Lecture 12

Thursday, October 10, 2024 10:04 AM

Chap 1 & 2

- → norms, angles, projections
- → var, cov, mean, eigenvalues/vectors ←
- → random variable, statistics

$\underbrace{\text{random variable, statistics}}_{\text{sample vector}}$

Chapter 6 : high dimensions

- ↳ vol ← Sphere
- ↳ angles Cubes
- ↳ normal rectangles

Chapter 7

PCA

- ↳ eigenvectors ←
- → eigenvalues ← significance

$\underbrace{\text{Var, Cov}}_{\text{input space}}$, mse, total variance, projected variance

eigen space

Chapter 8 & 9

Pattern mining — freq itemsets

↙ Apriori ↘ Eclat + diffset

(NO FPGrowth 8.2.3 8.2.4)

closed itemsets — charm

↳ closure operator

Chapter 13 k-mean, EM

Objective (SSE)

E-step: assignment (0/1 or probability)

M-step: update the parameters

] mixture of gaussians

Bayes theorem

Chapter 15 : density (skip 15.1)

↳ kernel density estimation $f(x)$

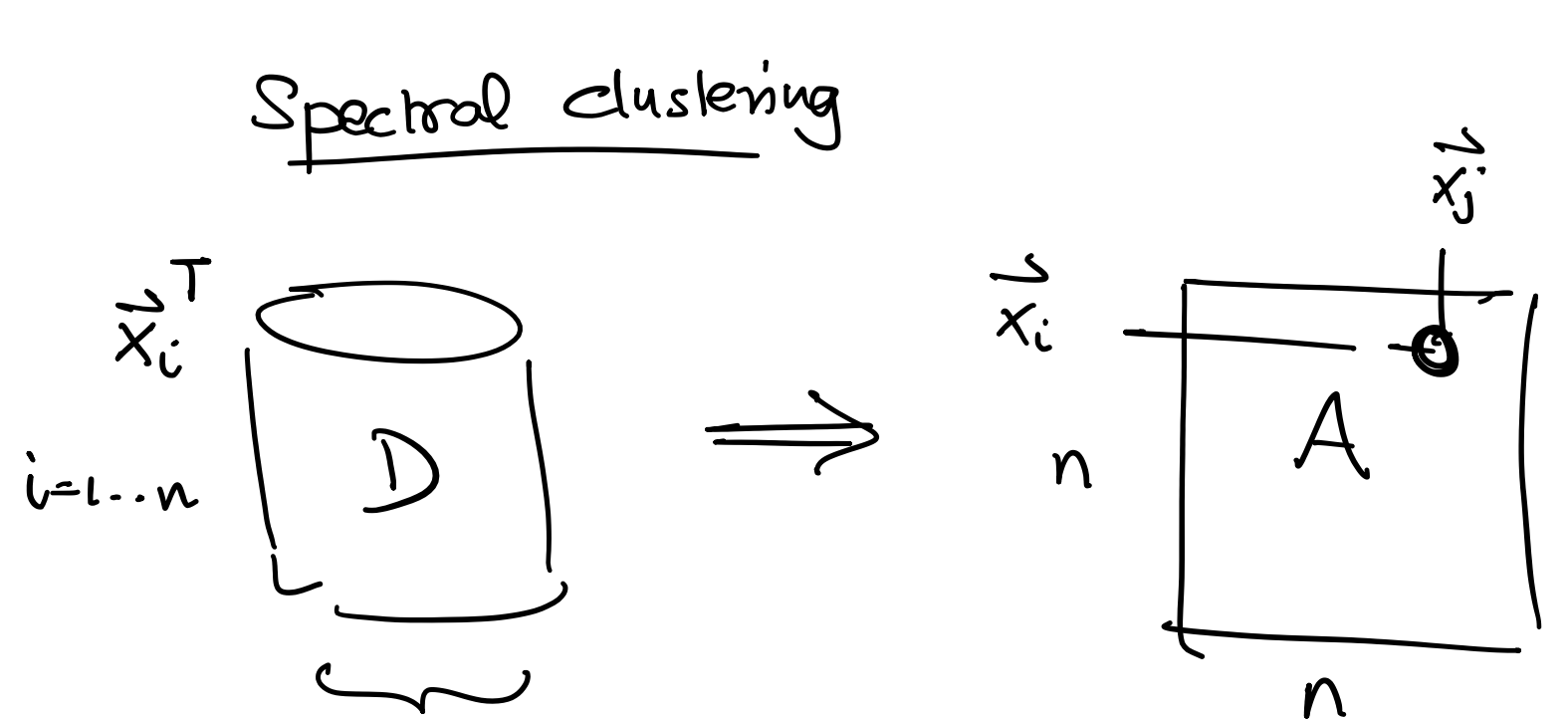
(smooth histograms)

