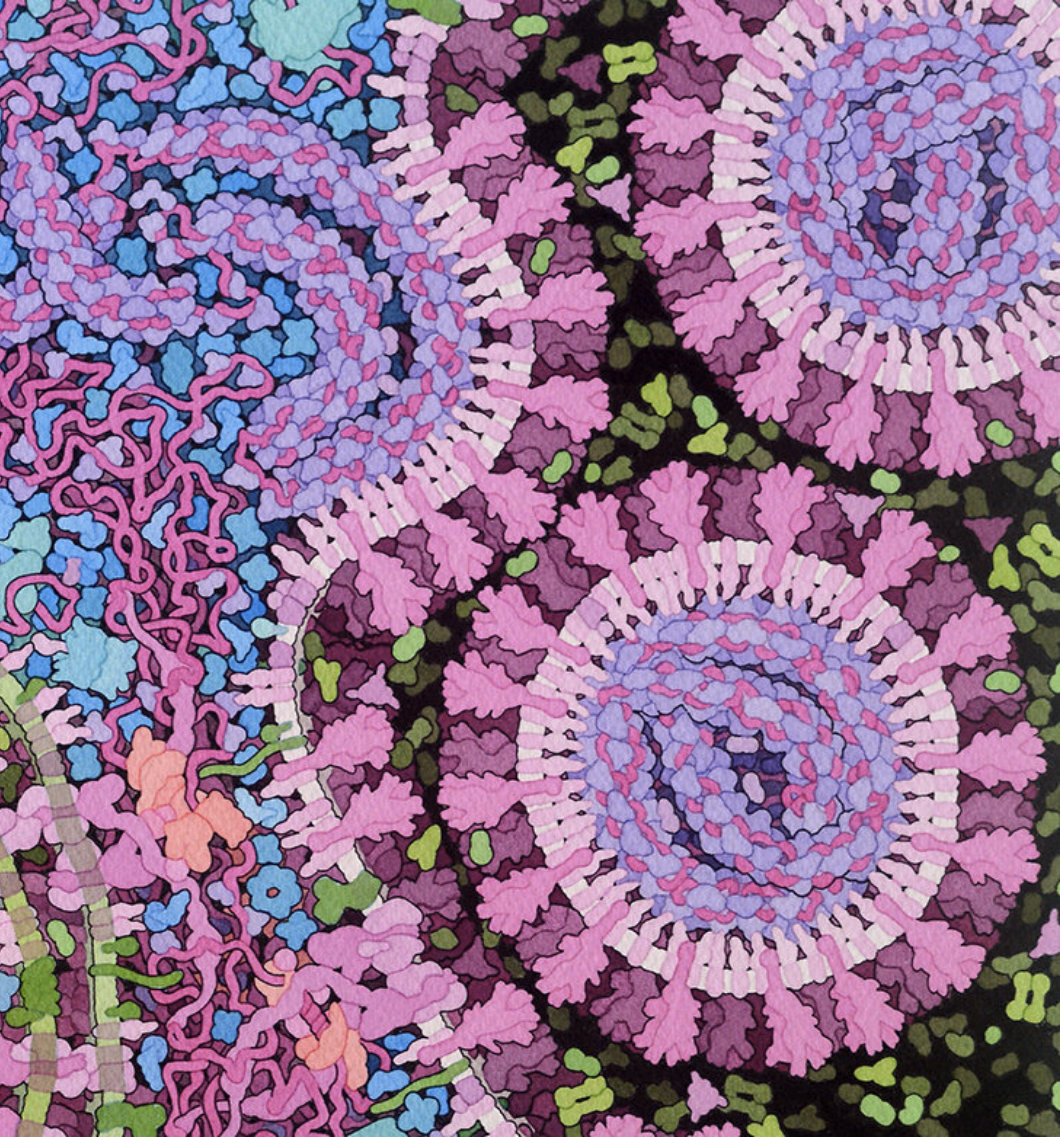


بسم الله الرحمن الرحیم

ژنومیک محاسباتی

