A Distributed Algorithm for Spectral Sparsification of Graphs with Applications to Data Clustering

Fabricio Mendoza-Granada, Marcos Villagra

Facultad Politécnica
Núcleo de Investigación y Desarrollo Tecnológico
Universidad Nacional de Asunción

September 16, 2020

Outline

Outline for Section 1

Graph Laplacian

We will work with two fundamental definitions from *Spectral Graph Theory*

Definition 1

Given an undirected weighted graph $G = (V, E, w)$ where $w : E \to \mathbb{R}_{\geq 0}$ the Laplacian matrix is defined as

$$L_G = D_G - A_G.$$

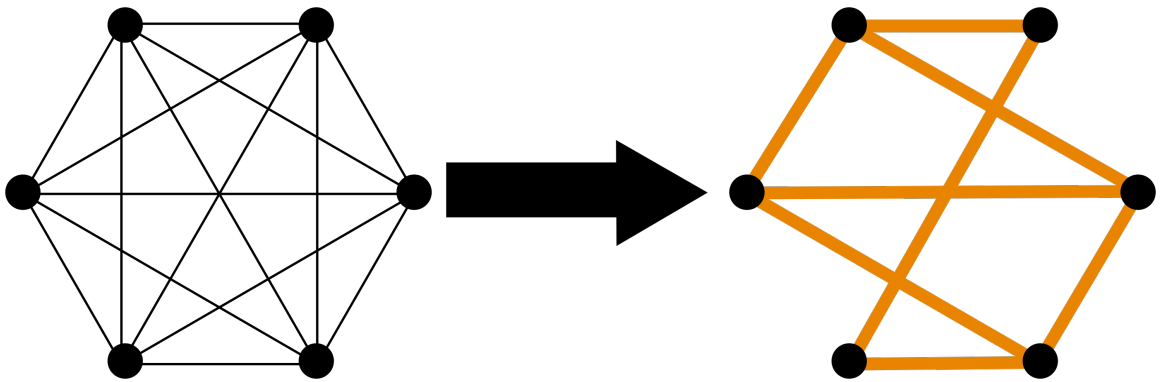
where D_G is the weighted degree matrix and A_G is the weighted adjacency matrix.

Outline for Section 1

Spectral Sparsification Problem

Given an input graph G, an spectral sparsifier of G is a subgraph H with two major characteristics:

1. H has fewer edges than G
2. The spectra of G and H are close up to a constant factor



The spectra of a graph G is the set of eigenvalues of L_G.

Spectral Sparsification Definition

Definition 2

Let $H = (V, E, \tilde{w})$ be a subgraph of G. We say that H is an ϵ-spectral sparsifier of G if

$$(1-\epsilon)x^T L_G x \leq x^T L_H x \leq (1+\epsilon)x^T L_G x. \tag{1}$$

Outline for Section 2

Problem Resolution Approach

- First, we will present a structure that captures the idea of repeated elements along a family of subsets. The set theoretic structure is composed of three parts which are
  - Occurrence Number
  - Overlapping Cardinality
  - Overlapping Cardinality Partition
- Second, we will use the overlapping cardinality partition for decomposing the sum of Laplacians of a family of subgraphs.
- Finally, we will show that the union of spectral sparsifiers of those subgraphs is an spectral sparsifier of their union.

Occurrence number

Definition 3

Let $\{A_1, \ldots, A_t\}$ be a family of subsets of A. For any $a \in A$, the occurrence number of a in $\{A_i\}_{i \leq t}$, denoted $\#(a)$, is the maximum number of sets from $\{A_i\}_{i \leq t}$ in which a appears.

Occurrence number

Definition 3

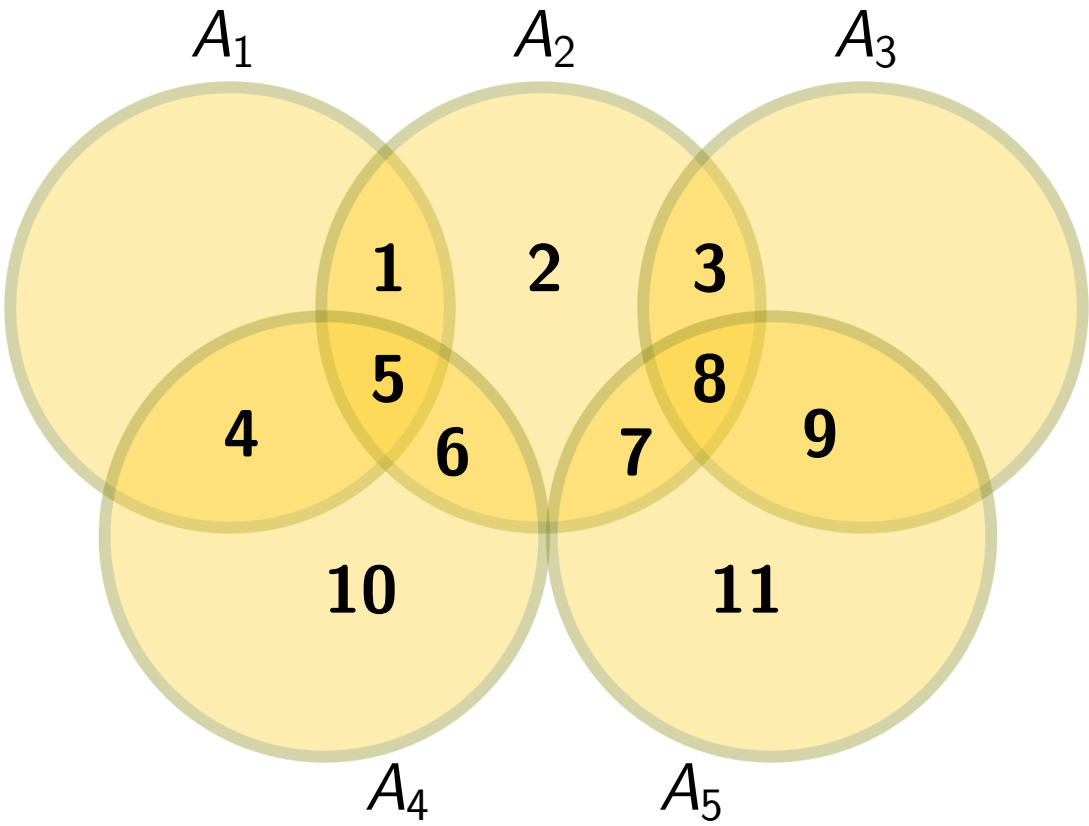
Let $\{A_1, \ldots, A_t\}$ be a family of subsets of A. For any $a \in A$, the occurrence number of a in $\{A_i\}_{i \leq t}$, denoted $\#(a)$, is the maximum number of sets from $\{A_i\}_{i \leq t}$ in which a appears.

Example 4

Consider the following family of subsets

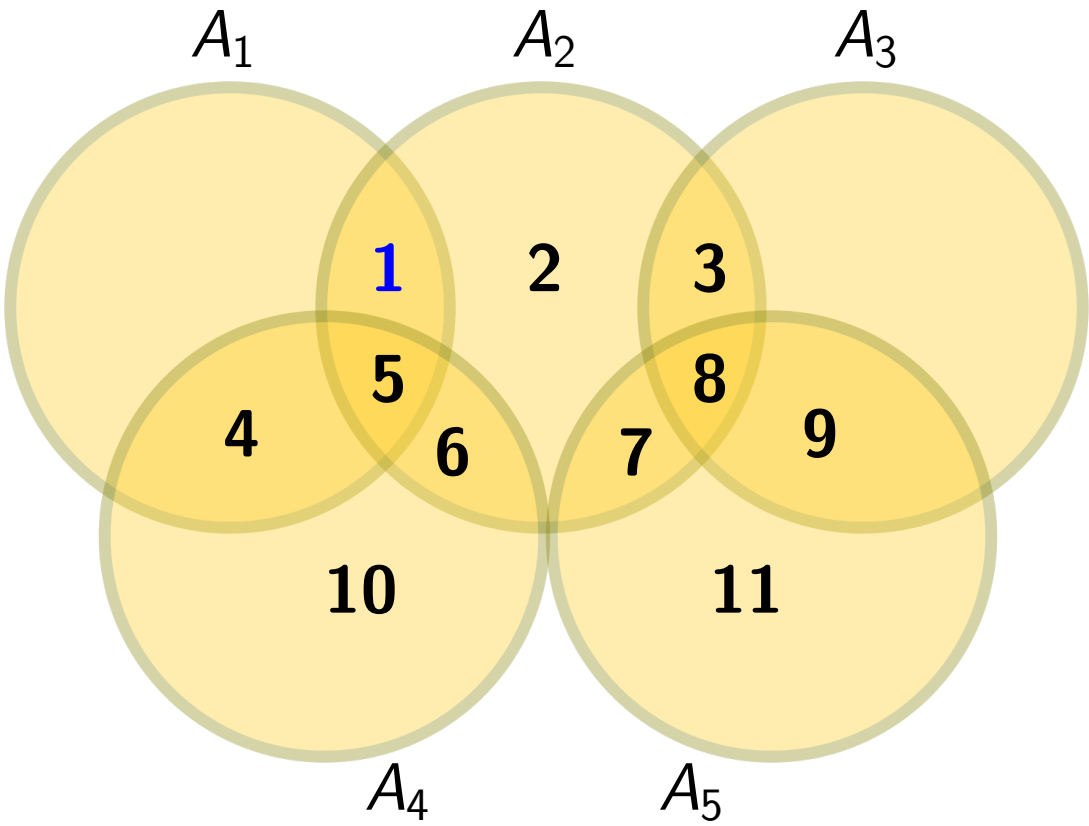
$$\{\{1,4,5\},\{1,2,3,5,6,7,8\},\{3,7,8,9\},\{4,5,6,10\},\{7,8,9,11\}\}.$$

Occurrence Number



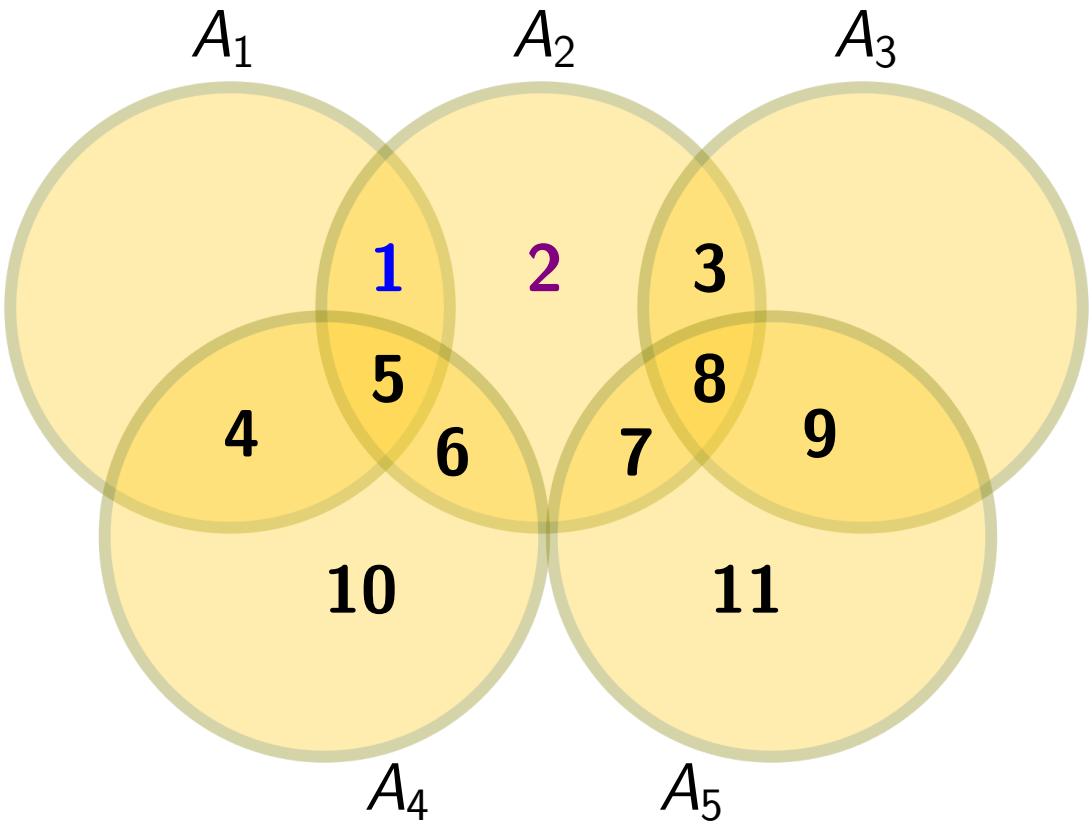
Here we have that $\#(1) = 2$, $\#(2) = 1$, $\#(3) = 2$, and so on.

Occurrence Number



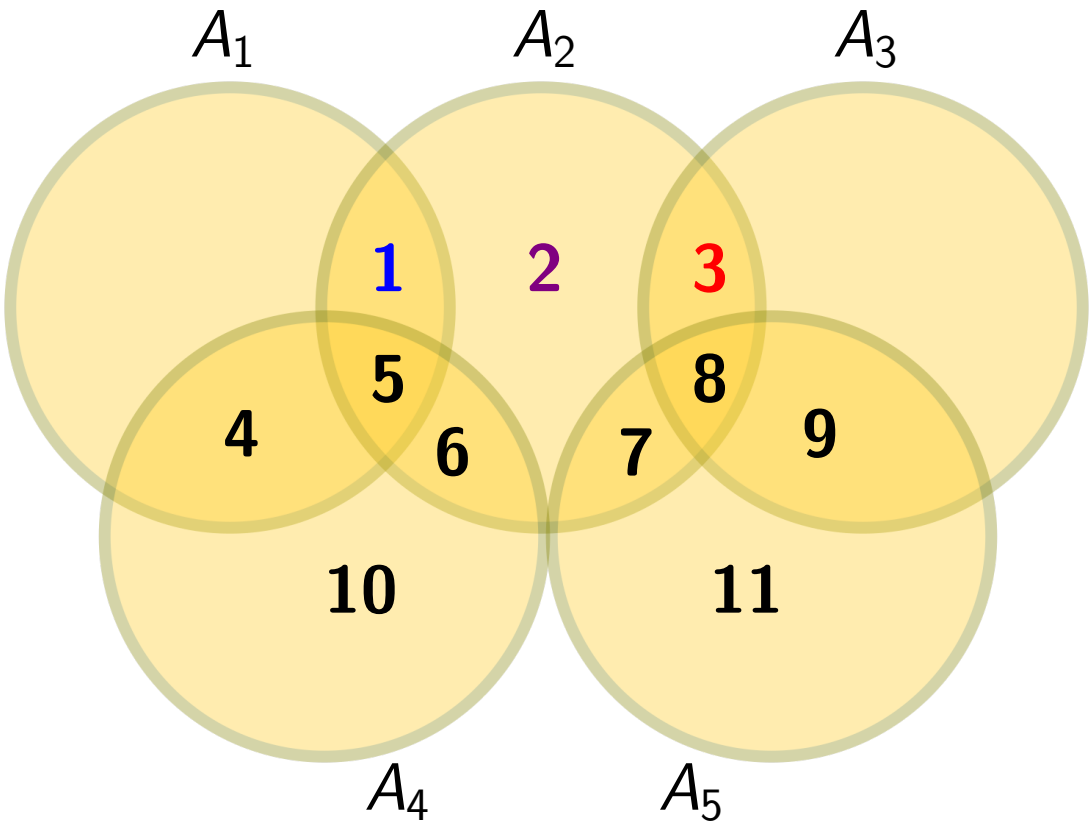
Here we have that $\#(1) = 2$, $\#(2) = 1$, $\#(3) = 2$, and so on.