↳ denclue/ mean-shift

find the attractors

find the clusters

Spectral clustering



$a_{ij} = \text{sim}(\vec{x}_i, \vec{x}_j)$

$= e^{-\|\vec{x}_i - \vec{x}_j\|^2 / 2\sigma^2}$

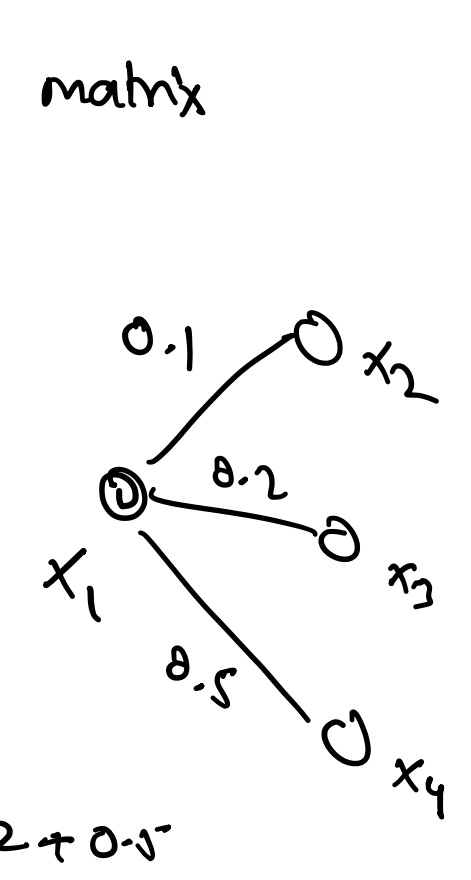
(weighted) graph

$A \equiv$ similarity matrix

$\equiv$ weighted adjacency matrix

$a_{ij} = a_{ji}$ symmetric

$a_{ij} \geq 0$ non-negative



$\Delta \equiv$ degree matrix

↳ diagonal matrix

$d(x_1) = 0.1 + 0.2 + 0.5$

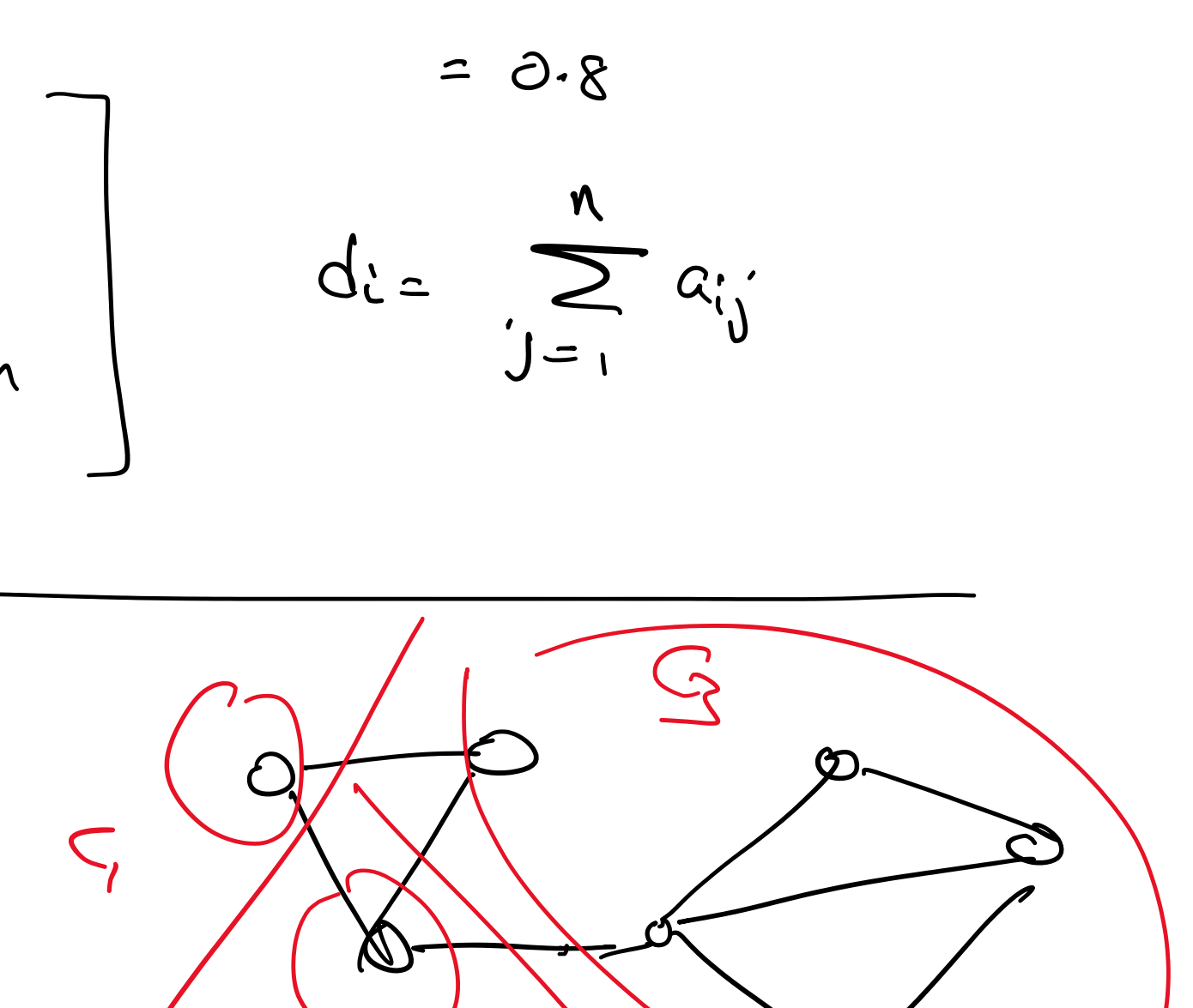
$= 0.8$

$$\Delta = \begin{bmatrix} d_1 & & & 0 \\ & d_2 & & \\ & & \ddots & \\ 0 & & & d_n \end{bmatrix}$$