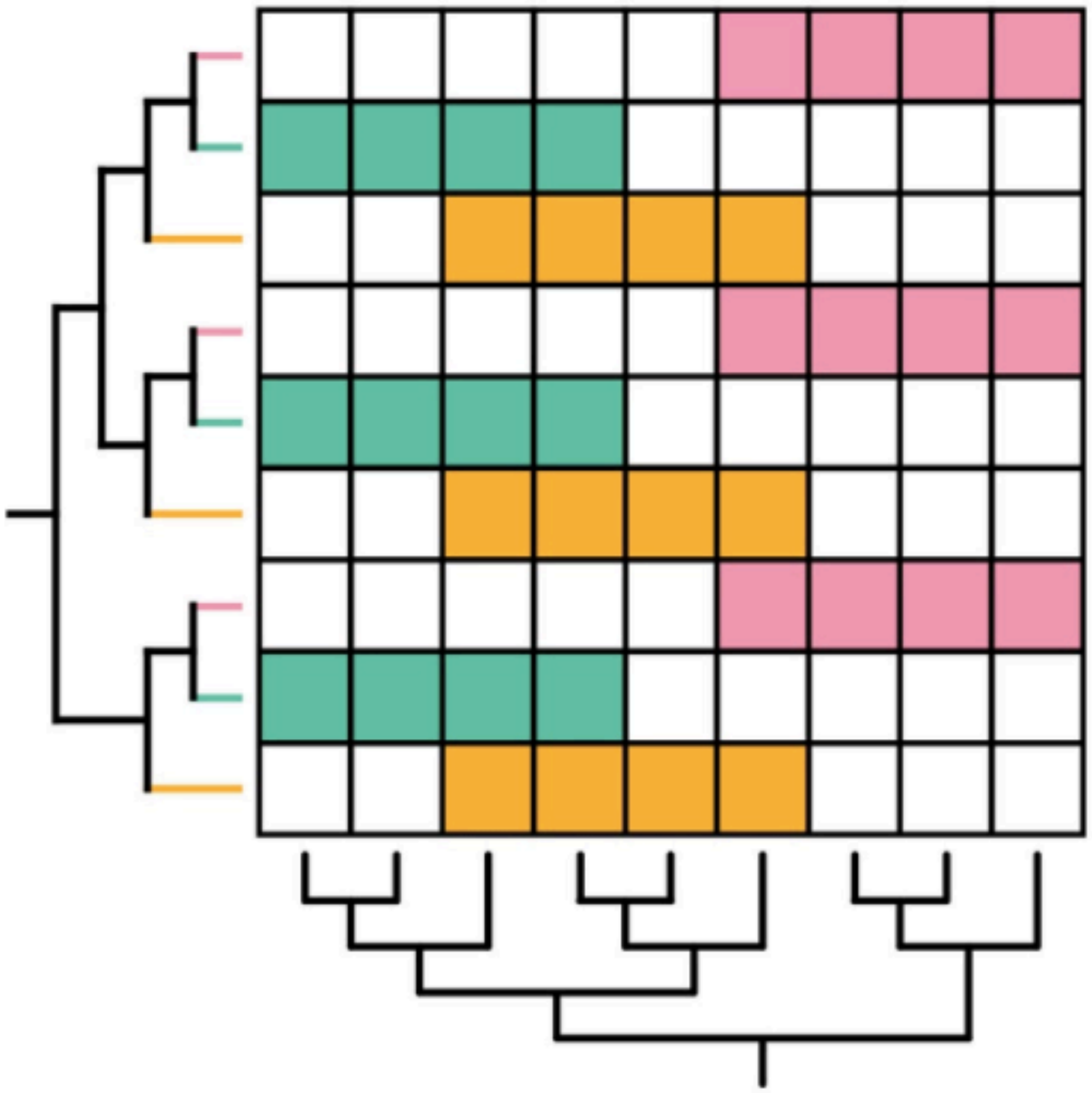
جلسه ۶: بازسازی درخت تبارزایی (۳)