Occurrence Number



Here we have that $\#(1) = 2$, $\#(2) = 1$, $\#(3) = 2$, and so on.

Occurrence Number



Here we have that $\#(1) = 2$, $\#(2) = 1$, $\#(3) = 2$, and so on.

Overlapping Cardinality Definition

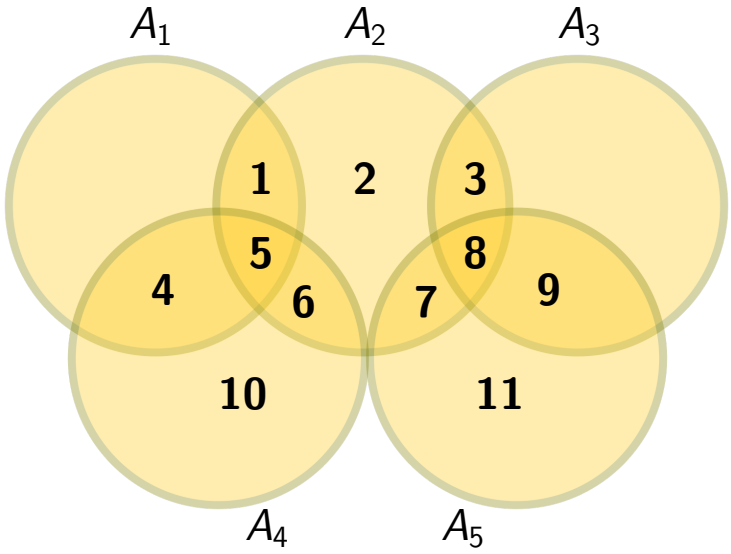
Definition 5

Let $\{A_1, \ldots, A_t\}$ be a family of subsets of $A = \bigcup_{i=1}^{t} A_i$. The overlapping cardinality of a subset $A' \subseteq A$ in $\{A_i\}_{i \leq t}$ is a positive integer c such that for each $a \in A'$ its occurrence number $\#(a) = c$; otherwise the overlapping cardinality of A' in $\{A_i\}_{i \leq t}$ is 0.

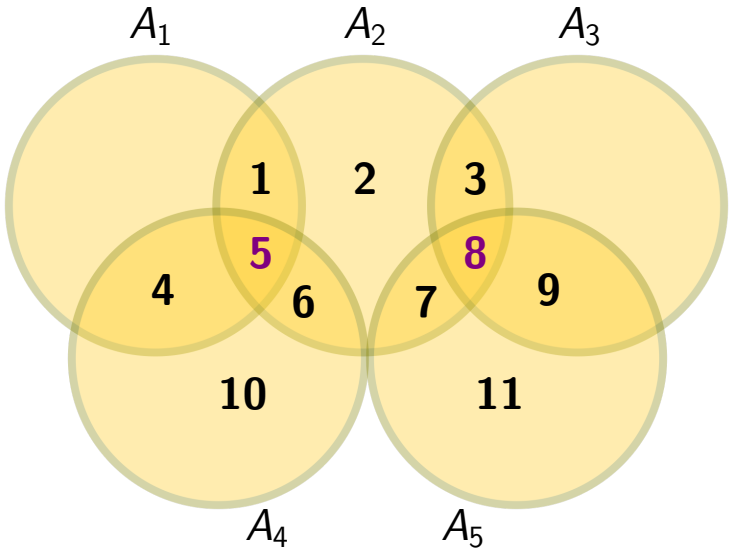
Observation 1

The overlapping cardinality of a given subset of A is a positive value c if and only if all its elements has the same occurrence number c

Overlapping Cardinality



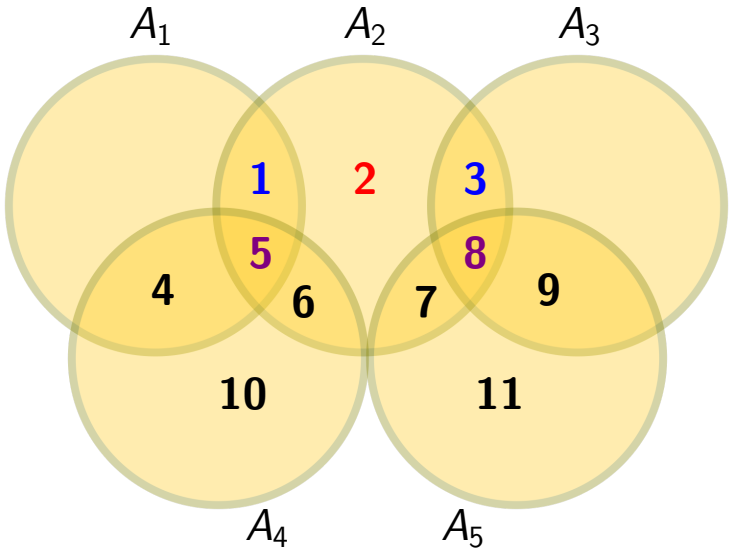
Overlapping Cardinality



Example 6

Take the subset $\{5, 8\}$, its overlapping cardinality is 3 because $\#5 = \#8 = 3$.

Overlapping Cardinality



Example 6

Take the subset $\{5, 8\}$, its overlapping cardinality is 3 because $\#5 = \#8 = 3$.

Example 7

Now, take $\{1, 2, 3\}$, its overlapping cardinality is 0 because $\#1 = \#3 = 2$ but $\#2 = 1$.

Overlapping Cardinality Partition

Overlapping Cardinality Partition

Definition 8

It is a way to partition a set A respect to the number of times each element $a \in A$ appears in a family $\{A_i\}_{i \leq t}$ where $t \in \mathbb{N}$ and $A_i \subseteq A$.

Example 9

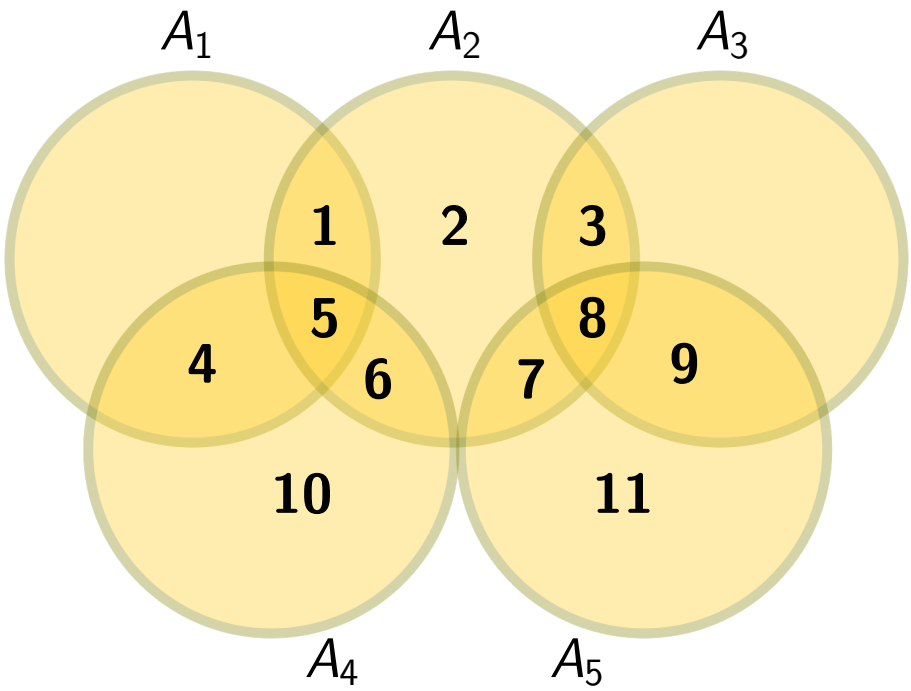
Let's take again the family

$$\{\{1,4,5\},\{1,2,3,5,6,7,8\},\{3,7,8,9\},\{4,5,6,10\},\{7,8,9,11\}\}.$$

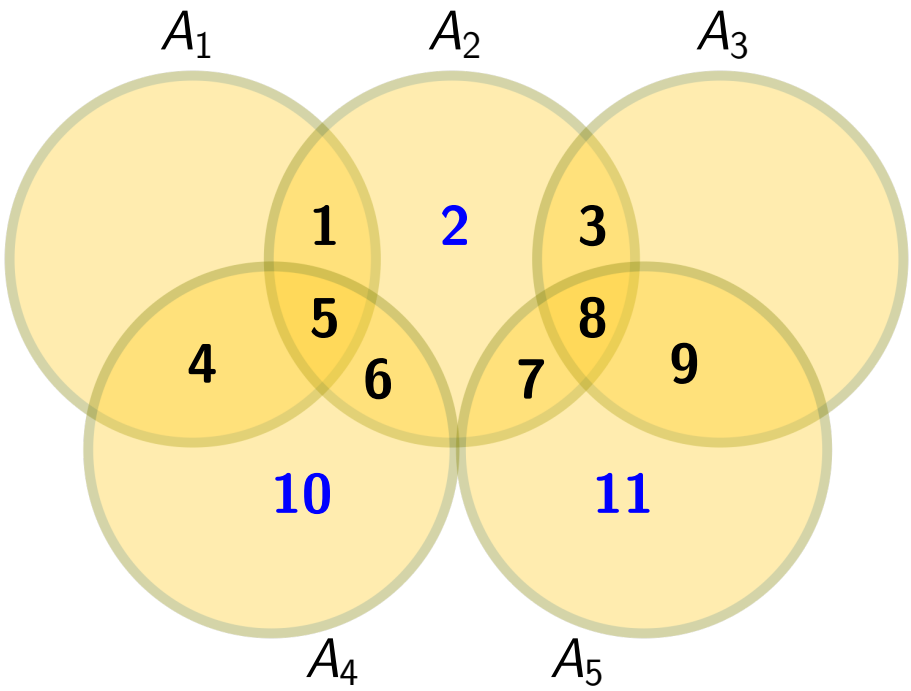
An overlapping cardinality partition is

$$\{\{2,10,11\},\{1,3,4,6,7,9\},\{5,8\}\}.$$

Overlapping Cardinality Partition

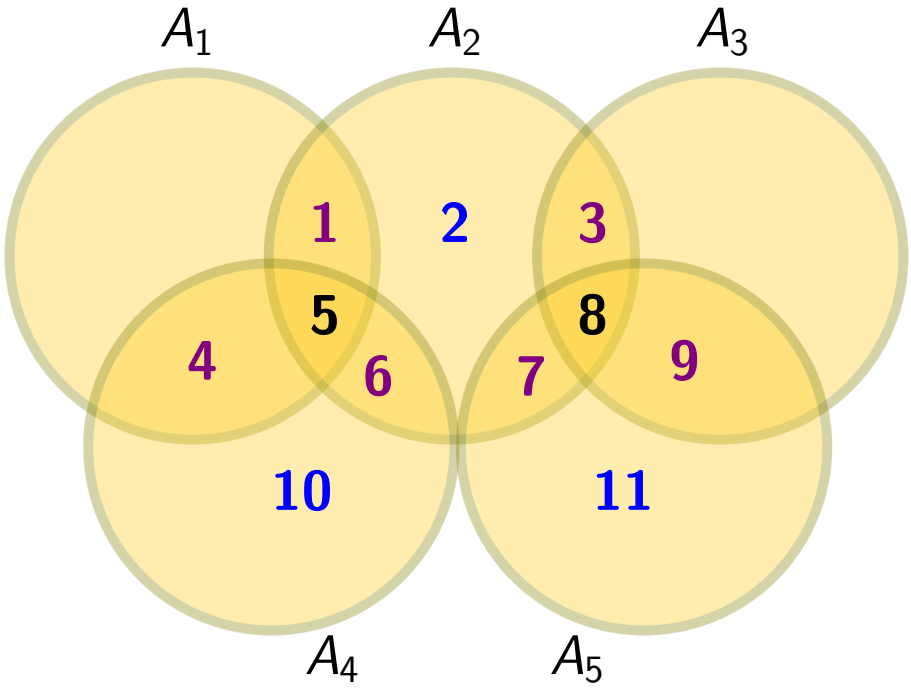


Overlapping Cardinality Partition



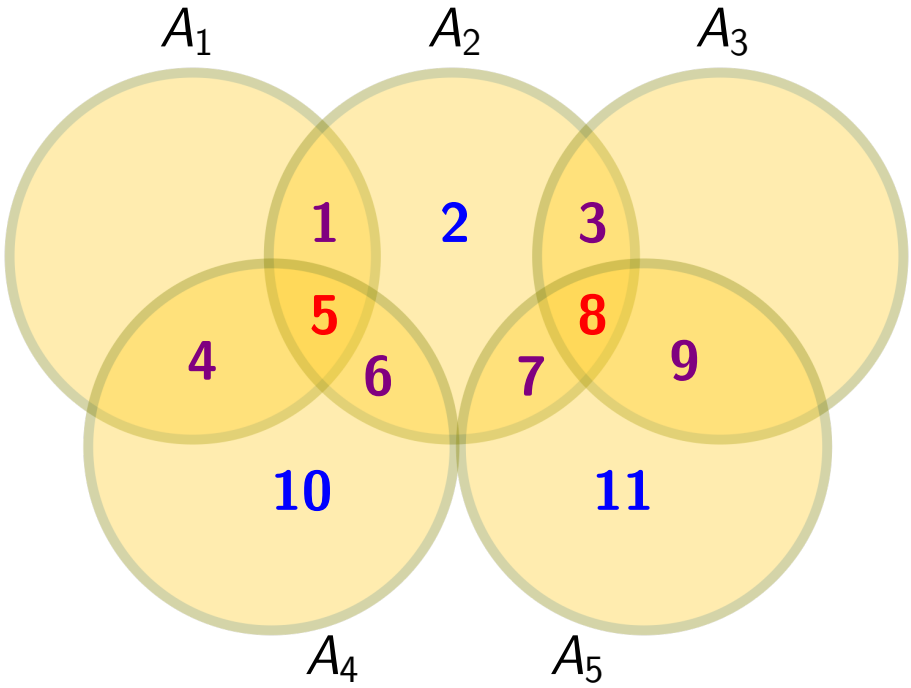
- $\{2, 10, 11\}$ has overlapping cardinality equals to 1 because $\#2 = \#10 = \#11 = 1$.