$$d_i = \sum_{j=1}^{n} a_{ij}$$

---

n [A] n



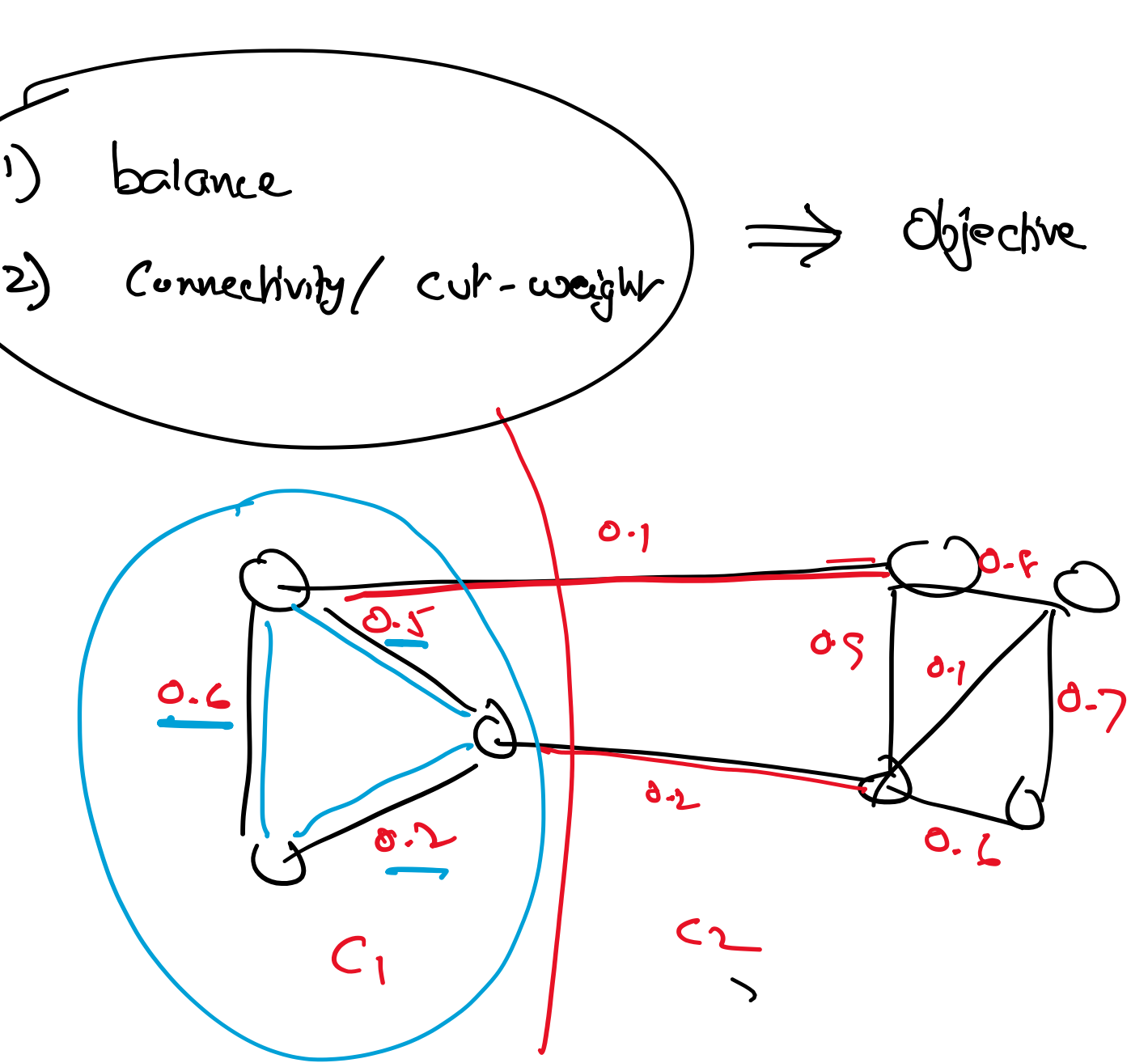
find k clusters

⇓

(k-way partitioning of the graph)

(1) balance 2) Connectivity/ cut-weight) ⇒ Objective

V = all the points/ nodes



$W(C_1, V) \equiv \text{volume}(C_1)$

$= \underset{i \in C_1}{\sum} \underset{j \in V}{\sum} a_{ij}$

= internal weights (1.3)

+

external weights (0.3)

= 1.6

$\text{Size}(C_1) = |C_1|$

= 3

$W(C_1, C_2) = \underset{i \in C_1}{\sum} \underset{j \in C_2}{\sum} a_{ij}$

Cut weight

$= 0.1 + 0.2 = 0.3$

Ratio Cut objective : J_{rc}

given A (n×n matrix of similarities / adjacency matrix)

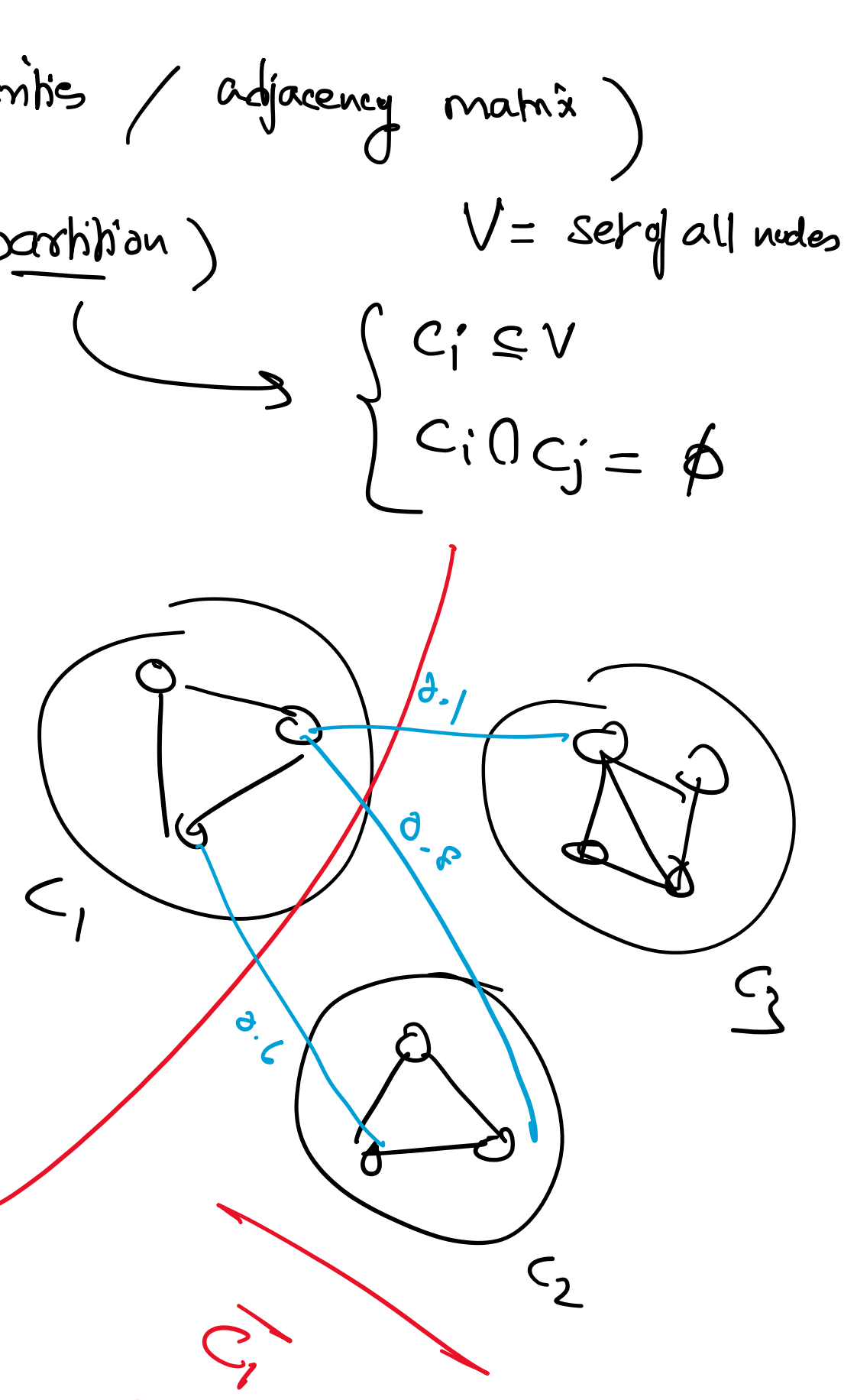
find $\underbrace{C_1, C_2, \ldots, C_k}_{\mathcal{C}}$ (partition)

V = set of all nodes

$\left\{ \begin{array}{l} C_i \subseteq V \\ C_i \cap C_j = \phi \end{array} \right.$

$\min_{\mathcal{C}} J_{rc} = \sum_{i=1}^{k} \frac{W(C_i, \bar{C}_i)}{|C_i|}$ ← min, ← max