ترم پاییز ۱۴۰۰-۱۴۰۱

بازسازی درخت تبارزایی

- ورودی: ماتریس فاصله
- ورودی: ویژگی‌ها

بازسازی فاصله‌مبنای درخت تبارزایی



مسئله

- ورودی: ماتریس فاصل دوبه‌دو

تعریف (ماتریس فاصله):

Symmetric: $M_{ij} = M_{ji}$ and $M_{ii} = 0$; and

Triangle Inequality: $M_{ij} + M_{jk} \geq M_{ik}$.

مسئله

- ورودی: ماتریس فاصل دوبه‌دو

تعریف (ماتریس فاصله):

Symmetric: $M_{ij} = M_{ji}$ and $M_{ii} = 0$; and

Triangle Inequality: $M_{ij} + M_{jk} \geq M_{ik}$.

- خروجی: درخت فیلوژنی
  - بدون ریشه
  - همه راس‌های میانی درجه ۳

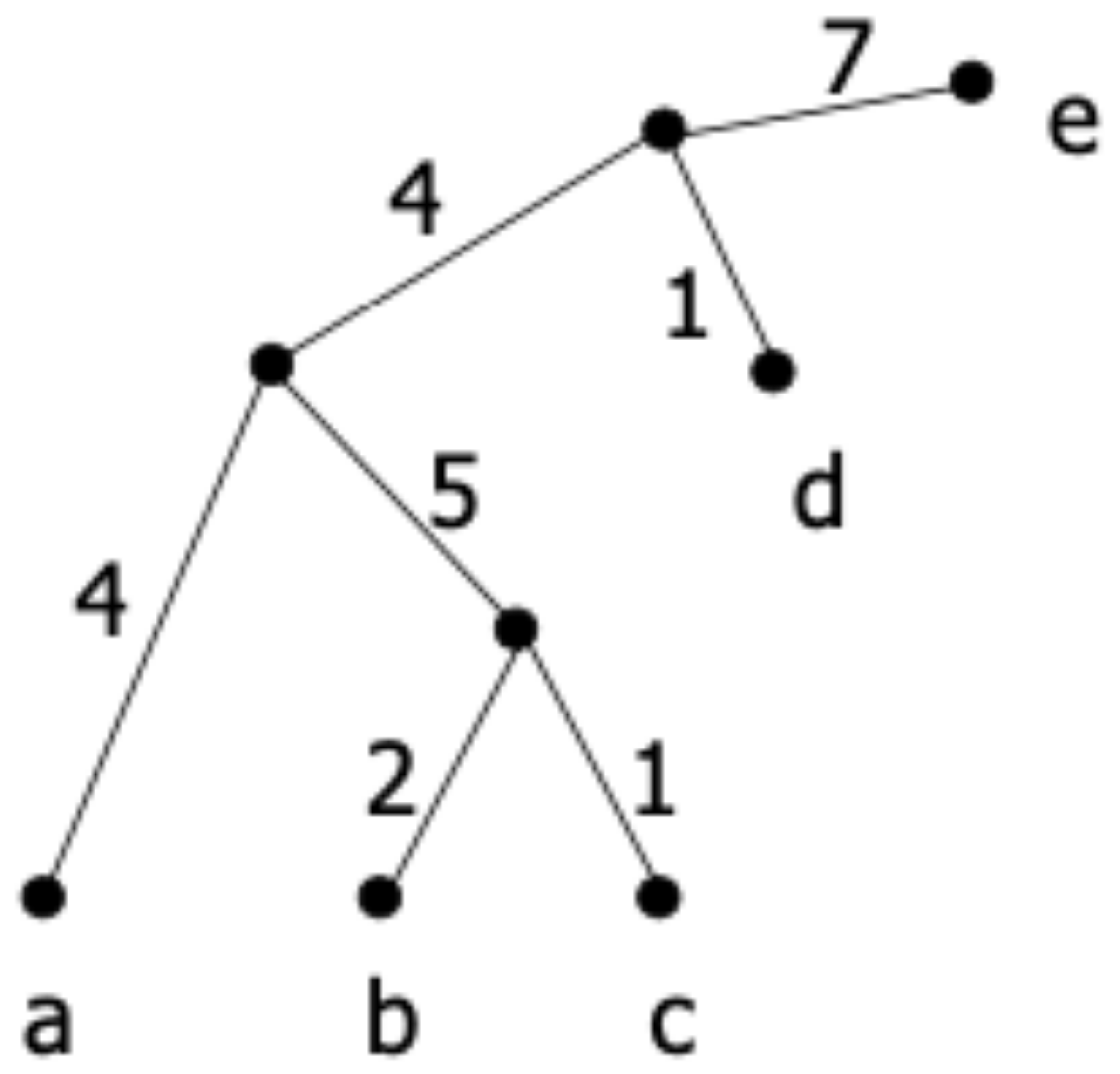
حالت خاص: ماتریس فاصله جمعی

- ماتریس M، ماتریس فاصله جمعی است اگر و فقط اگر
  - یک درخت با وزن یال‌های مثبت وجود داشته باشد
  - فاصله دو برگ = جمع فاصله یال‌های بین آن‌ها