Overlapping Cardinality Partition



- $\{2, 10, 11\}$ has overlapping cardinality equals to 1 because $\#2 = \#10 = \#11 = 1$.
- $\{1, 3, 4, 6, 7, 9\}$ has overlapping cardinality equals to 2 because $\#1 = \#3 = \#4 = \#6 = \#7 = \#9 = 2$.

Overlapping Cardinality Partition



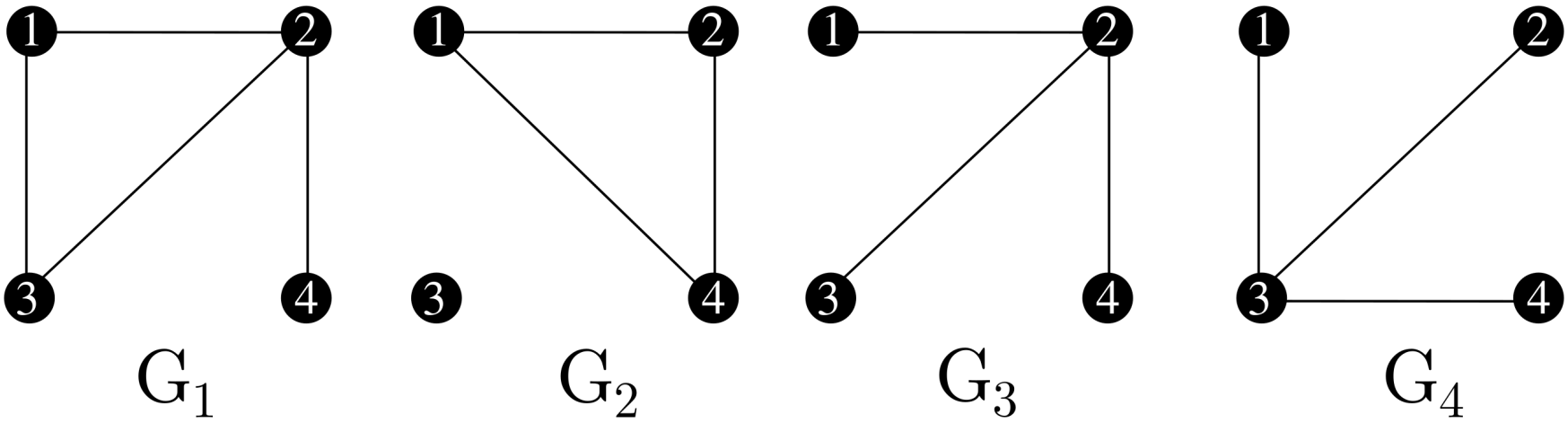
- $\{2, 10, 11\}$ has overlapping cardinality equals to 1 because $\#2 = \#10 = \#11 = 1$.
- $\{1, 3, 4, 6, 7, 9\}$ has overlapping cardinality equals to 2 because $\#1 = \#3 = \#4 = \#6 = \#7 = \#9 = 2$.
- $\{5, 8\}$ has overlapping cardinality equals to 3 because $\#5 = \#8 = 3$.

Outline for Section 2

Laplacian Decomposition

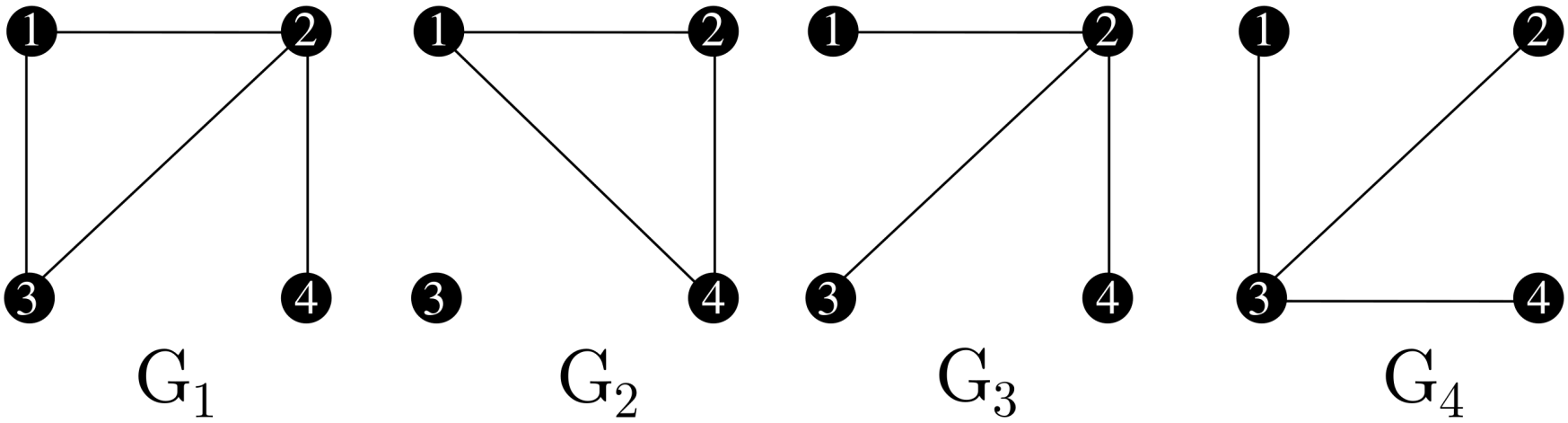
Laplacian Decomposition

Suppose a complete graph G of 4 vertices and consider four different subgraphs of it.



Laplacian Decomposition

Suppose a complete graph G of 4 vertices and consider four different subgraphs of it.



We will consider the union of graphs as a sum of Laplacians for technical reasons.

Laplacian Decomposition

Now, applying Definition 1 in every subgraph and adding them up we get $L_{G_1} + L_{G_2} + L_{G_3} + L_{G_4}$ equals

$$\begin{pmatrix} d_1^1 & -w_{12} & -w_{13} & 0 \\ -w_{12} & d_2^1 & -w_{23} & -w_{24} \\ -w_{13} & -w_{23} & d_3^1 & 0 \\ 0 & -w_{24} & 0 & d_4^1 \end{pmatrix} + \begin{pmatrix} d_1^2 & -w_{12} & 0 & -w_{14} \\ -w_{12} & d_2^2 & 0 & -w_{24} \\ 0 & 0 & d_3^2 & 0 \\ -w_{14} & -w_{24} & 0 & d_4^2 \end{pmatrix} +$$

$$\begin{pmatrix} d_1^3 & -w_{12} & 0 & 0 \\ -w_{12} & d_2^3 & -w_{23} & -w_{24} \\ 0 & -w_{23} & d_3^3 & 0 \\ 0 & -w_{24} & 0 & d_4^3 \end{pmatrix} + \begin{pmatrix} d_1^4 & 0 & -w_{13} & 0 \\ 0 & d_2^4 & -w_{23} & 0 \\ -w_{13} & -w_{23} & d_3^4 & -w_{34} \\ 0 & 0 & -w_{34} & d_4^4 \end{pmatrix}$$

where the supraindex i in d_v indicate the corresponding subgraph.

Sum of Laplacians

So we can express the sum of Laplacians as

$$\begin{pmatrix} d_1^{1'} & -3w_{12} & 0 & 0 \\ -3w_{12} & d_2^{1'} & -3w_{23} & -3w_{24} \\ 0 & -3w_{23} & d_3^{1'} & 0 \\ 0 & -3w_{24} & 0 & d_4^{1'} \end{pmatrix} + \begin{pmatrix} d_1^{2'} & 0 & -2w_{13} & 0 \\ 0 & d_2^{2'} & 0 & 0 \\ -2w_{13} & 0 & d_3^{2'} & 0 \\ 0 & 0 & 0 & d_4^{2'} \end{pmatrix}$$

$$+ \begin{pmatrix} d_1^{3'} & 0 & 0 & -w_{14} \\ 0 & d_2^{3'} & 0 & 0 \\ 0 & 0 & d_3^{3'} & -w_{34} \\ -w_{14} & 0 & -w_{34} & d_4^{3'} \end{pmatrix} = 3L_{G_1'} + 2L_{G_2'} + L_{G_3'}.$$

Where the coefficients are the overlapping cardinalities of each subset of edges.

Laplacian Decomposition Theorem

The generalization of the last example can be summarized in the following result

Theorem

If $1 \leq c_1 < c_2 < \cdots < c_k$ are the overlapping cardinalities over the family $\mathcal{E} = \{E_i\}_{i \leq t}$ with an overlapping cardinality partition $\{E'_{c_j}\}_{j \leq k}$, then $\sum_{i=1}^{t} L_{G_i} = \sum_{j=1}^{k} c_j L_{G'_{c_j}}$ where $L_{G'_{c_j}}$ is the Laplacian of $G'_{c_j} = (V, E'_{c_j}, w'_{c_j})$.

The sum of Laplacians of subgraphs equals to the sum of Laplacians of subgraphs induced by the overlapping cardinality partition times the overlapping cardinality of its associated set

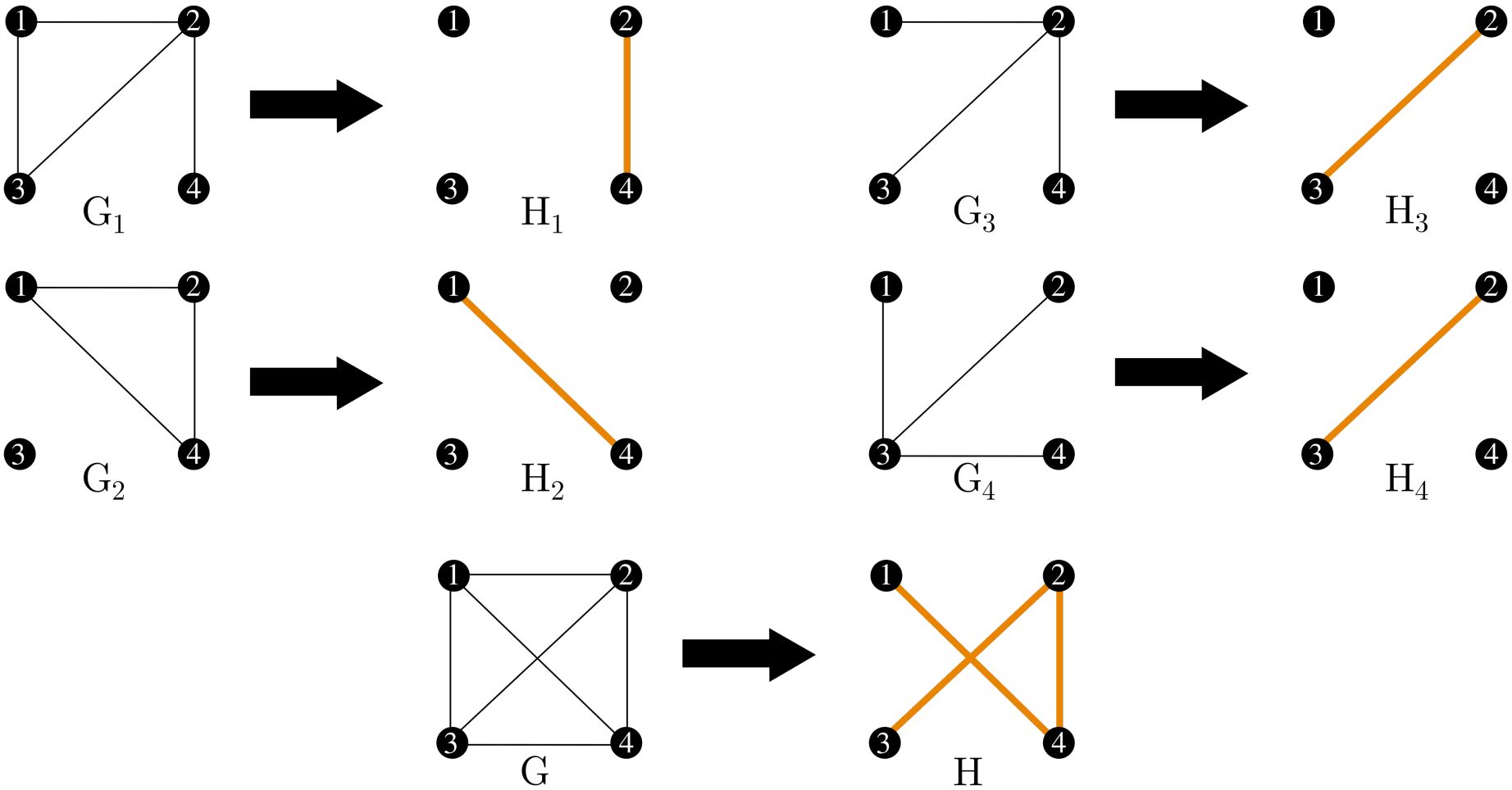
Union of Spectral Sparsifiers

Union of Spectral Sparsifiers

Theorem

Let $(1 = c_1 < c_2 < \cdots < c_k)$ *be the overlapping cardinalities over the family* $\mathcal{E} = \{E_i\}_{i \leq t}$ *with* $\{E'_{c_j}\}_{j \leq k}$ *its associated overlapping cardinality partition and* $L_{G_1}, \ldots, L_{G_t}$ *the Laplacians of* $G_1, \ldots, G_t$. *If* $H_i = (V, D_i, h_i)$ *is an* ϵ*-spectral sparsifier of* G_i, *then* $H = (V, \bigcup_i^t D_i, h)$ *is an* ϵ'*-spectral sparsifier of* G *where* $h(e) = \frac{\sum_i^t h_i(e)}{c_1 c_k}$ *and* $\epsilon' \geq 1 - \frac{1-\epsilon}{c_k}$.