$W(C_i, \bar{C}_i) \equiv$ Sum of all external weights



Normalized cut : J_{nc}

$$J_{nc} = \sum_{i=1}^{k} \frac{W(C_i, \bar{C}_i)}{\text{vol}(C_i)}$$ min, max

---

$$J_{rc} = \sum_{i=1}^{k} \frac{W(C_i, \bar{C}_i)}{|C_i|}$$

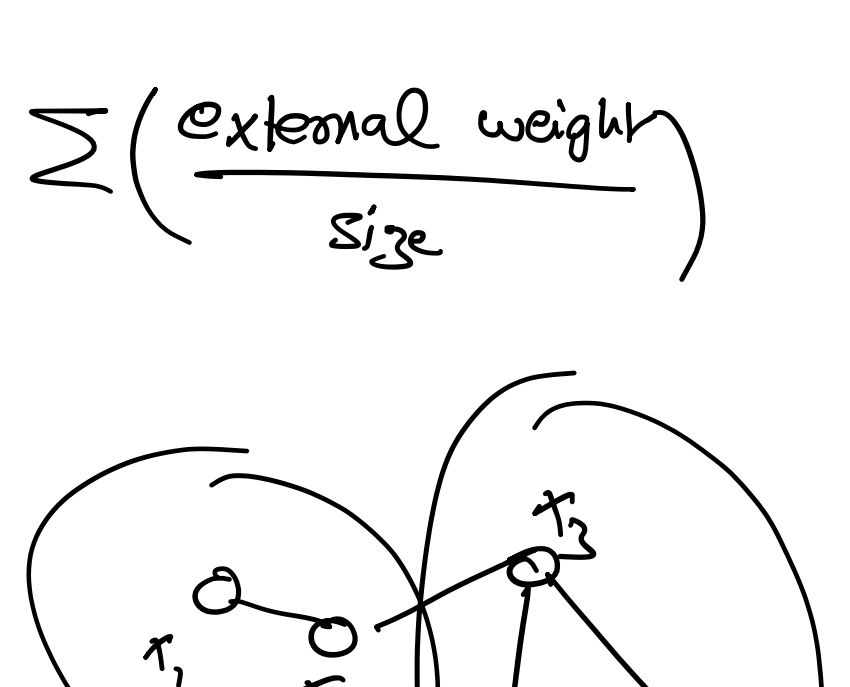
$\Sigma\left(\frac{\text{external weight}}{\text{size}}\right)$

A = n×n adjacency matrix

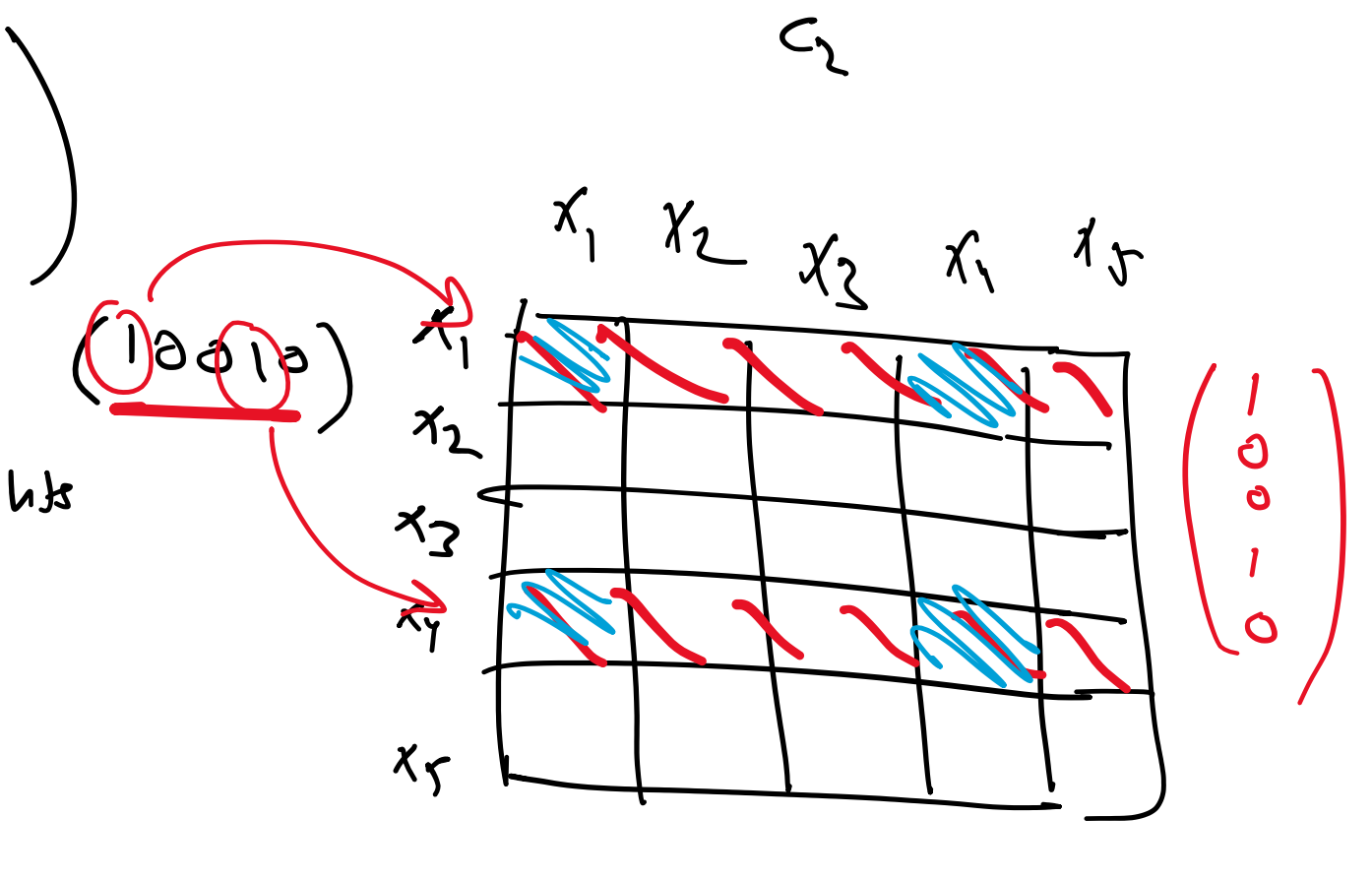
$\vec{c}_i \equiv$ cluster indicator vector

$\in \{0/1\}^n$

binary



$$\vec{c}_1 = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \vec{c}_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$



$W(C_i, C_i) \equiv$ internal weights

$= \vec{c}_i^T A \vec{c}_i$

$|C_i| = \boxed{\vec{c}_i^T \vec{c}_i}$

$\text{vol}(C_i) = \vec{c}_i^T \Delta \vec{c}_i = W(C_i, V) \equiv$ internal + external