حالت خاص: ماتریس فاصله جمعی

- ماتریس M، ماتریس فاصله جمعی است اگر و فقط اگر
  - یک درخت با وزن یال‌های مثبت وجود داشته باشد
  - فاصله دو برگ = جمع فاصله یال‌های بین آن‌ها

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



شرط لازم و کافی برای جمعی بودن فاصله

ماتریس M، فاصله جمعی است برای درخت یکتای T اگر و فقط اگر

به ازای هر چهارگونه i و j و k و l، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$

شرط لازم و کافی برای جمعی بودن فاصله

ماتریس M، فاصله جمعی است برای درخت یکتای T اگر و فقط اگر

به ازای هر چهار گونه i و j و k و l، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$

شرط چهار نقطه

شرط لازم و کافی برای جمعی بودن فاصله

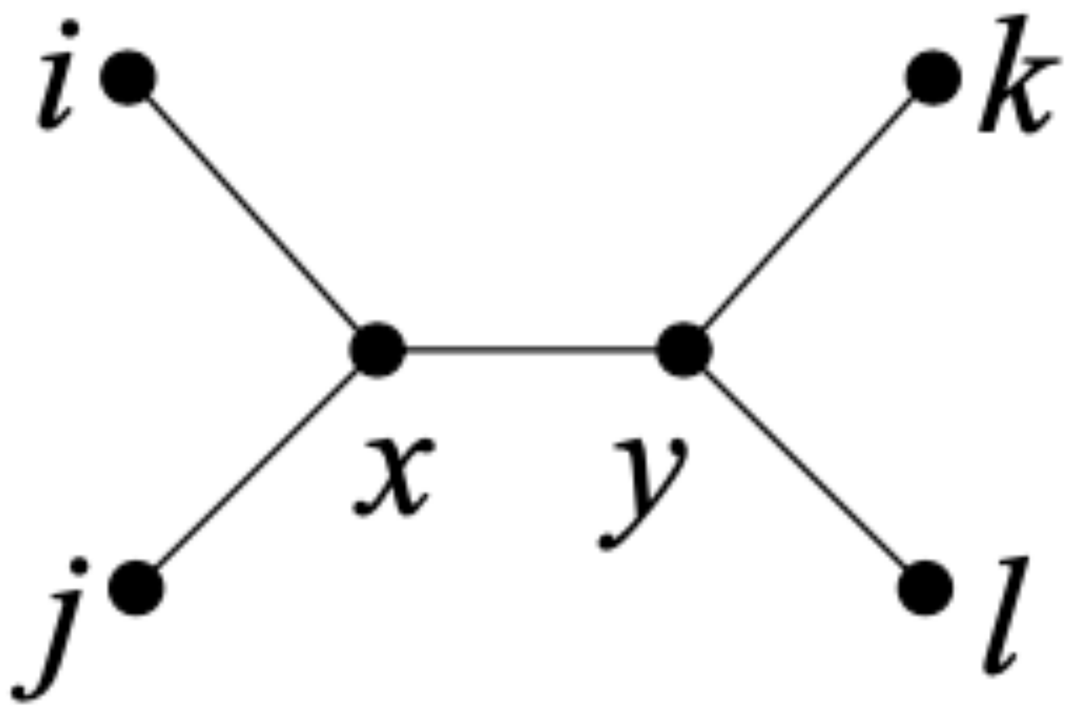
ماتریس M، فاصله جمعی است برای درخت یکتای T اگر و فقط اگر