Union of Spectral Sparsifiers



Outline of the proof

Outline of the proof

First, for pair of graphs G_i and H_i, by definition of Spectral Sparsifier it holds

$$(1-\epsilon)x^T L_{G_i} x \leq x^T L_{H_i} x \leq (1+\epsilon)x^T L_{G_i} x. \tag{2}$$

Outline of the proof

First, for pair of graphs G_i and H_i, by definition of Spectral Sparsifier it holds

$$(1-\epsilon)x^T L_{G_i} x \leq x^T L_{H_i} x \leq (1+\epsilon)x^T L_{G_i} x. \tag{2}$$

Then, taking the sum over all sparsifier we get

$$(1-\epsilon)\sum_{i=1}^{t} x^T L_{G_i} x \leq \sum_{i=1}^{t} x^T L_{H_i} x \leq (1+\epsilon)\sum_{i=1}^{t} x^T L_{G_i} x. \tag{3}$$

Here is where we simulate the union of graphs by the sum of Laplacians.

Outline of the proof

First, for pair of graphs G_i and H_i, by definition of Spectral Sparsifier it holds

$$(1-\epsilon)x^T L_{G_i} x \leq x^T L_{H_i} x \leq (1+\epsilon)x^T L_{G_i} x. \tag{2}$$

Then, taking the sum over all sparsifier we get

$$(1-\epsilon)\sum_{i=1}^{t} x^T L_{G_i} x \leq \sum_{i=1}^{t} x^T L_{H_i} x \leq (1+\epsilon)\sum_{i=1}^{t} x^T L_{G_i} x. \tag{3}$$

Here is where we simulate the union of graphs by the sum of Laplacians.

Now, applying Laplacian Decomposition theorem on the left and right terms of the last inequality we may conclude that

$$(1-\epsilon)\sum_{i=1}^{t} x^T L_{G_i} x \geq (1-\epsilon)c_1 x^T L_G x \text{ and} \tag{4}$$

$$(1+\epsilon)\sum_{i=1}^{t} x^T L_{G_i} x \leq (1+\epsilon)c_k x^T L_G x \tag{5}$$

Outline of the proof (cont'd)

Outline of the proof (cont'd)

Then, the resultant inequality is

$$(1-\epsilon)c_1 x^T L_G x \leq \sum_{i=1}^{t} x^T L_{H_i} x \leq (1+\epsilon)c_k x^T L_G x, \tag{6}$$

Outline of the proof (cont'd)

Then, the resultant inequality is

$$(1-\epsilon)c_1 x^T L_G x \leq \sum_{i=1}^{t} x^T L_{H_i} x \leq (1+\epsilon)c_k x^T L_G x, \tag{6}$$

and multiplying it by $\frac{1}{c_1 c_k}$ we get

$$(1-\epsilon)\frac{x^T L_G x}{c_k} \leq x^T L_H x \leq (1+\epsilon)\frac{x^T L_G x}{c_1} \tag{7}$$

Outline of the proof (cont'd)

Then, the resultant inequality is

$$(1-\epsilon)c_1 x^T L_G x \leq \sum_{i=1}^{t} x^T L_{H_i} x \leq (1+\epsilon)c_k x^T L_G x, \tag{6}$$

and multiplying it by $\frac{1}{c_1 c_k}$ we get

$$(1-\epsilon)\frac{x^T L_G x}{c_k} \leq x^T L_H x \leq (1+\epsilon)\frac{x^T L_G x}{c_1} \tag{7}$$

Finally, taking an

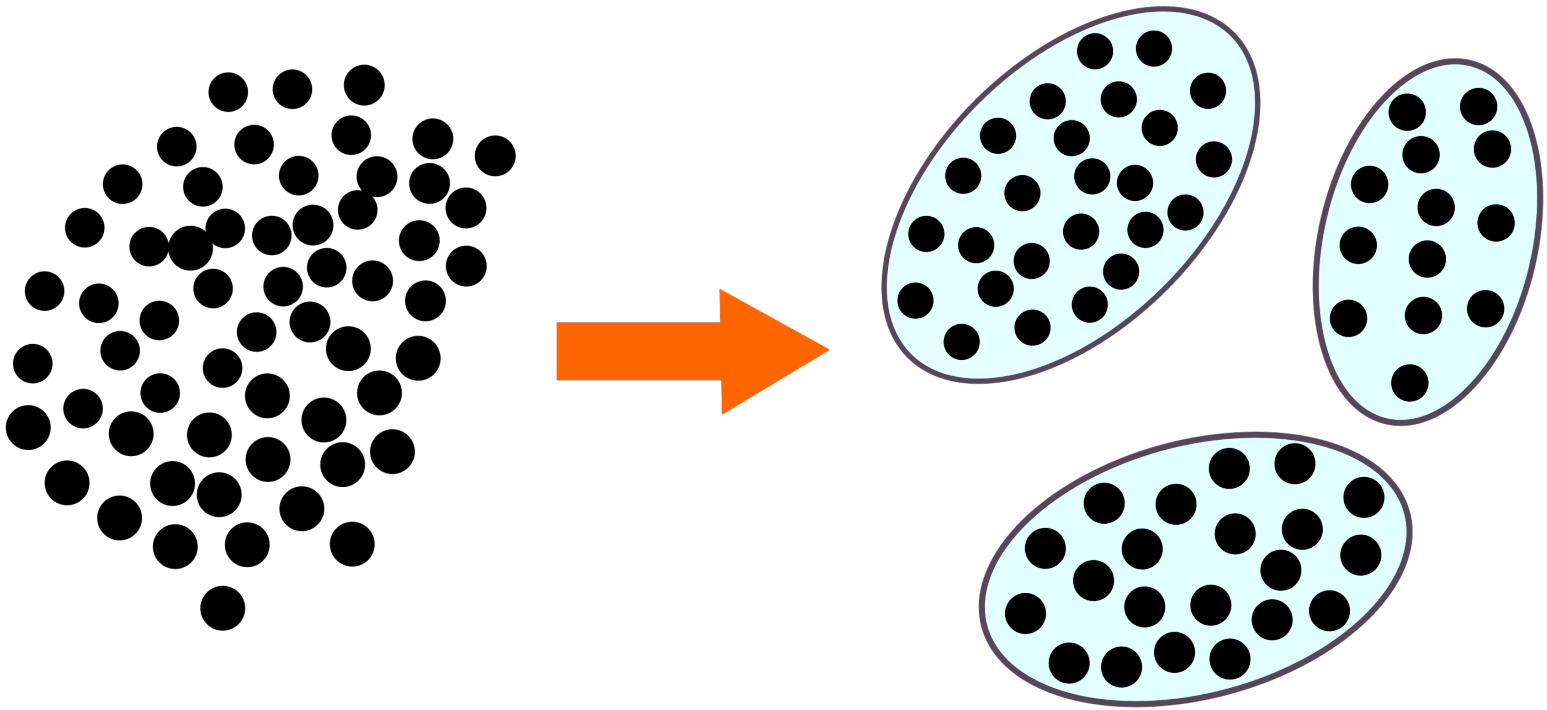
$$\epsilon' \geq 1 - \frac{1-\epsilon}{c_k} \tag{8}$$

we get an inequality in the form of Definition 2.

Outline for Section 3

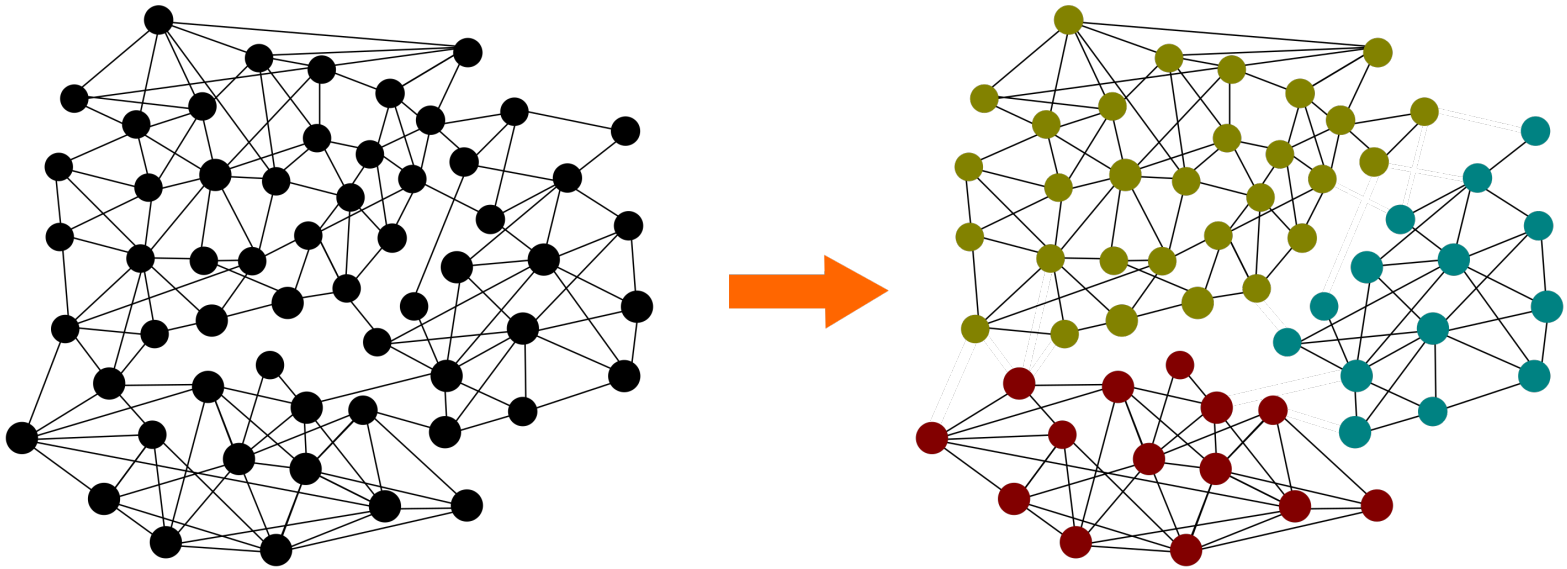
Clustering Problem

Clustering seeks to find a partition on a subset of points $X \subset \mathbb{R}^d$



Graph Clustering

Graph Clustering solves the Clustering problem by transforming the set X into a similarity graph and seeks for a minimum cut



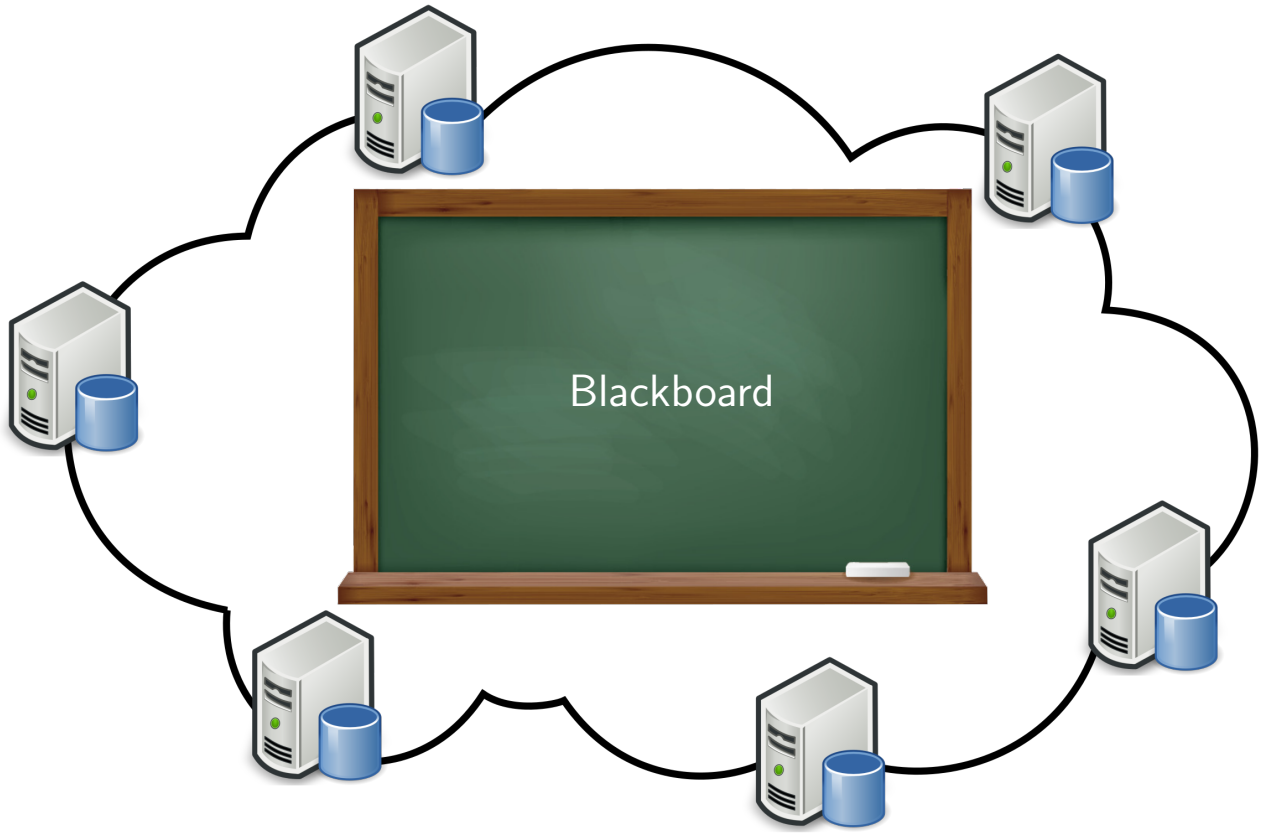
It is known that the eigenvectors of the first k eigenvalues of L_G taking in nondecreasing order are used to approximate a minimum multicut in G[1].

[1] James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. "Multiway Spectral Partitioning and higher-order cheeger inequalities.". In: *Journal of the ACM (JACM)* 61.6 (2014), p. 37.