$W(C_i, \bar{C}_i)$ = external

= vol − internal

$= W(C_i, V) - W(C_i, C_i)$

$= \vec{c}_i^T \Delta \vec{c}_i - \vec{c}_i^T A \vec{c}_i$

$= \vec{c}_i^T \underbrace{(\Delta - A)}_{L} \vec{c}_i$

$= \vec{c}_i^T L \vec{c}_i$

Laplacian matrix

$L = \Delta - A$

Positive semi-definite matrix

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$

$\downarrow \quad \downarrow$

$\vec{u}_1 \quad \vec{u}_2 \quad \ldots \quad \vec{u}_n$

$$J_{rc} = \sum_{i=1}^{k} \frac{W(C_i, \bar{C}_i)}{|C_i|}$$

$$J_{rc} = \sum_{i=1}^{k} \frac{\vec{c}_i^T L \vec{c}_i}{\|\vec{c}_i\|^2}$$

~~$\vec{c} \in \{0,1\}^n$ binary vector~~

$$= \sum_{i=1}^{k} \left(\frac{\vec{c}_i}{\|\vec{c}_i\|}\right)^T L \left(\frac{\vec{c}_i}{\|\vec{c}_i\|}\right)$$

relaxation

$\vec{c}_i$ can be $\mathbb{R}$ now!

$$\min J_{rc} = \sum_{i=1}^{k} \vec{u}_i^T L \vec{u}_i$$

$\vec{u}_i$ = unit vector in direction of $\vec{c}_i$

Solution is k eigenvalues + eigenvectors of $\overset{J_{rc}}{\boxed{L}}$ vs $\underbrace{L^a \text{ vs } L^s}_{J_{nc}}$