به ازای هر چهارگونه i و j و k و l، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$

شرط چهار نقطه

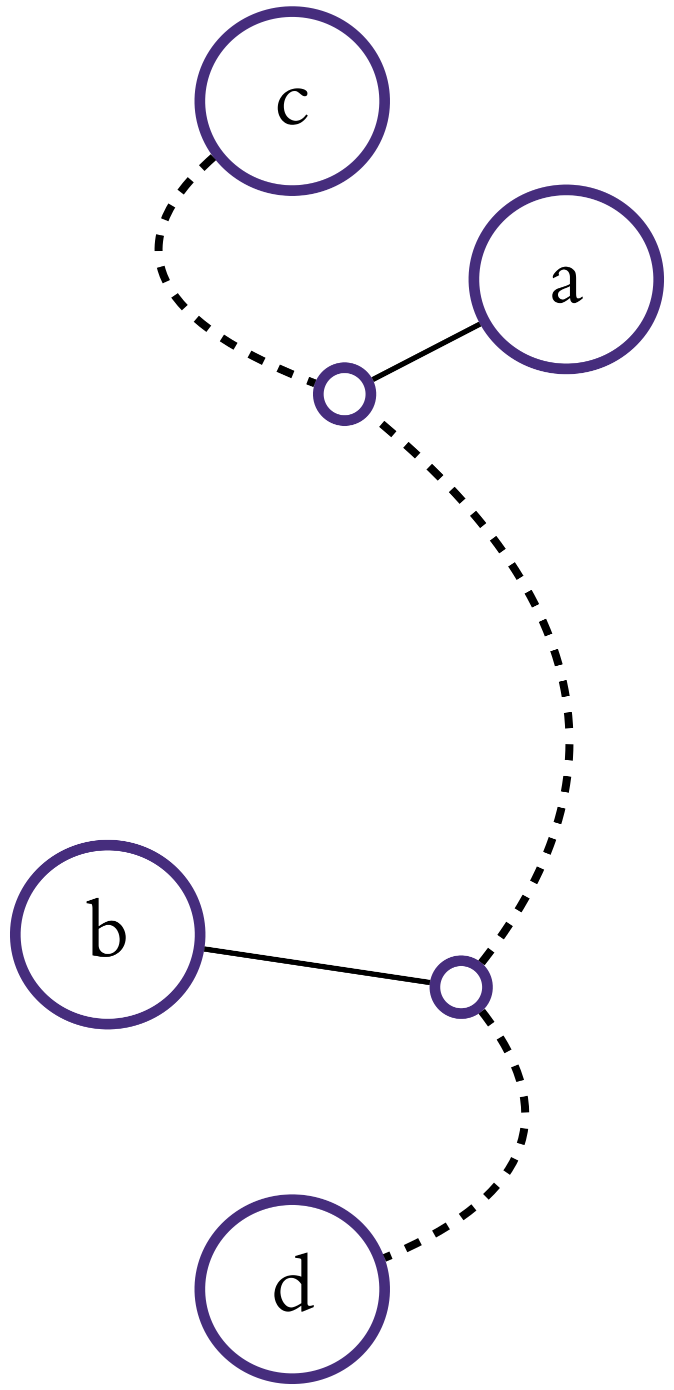
- فاصله جمعی => شرط چهار نقطه



ماتریس M، فاصله جمعی است برای درخت یکتای T اگر و فقط اگر

به ازای هر چهارگونه i و j و k و l، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$