Outline for Section 3

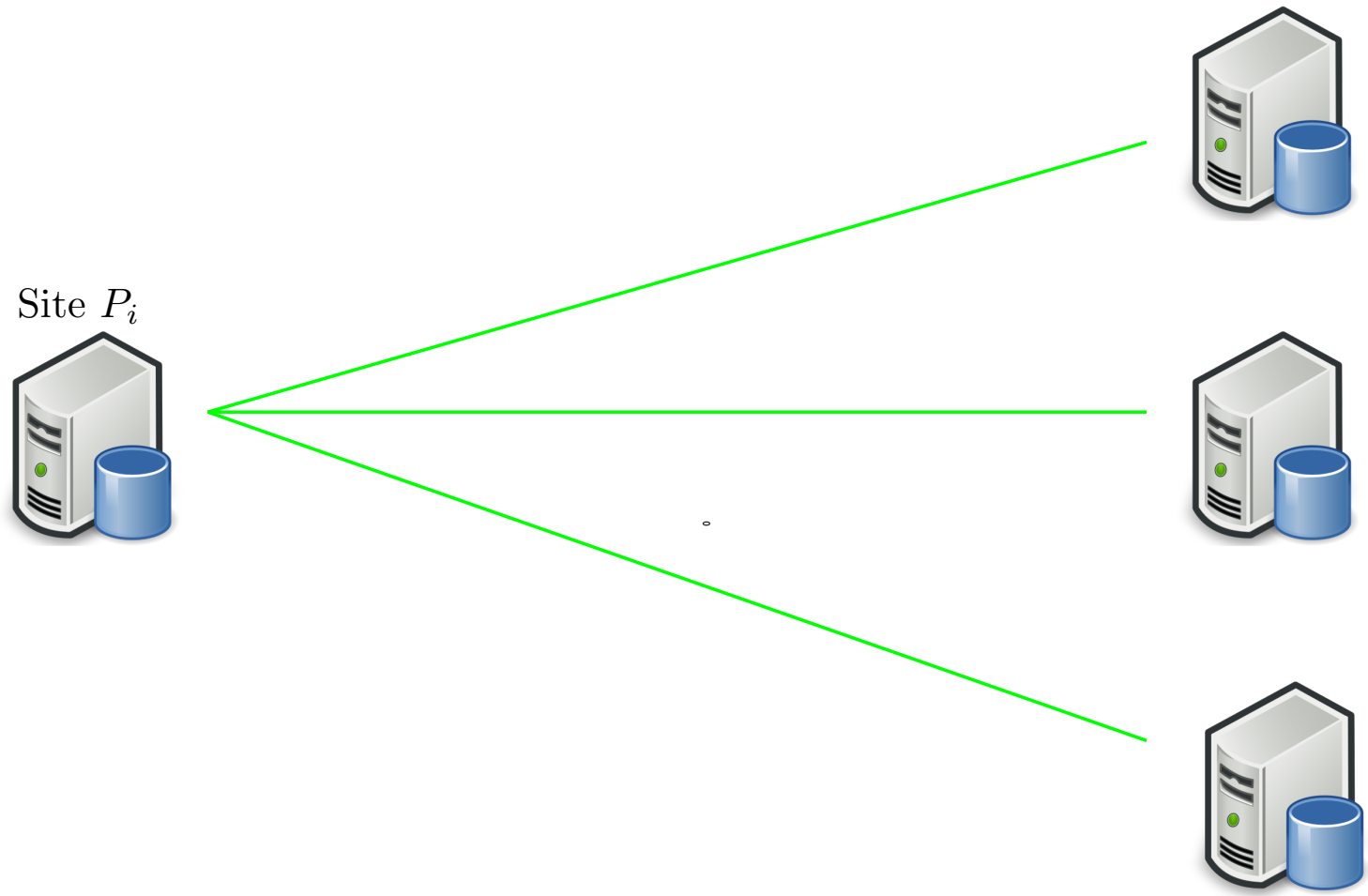
Communication Complexity

There are s sites which want to compute a function $f : X_1 \times X_2 \times \ldots X_s \to Z$ where X_i is the input of site P_i.
Each site sends bits to the others so that they can compute the function f, the way in which the bits are sent defined a so called communication protocol.



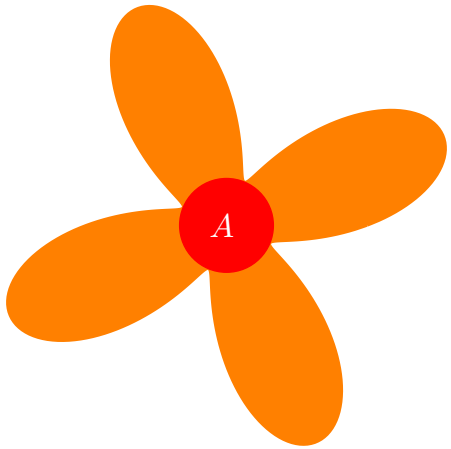
Number-On-Forehead Model (NOF)

The site P_i has access to the input $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_s)$.



Δ-Systems

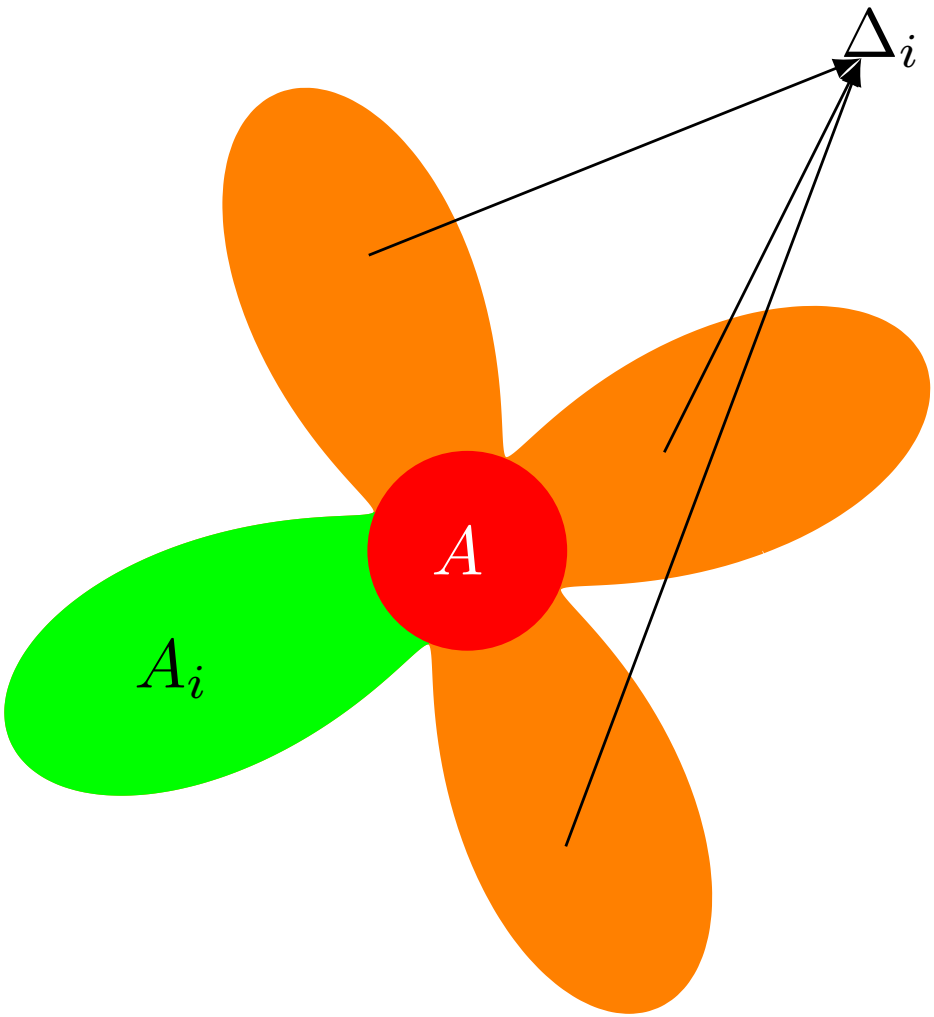
- A Sunflower or Δ-System is a family of sets $F = \{A_1, ..., A_t\}$ where $(A_i \cap A_j) = \bigcap_k^t A_k$ for all $i \neq j$.
- If the members of F are of size ℓ and $|A_i \cap A_j| = \lambda$ for all $i \neq j$ the F is a Weak Δ-System.



It is known that if F is a weak Δ-System and $|F| \geq \ell^2 - \ell + 2$, then F is a Sunflower[2].

[2] Michel Deza. "Solution d'un problème de Erdös-Lovász". In: *Journal of Combinatorial Theory, Series B* 16.2 (1974), pp. 166–167.

Δ-Systems



$\Rightarrow$ $A = \bigcap_{i=1}^{t} A_i$ is the Kernel

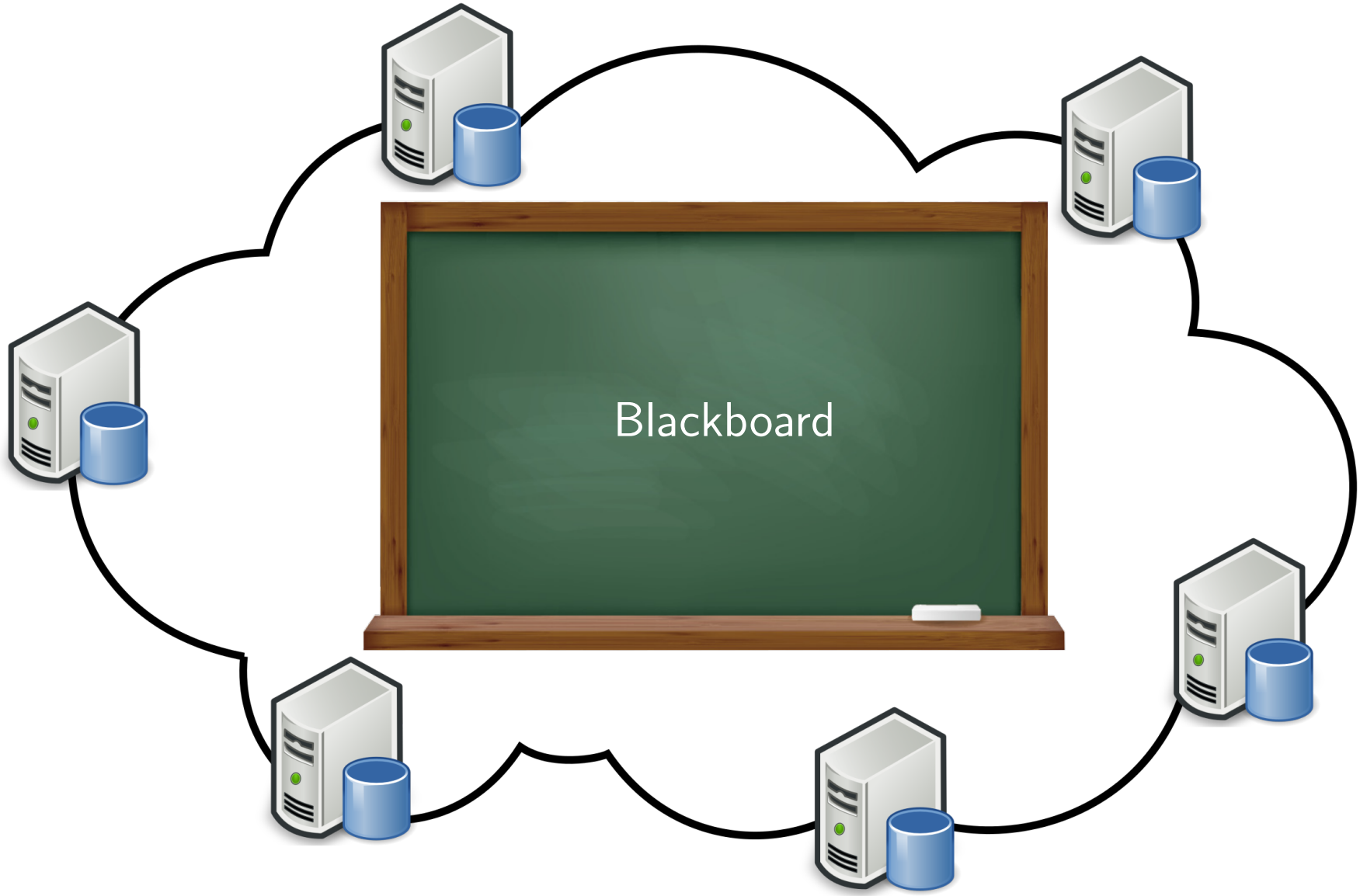
$\Rightarrow$ $\Delta_i = \bigcup_{j \neq i}^{t} (A_j - A)$ is the Generalized Symmetric Difference

$\Rightarrow$ $\delta_i = \frac{|A|}{|\bigcup_{j \neq i}^{t} A_j|}$ is the Overlapping Coefficient

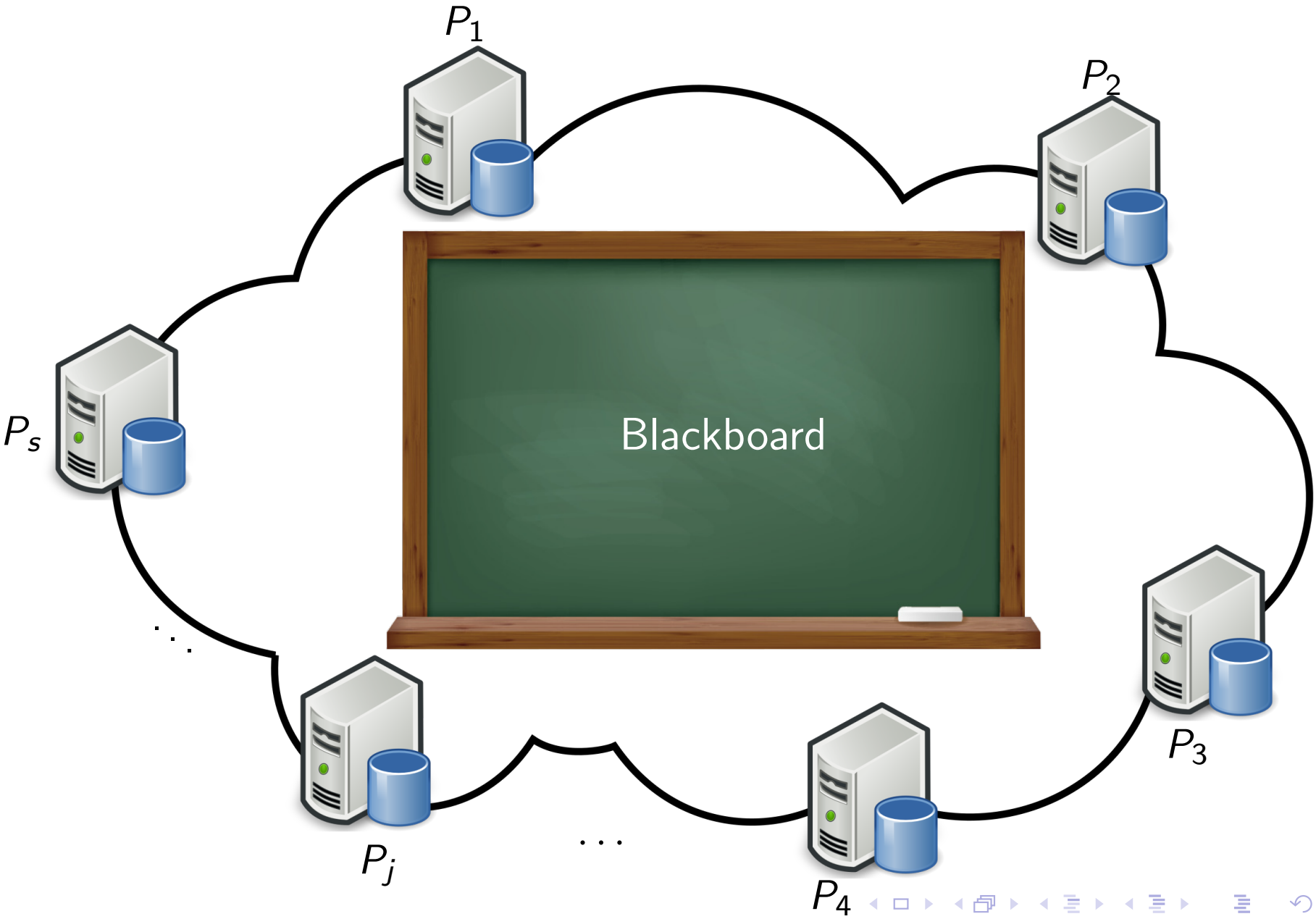
$\Rightarrow$ $\delta = \max_i \{\delta_i\}$ is the maximum overlapping coefficient