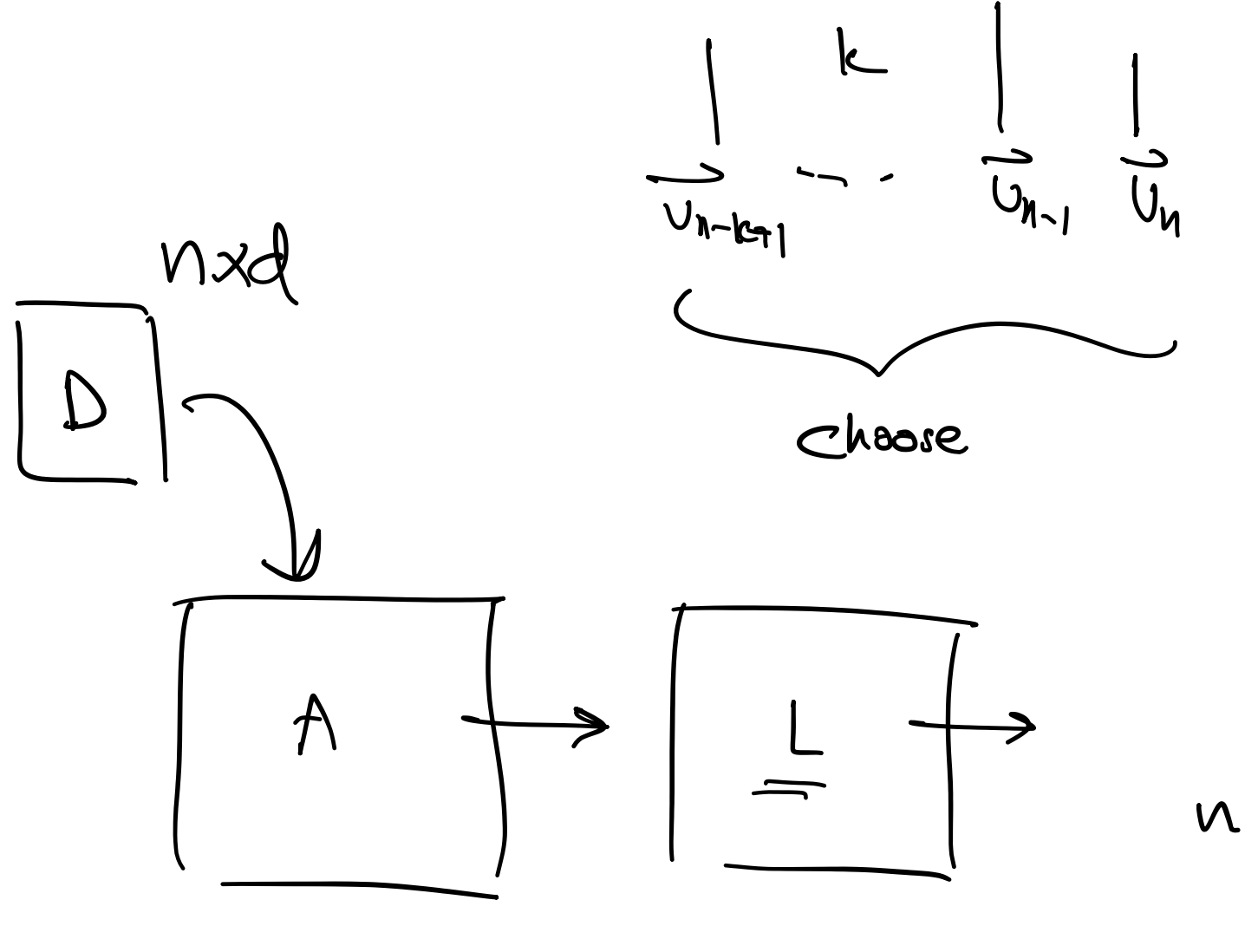
choose the k − smaller eigenvectors

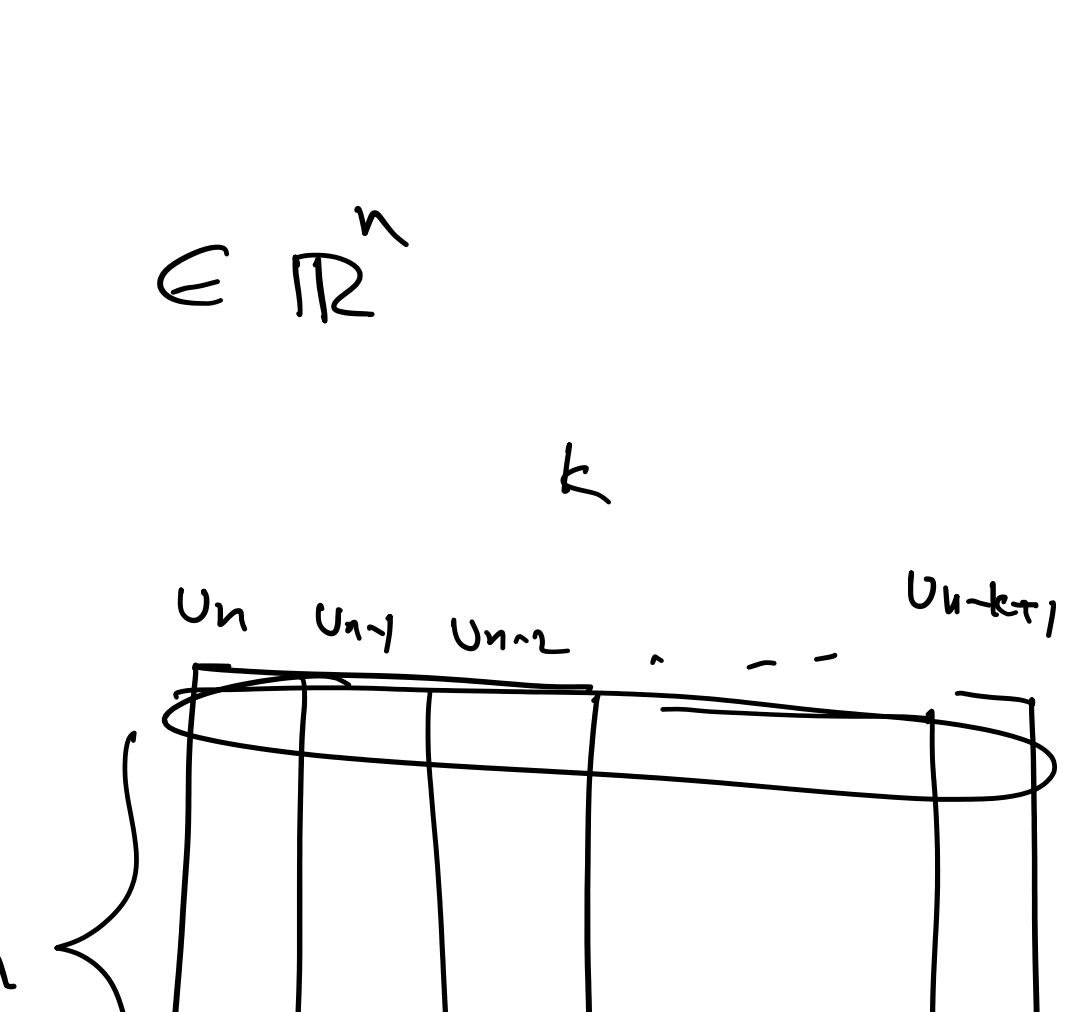
L

↳ $\lambda_1 \geq \lambda_2 \ldots \underbrace{\lambda_{n-k+1} \geq \lambda_{n-2} \geq \lambda_{n-1} \geq \lambda_n}_{k} \geq 0$

L : n×n

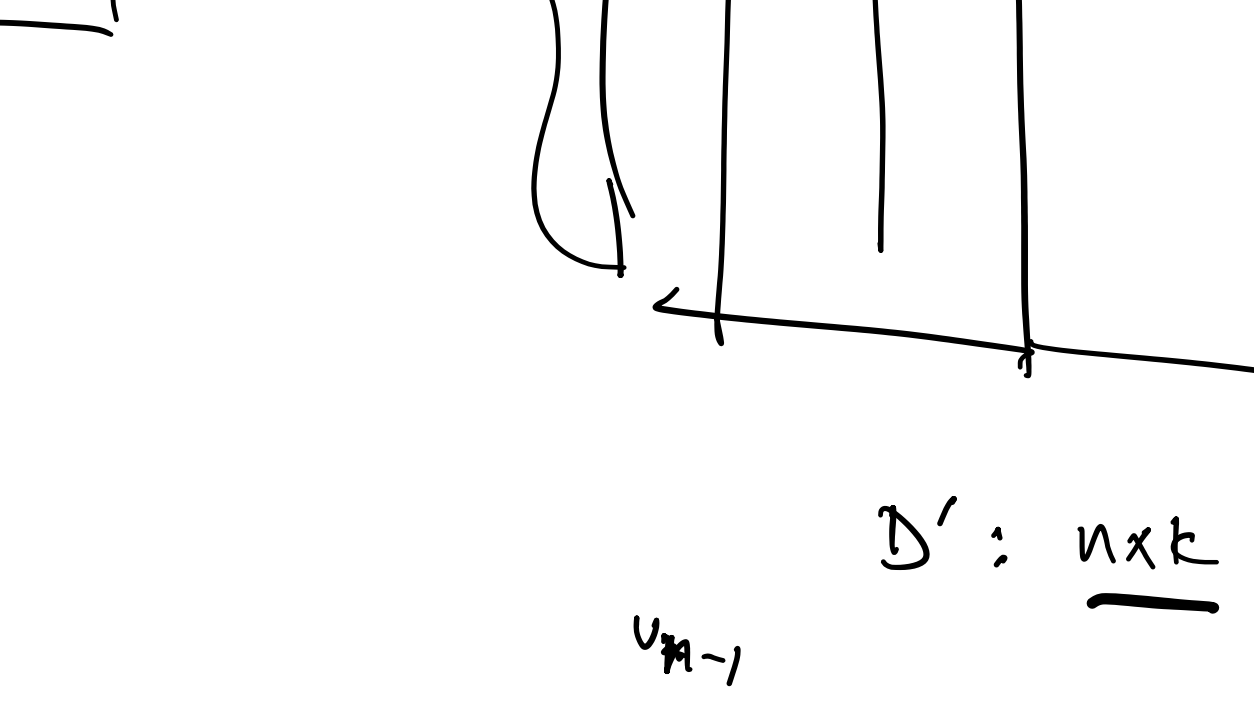
$\underbrace{\vec{u}_{n-k+1} \ldots \vec{u}_{n-1} \quad \vec{u}_n}_{\text{choose}} \in \mathbb{R}^n$





D' : n×k

Use simple k-means on D'



$\mathbb{R}^k$

$J_{nc} \rightarrow L^s \equiv$ symmetric laplacian

$\Delta^{-1/2} L \Delta^{-1/2}$

$\boxed{L = \Delta - A}$

Laplacian

Δ = degree matrix diagonal

$L^a \equiv$ asymmetric Laplacian

$\boxed{\Delta^{-1} L} \equiv$ normalized Laplacian

every row is divided by degree of that node

$$\Delta = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \quad \Delta^{-1} = \begin{bmatrix} 1/d_1 & 0 \\ 0 & 1/d_2 \end{bmatrix} \quad \Delta^{-1/2} = \begin{bmatrix} 1/\sqrt{d_1} & 0 \\ 0 & 1/\sqrt{d_2} \end{bmatrix}$$

