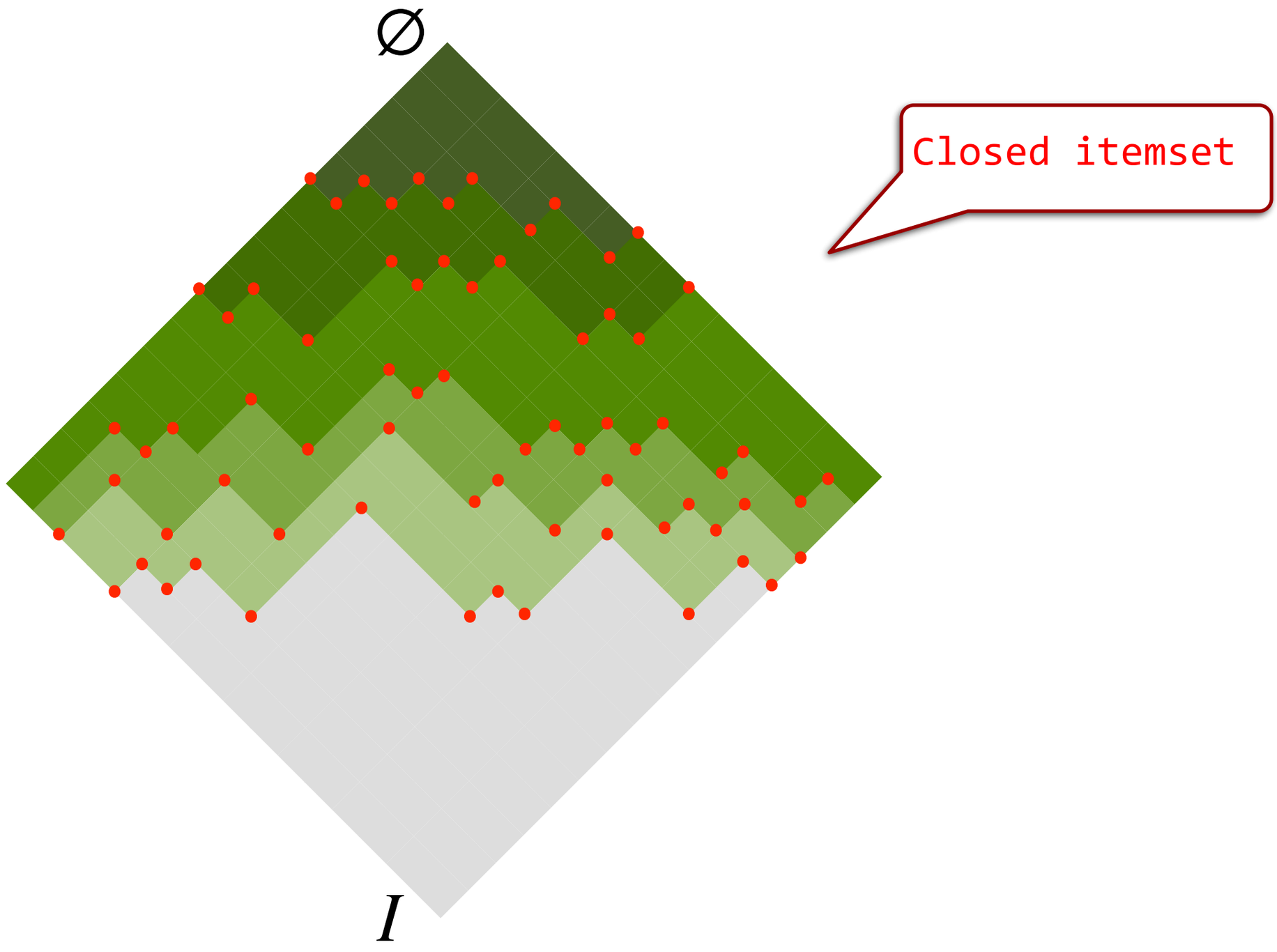

∅
Maximal at α
I

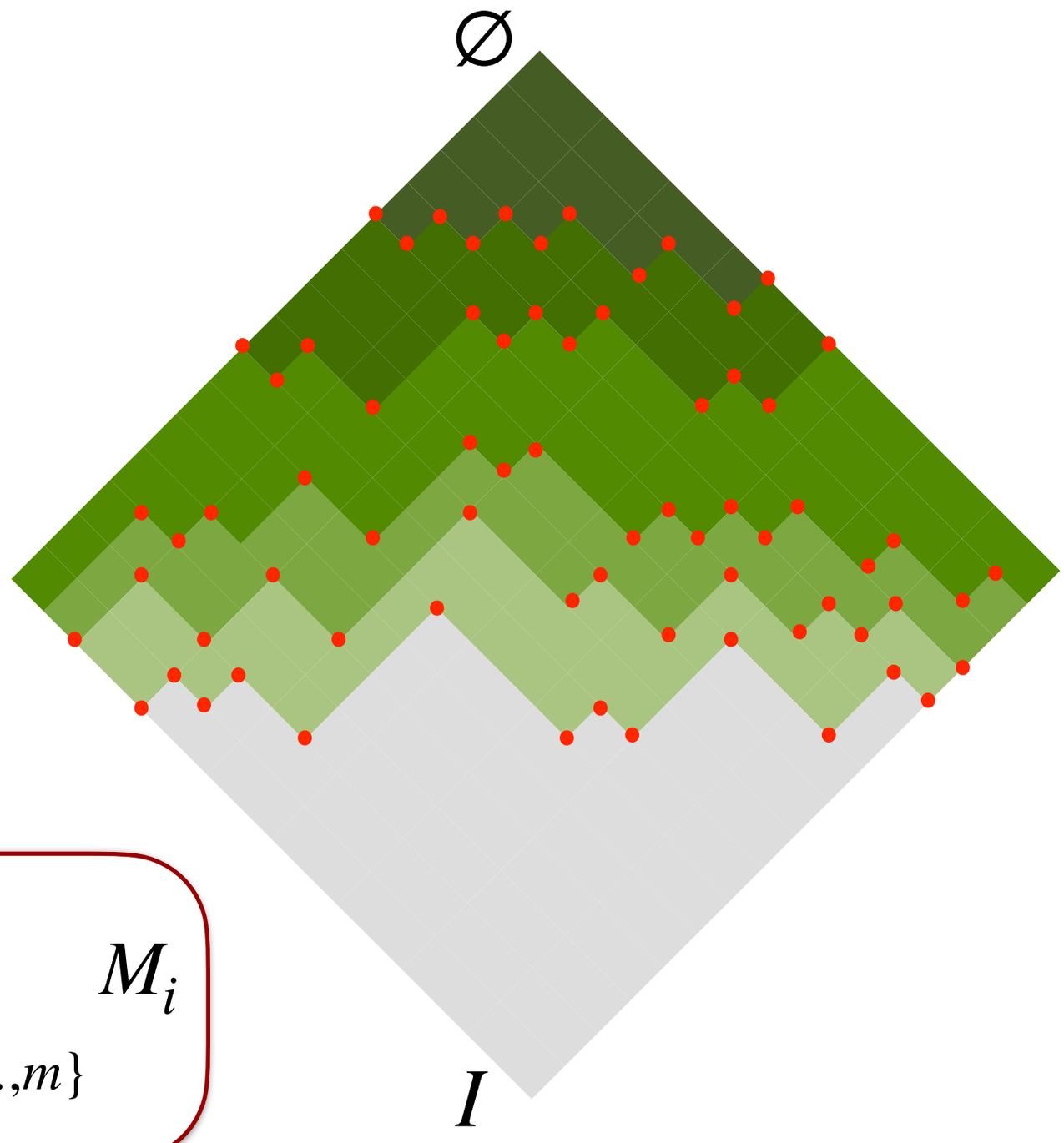

∅
Maximal at $(\alpha + 1)$
I


∅
Maximal at $(\alpha + 2)$
I


∅
Maximal at $(m-1)$
I


∅
Maximal at m
I


∅
Closed itemset
I



$$C_\alpha = \bigcup_{i \in \{\alpha,\ldots,m\}} M_i$$

Frequent/Closed/Maximal

Itemsets

Dataset	#Frequent	#Closed	#Maximal
Zoo-1	151 807	3 292	230
Mushroom	155 734	3 287	453
Lymph	9 967 402	46 802	5 191
Hepatitis	$27 . 10^{7}+$	1 827 264	189 205

http://fimi.ua.ac.be/data/

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

université PARIS-SACLAY

cnrs

Artificial Intelligence

Cours7 - Frequent Itemset Mining

L3 - Informatique

Nadjib Lazaar
Ing - Phd - HDR - Professor - Paris-Saclay University - LISN - LaHDAK
lazaar@lisn.fr https://perso.lisn.upsaclay.fr/lazaar/

14/03/2025