Clustering in NOF

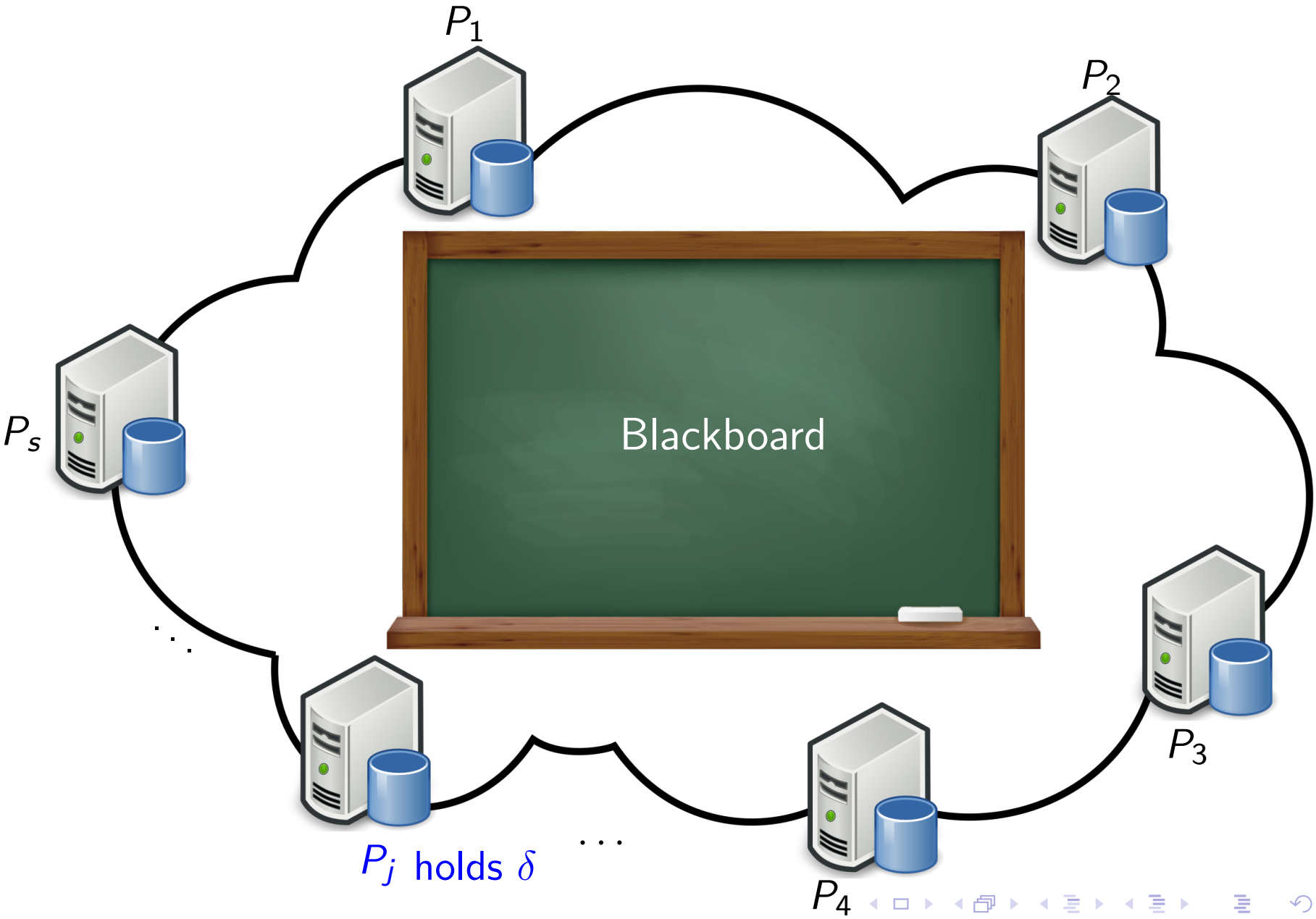
Clustering in NOF



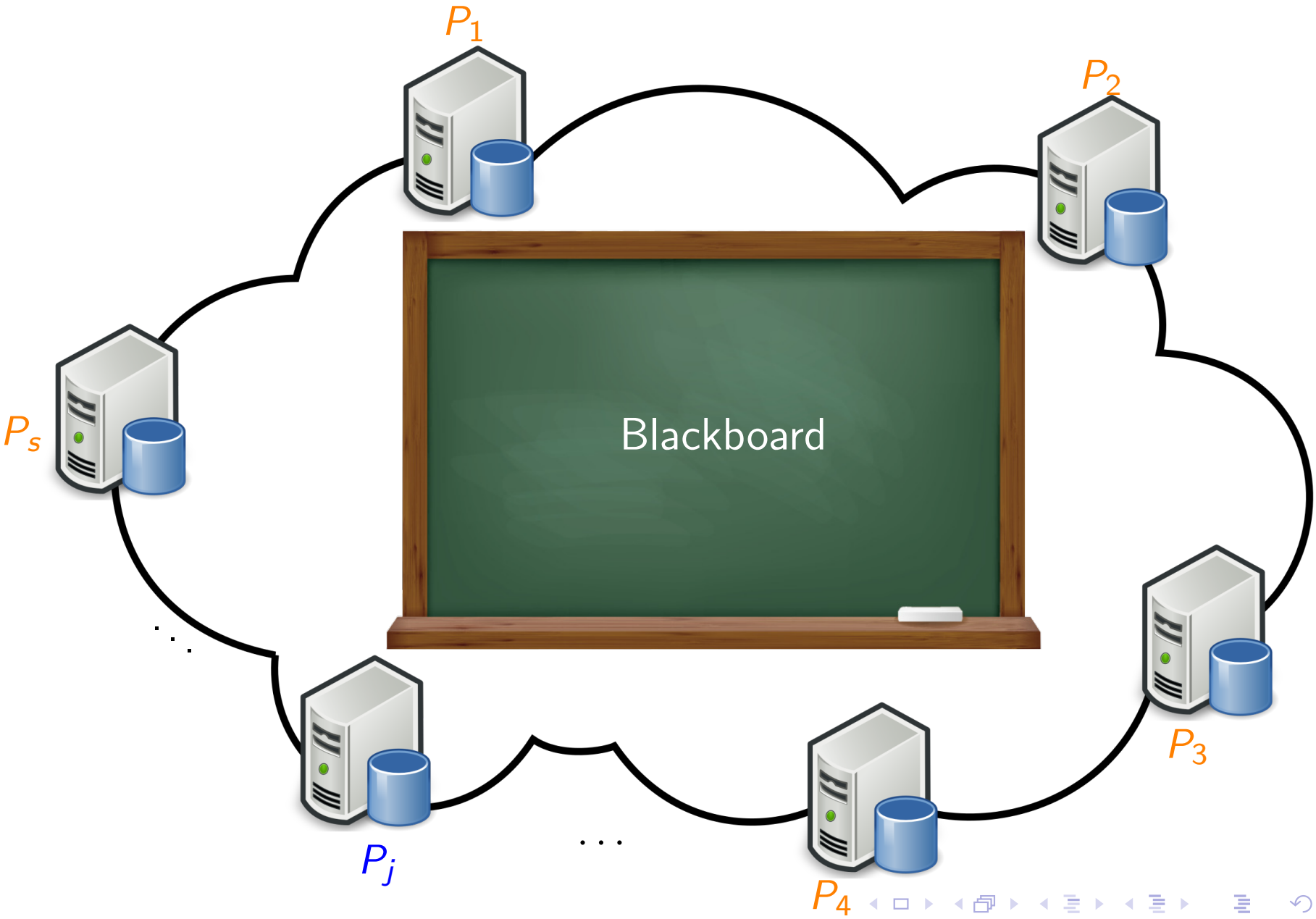
Clustering in NOF



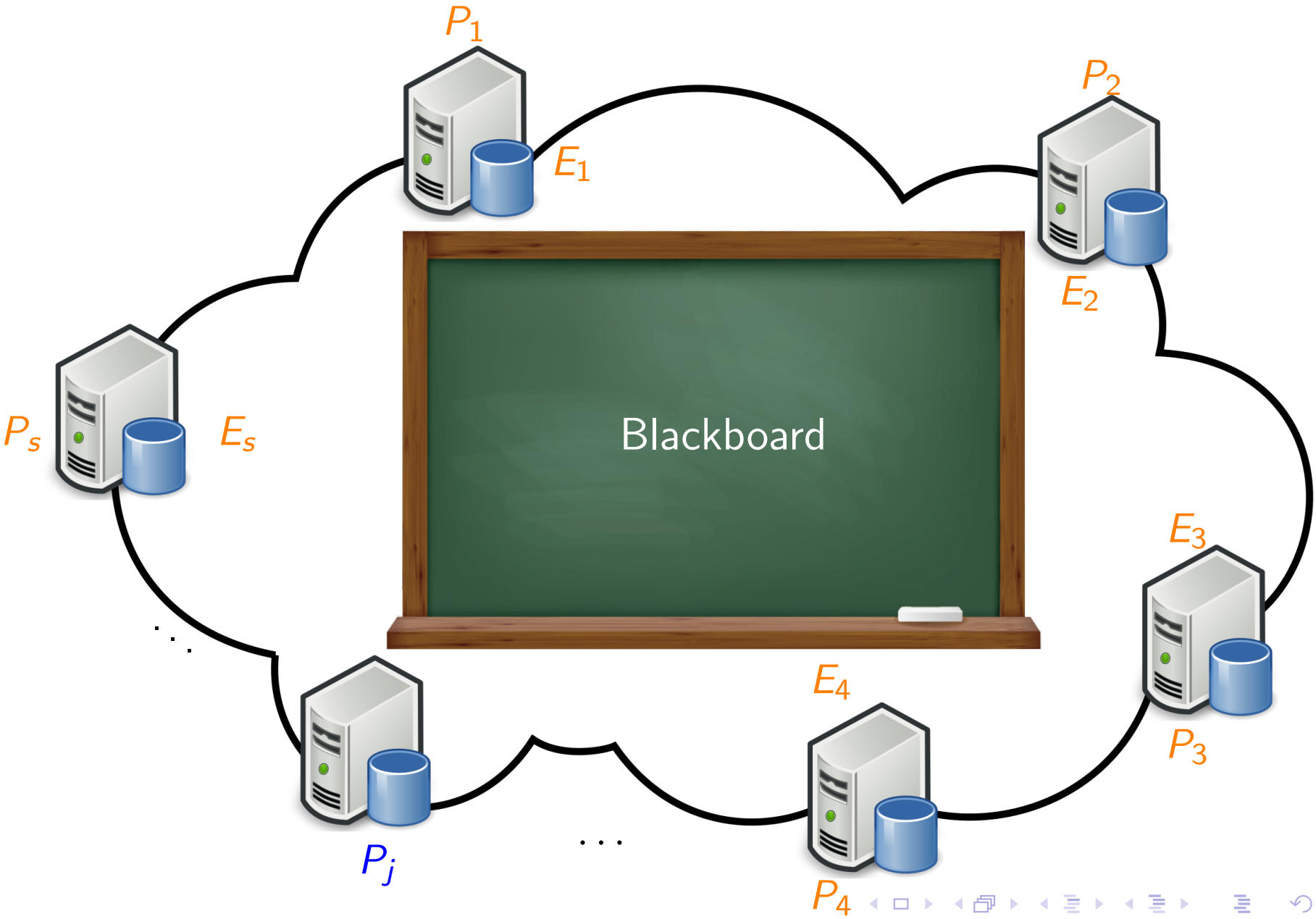
Clustering in NOF



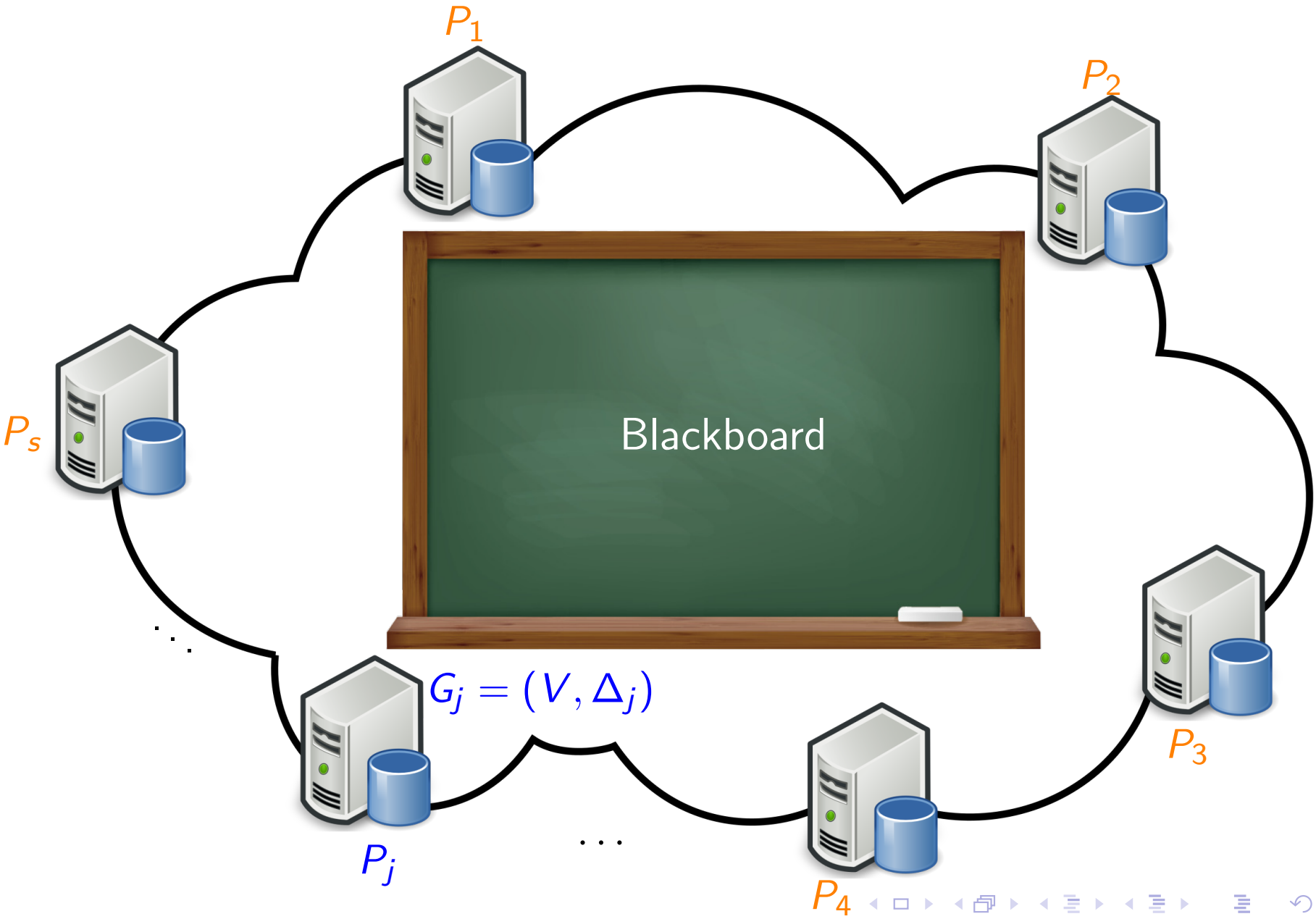
Clustering in NOF



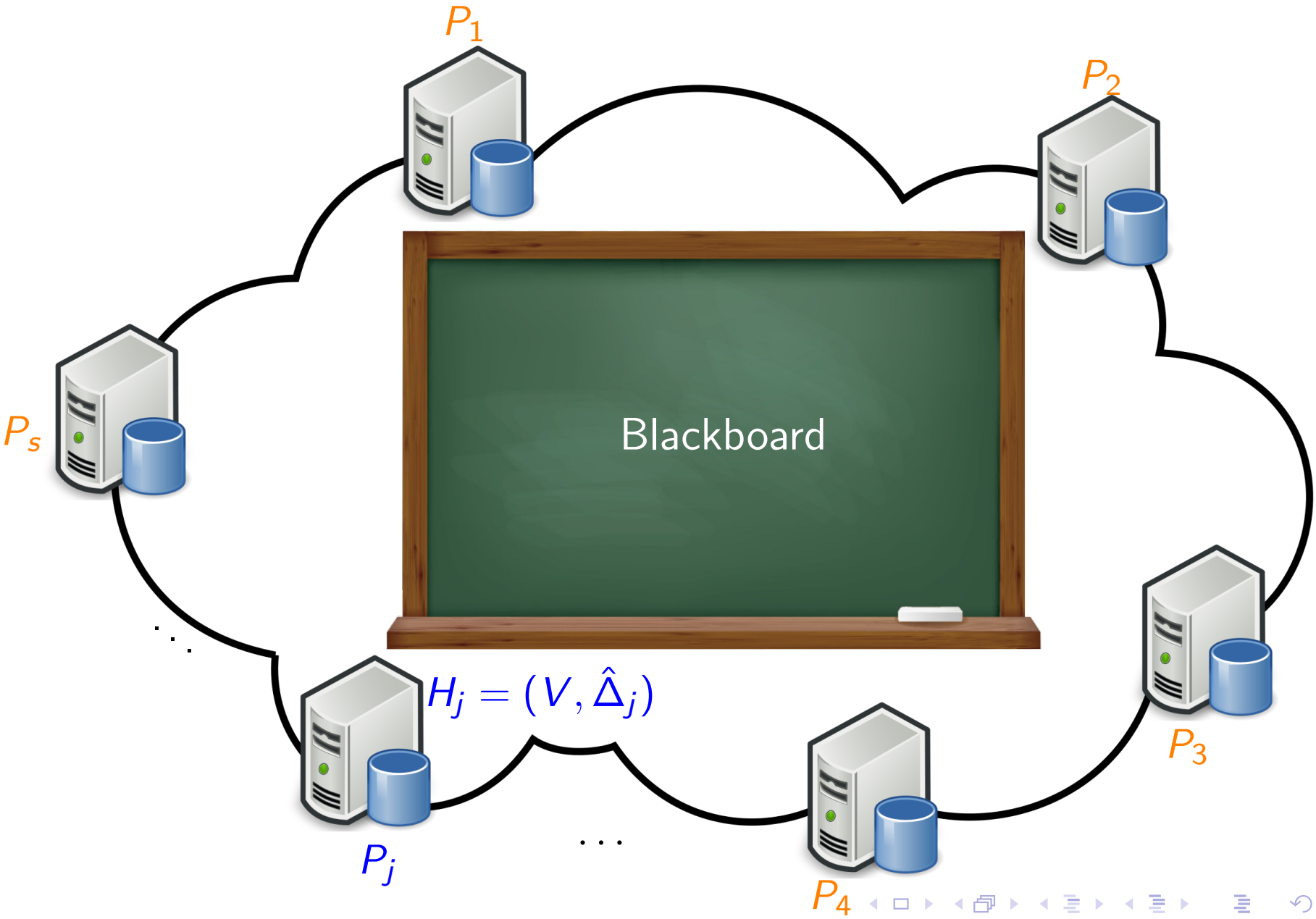
Clustering in NOF



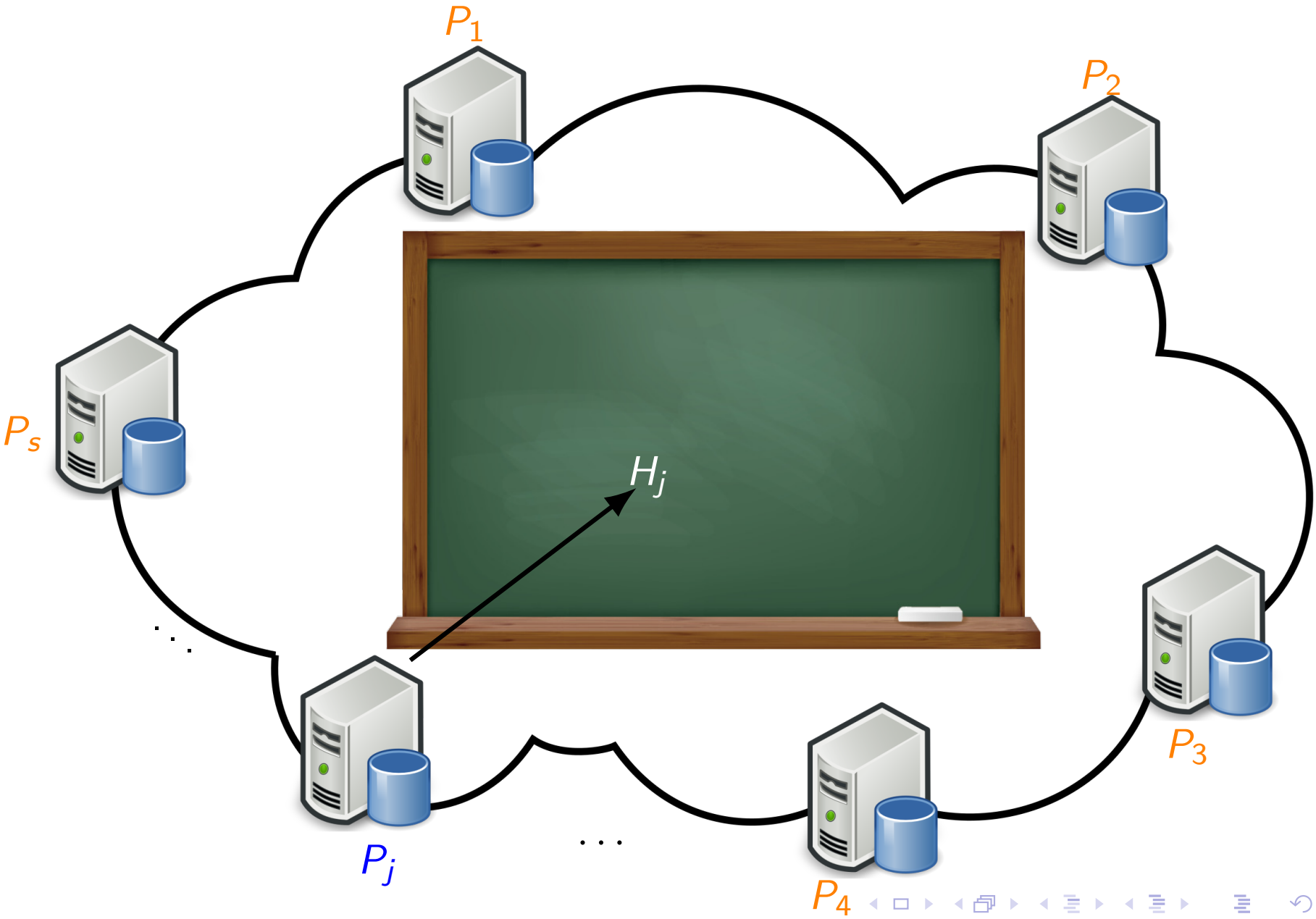
Clustering in NOF



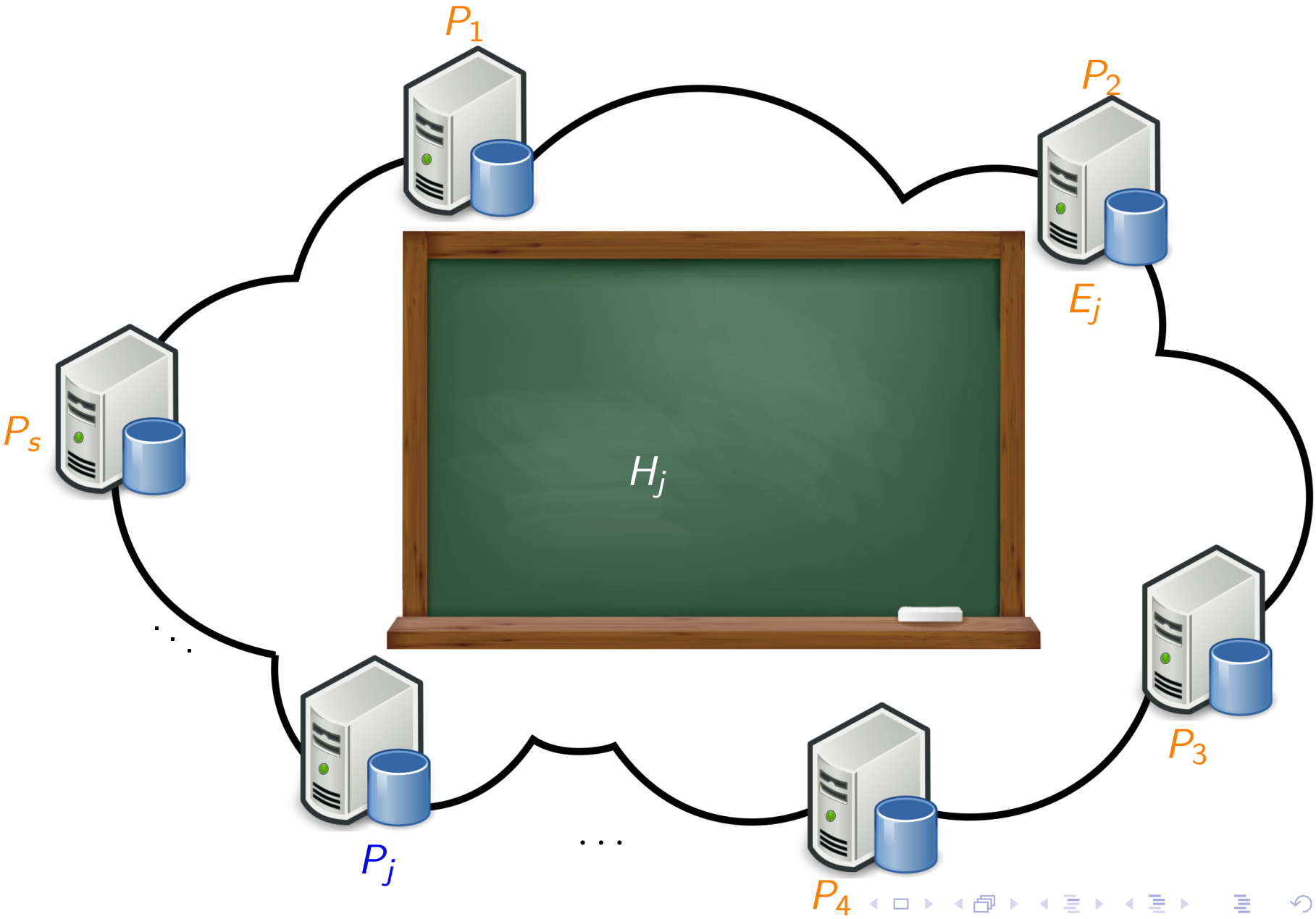
Clustering in NOF