Markov-chain clustering

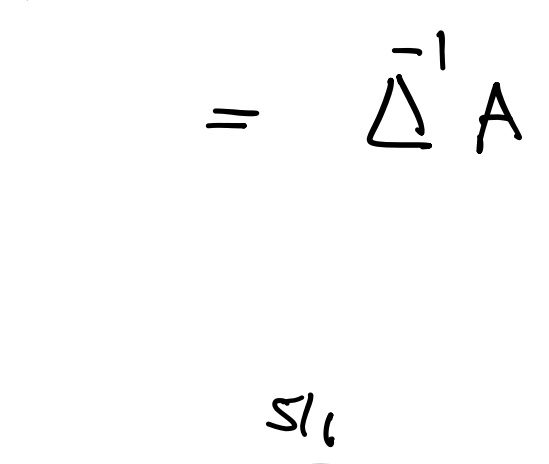
A → add self loops

M ≡ Markov matrix

$= \Delta^{-1} A$

$$A = \begin{array}{c|cc} & x_1 & x_2 \\ \hline x_1 & 0.1 & 0.5 \\ x_2 & 0.5 & 0.4 \end{array}$$

$$\Delta = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.9 \end{bmatrix}$$



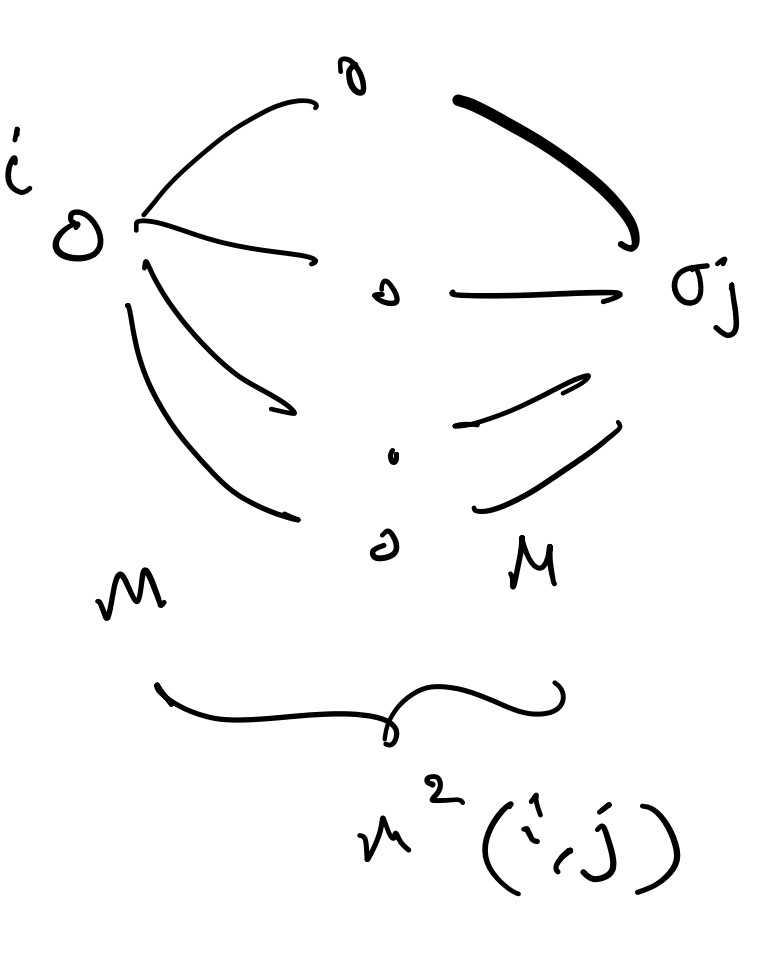
$$M = \Delta^{-1} A = \begin{bmatrix} 1/6 & 5/6 \\ 0.5/0.9 & 0.4/0.9 \end{bmatrix}$$

Sum of each row = 1

M ≡ one step transition matrix

(dominant eigen vector of M is pagerank vector)

M^2 = 2 step probability



random walk + bias (inflation)

$x^0 \xrightarrow{M} x^1 \xrightarrow{M} x^2 \longrightarrow \cdots \longrightarrow \boxed{x^t}$

($M^t \rightarrow$ top eigenvector → pagerank M^∞)

MCL (A)

0) add self loops ≡ new A

1) $M = \Delta^{-1} A$ (normalize A by degree of each node)