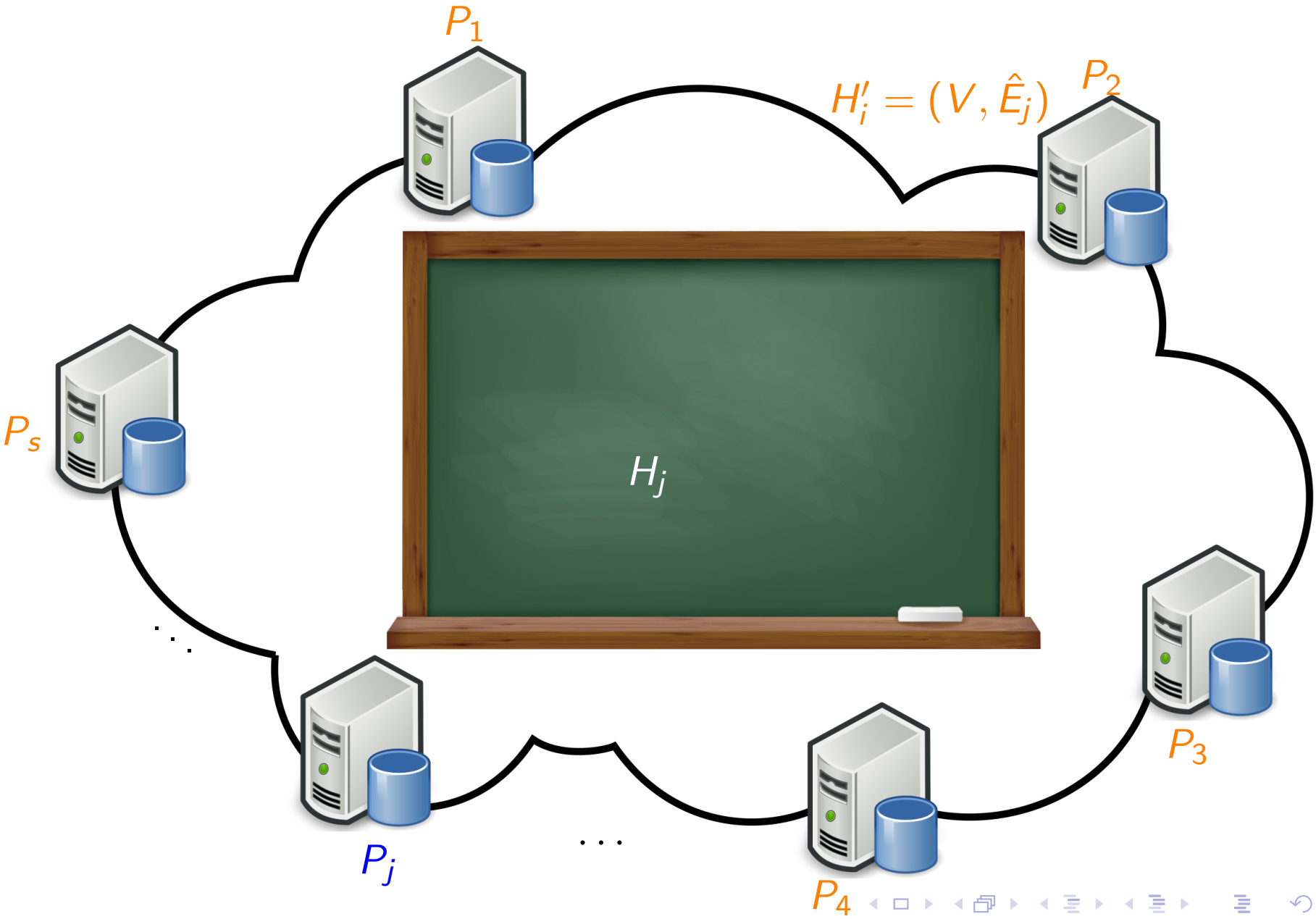
Clustering in NOF



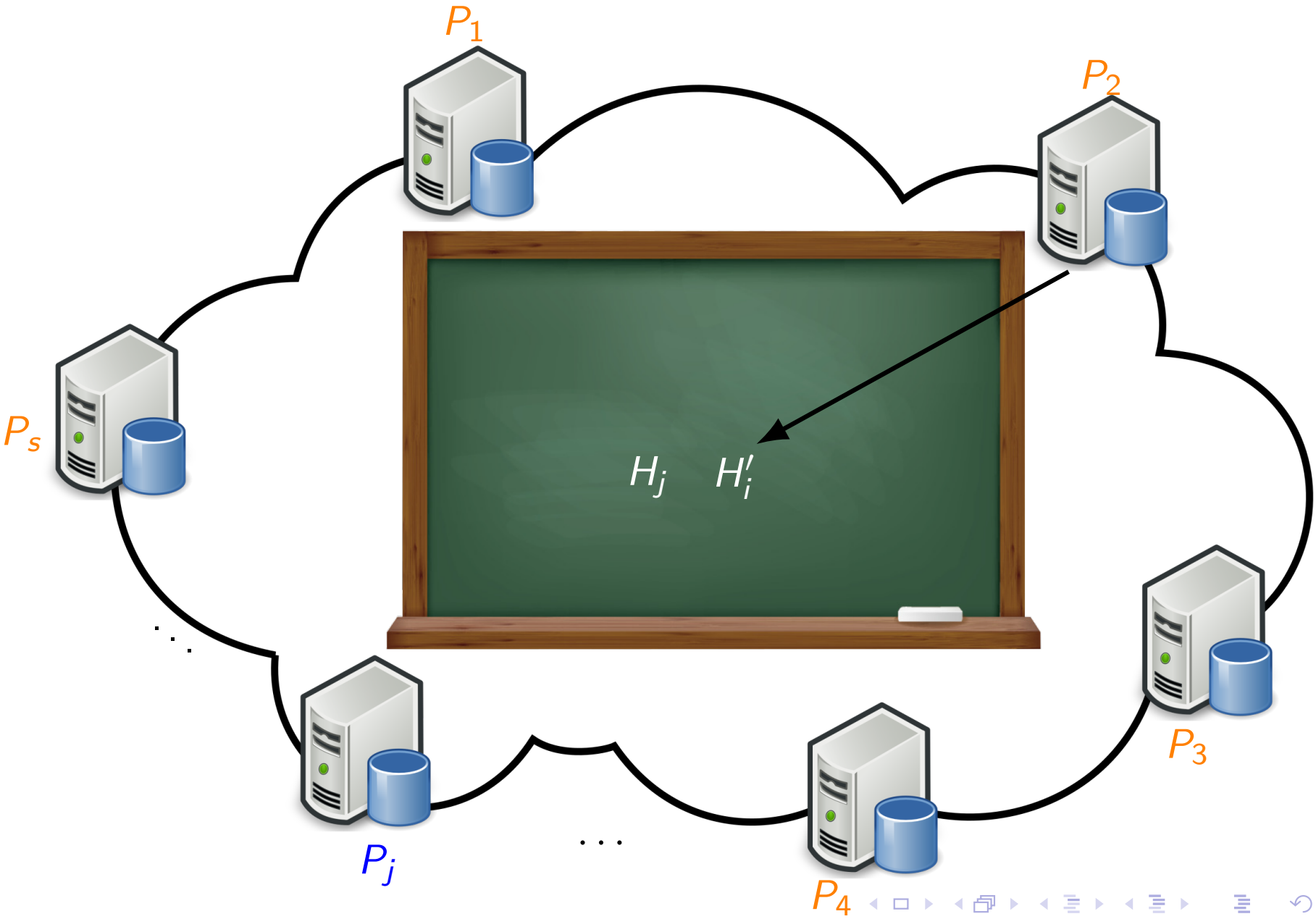
Clustering in NOF



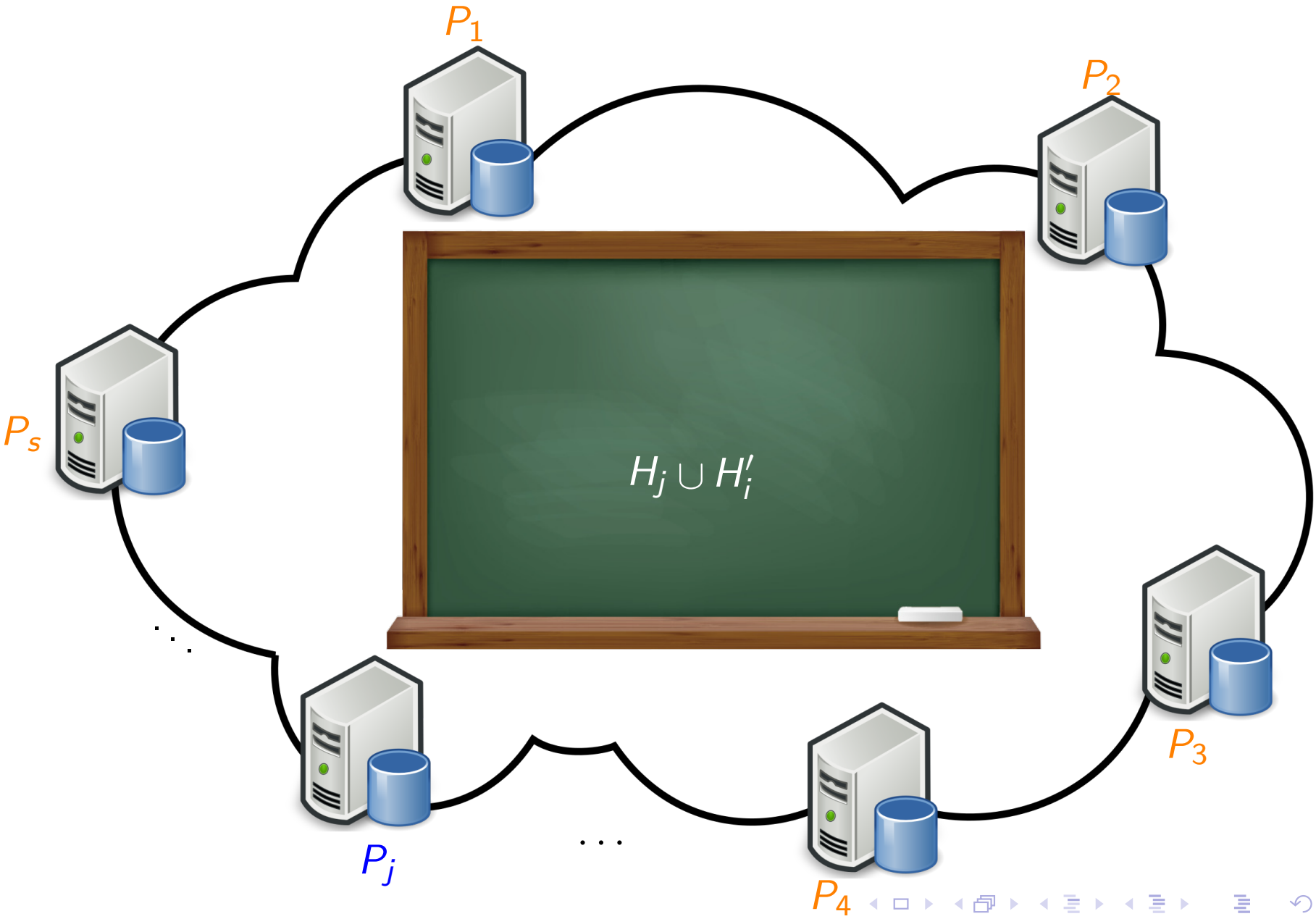
Clustering in NOF



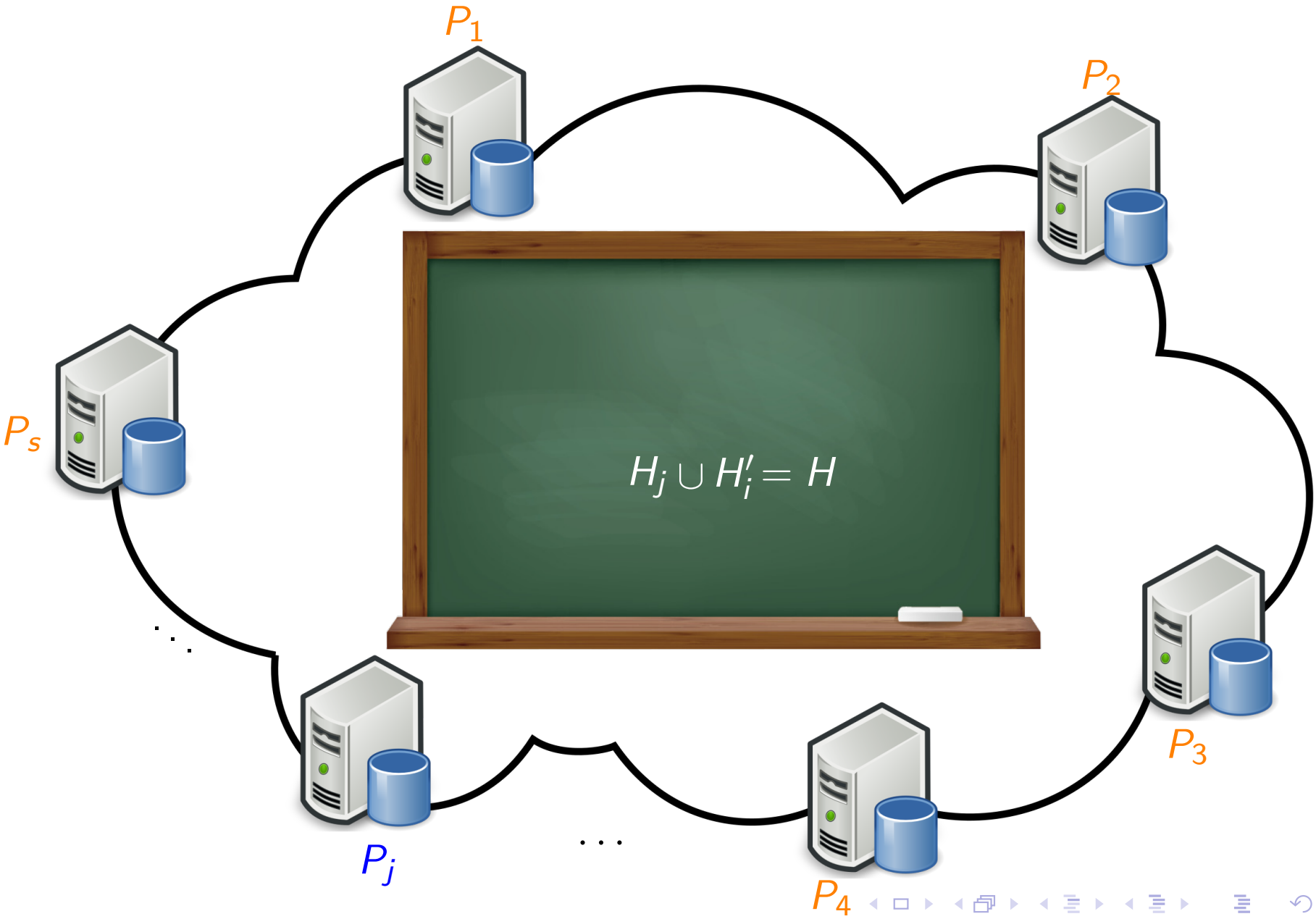
Clustering in NOF



Clustering in NOF



Clustering in NOF



Communication Cost

Finally, the communication cost of our protocol is

$$O(\log(\frac{n}{\epsilon^2}\sqrt{1-\delta})),$$

where n is the number of vertices, ϵ is the spectral approximation factor and δ is the maximum overlapping coefficient.

Summary

- We showed that the union of spectral sparsifiers of subgraphs of a given graph G is a spectral sparsifier of the graph G as well.
- We gave an exact computation of the spectral approximation factor ϵ' for the union of spectral sparsifiers.
- We gave an application of the union of Spectral Sparsifier for computing Clustering in a distributed problem with overlapping data.

Thank you for your attention!