Iteration

$M^0 = M$

2) 1 step walk

$M^t = M^{t-1} * M^{t-1} = (M^{t-1})^2$

3) bias / inflation

$M^t = \text{Inflate}(M^t, r)$

} t : current step

4) repeat until Convergence

$\|M^t - M^{t-1}\|_F^2 \leq \varepsilon$ stop

5) extract clusters

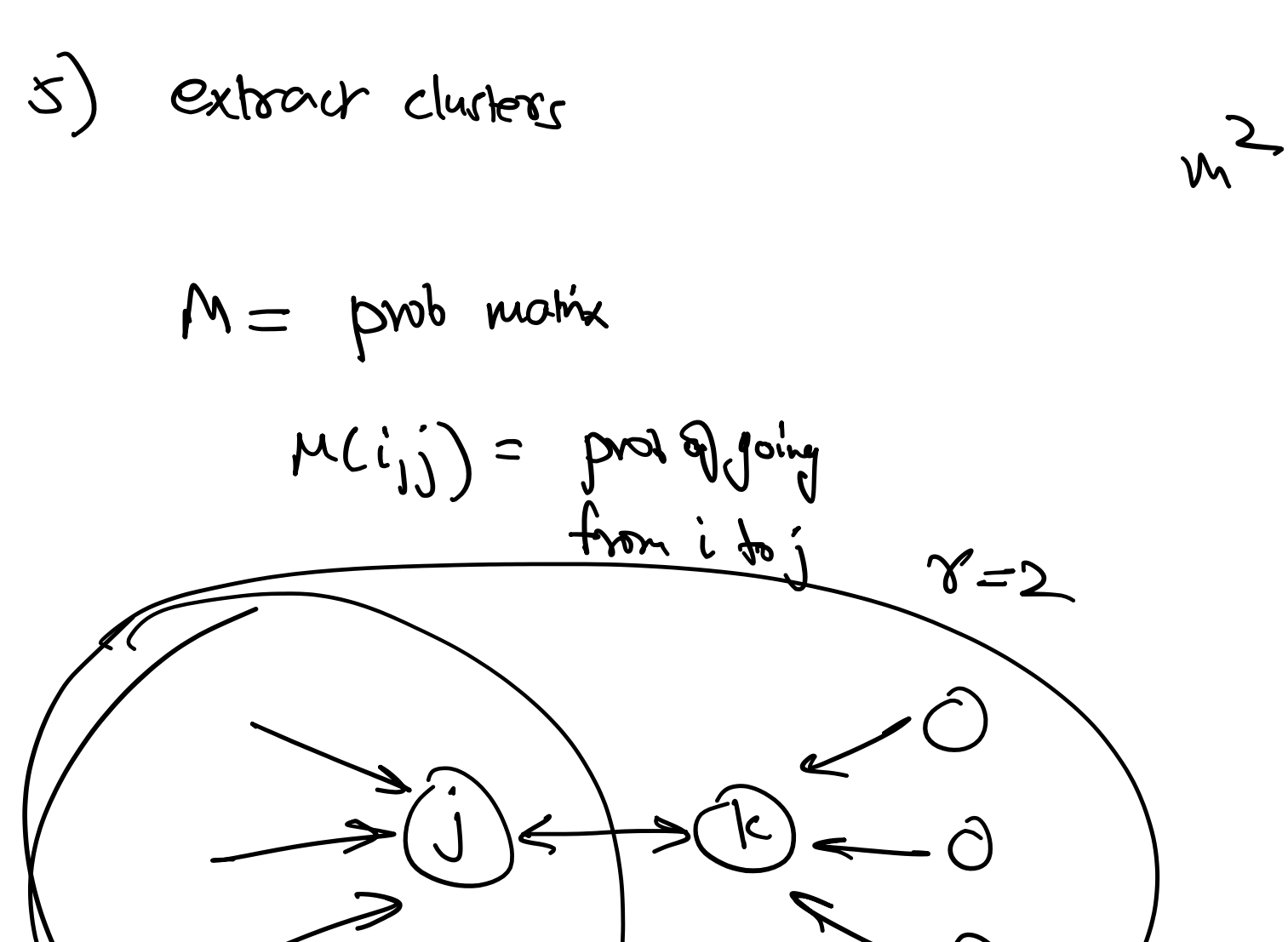
$$M^2 = \begin{bmatrix} 0.16 & 0.84 \\ 0.55 & 0.45 \end{bmatrix}$$

$$\begin{bmatrix} 0.16^2 & 0.84^2 \\ 0.55^2 & 0.45^2 \end{bmatrix}$$

M = prob matrix

M(i,j) = prob of going from i to j

r=2



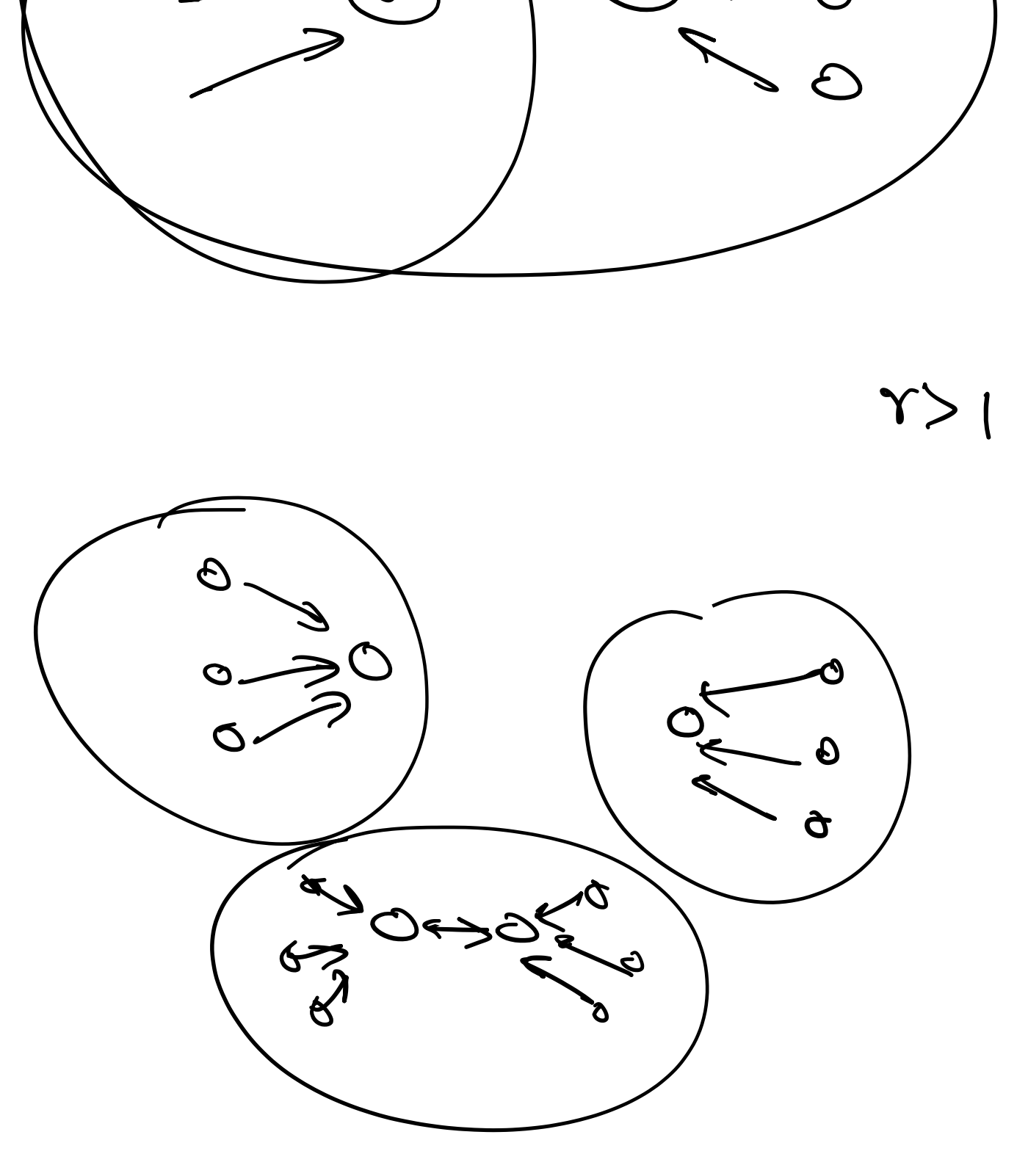
0.02 0.6 0.62

0.36 0.16 0.52

⇓ normalize

$$\begin{bmatrix} 0.1 & 0.9 \\ 0.8 & 0.2 \end{bmatrix}$$

r>1



$\boxed{r}$

r=1 (pure random walk)

directed graph from final M

find all weakly connected components

⇒ Clusters

r>1