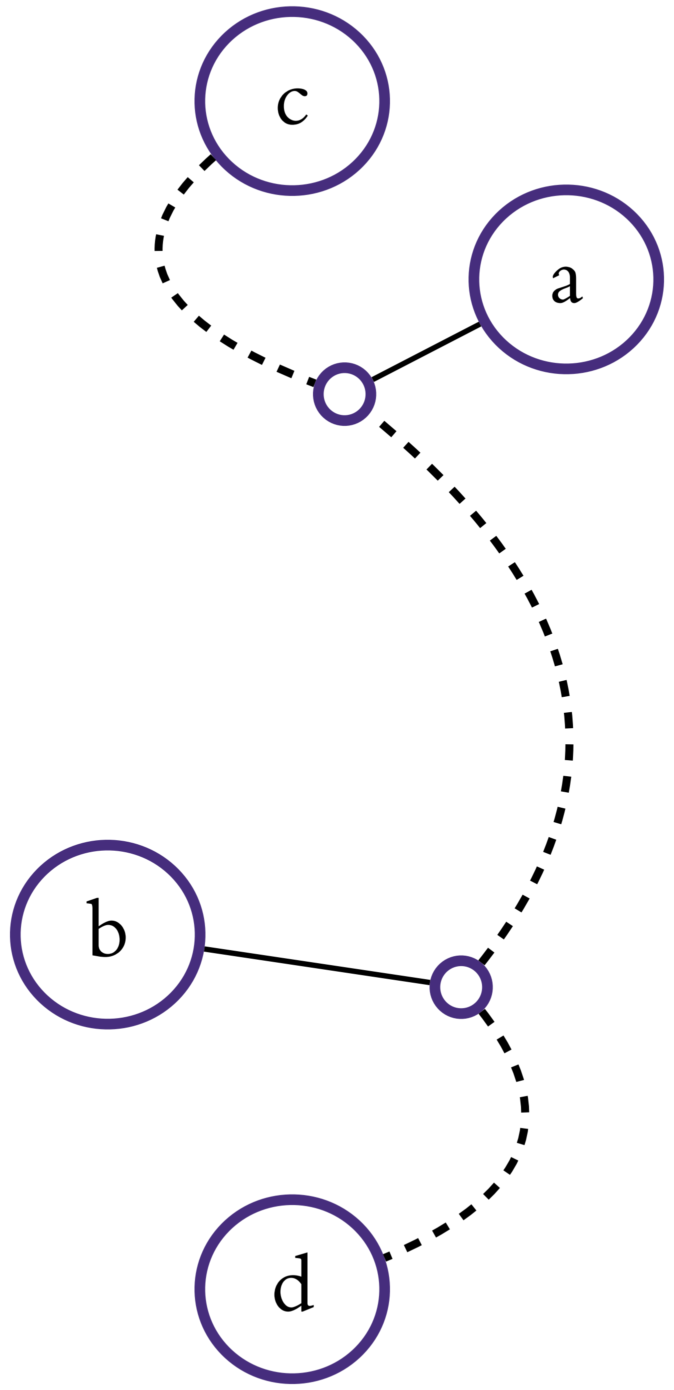


- شرط چهار نقطه => فاصله جمعی
- T: درخت S-a
- T’: درخت S-b

ماتریس M، فاصله جمعی است برای درخت یکتای T اگر و فقط اگر

به ازای هر چهارگونه i و j و k و l، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

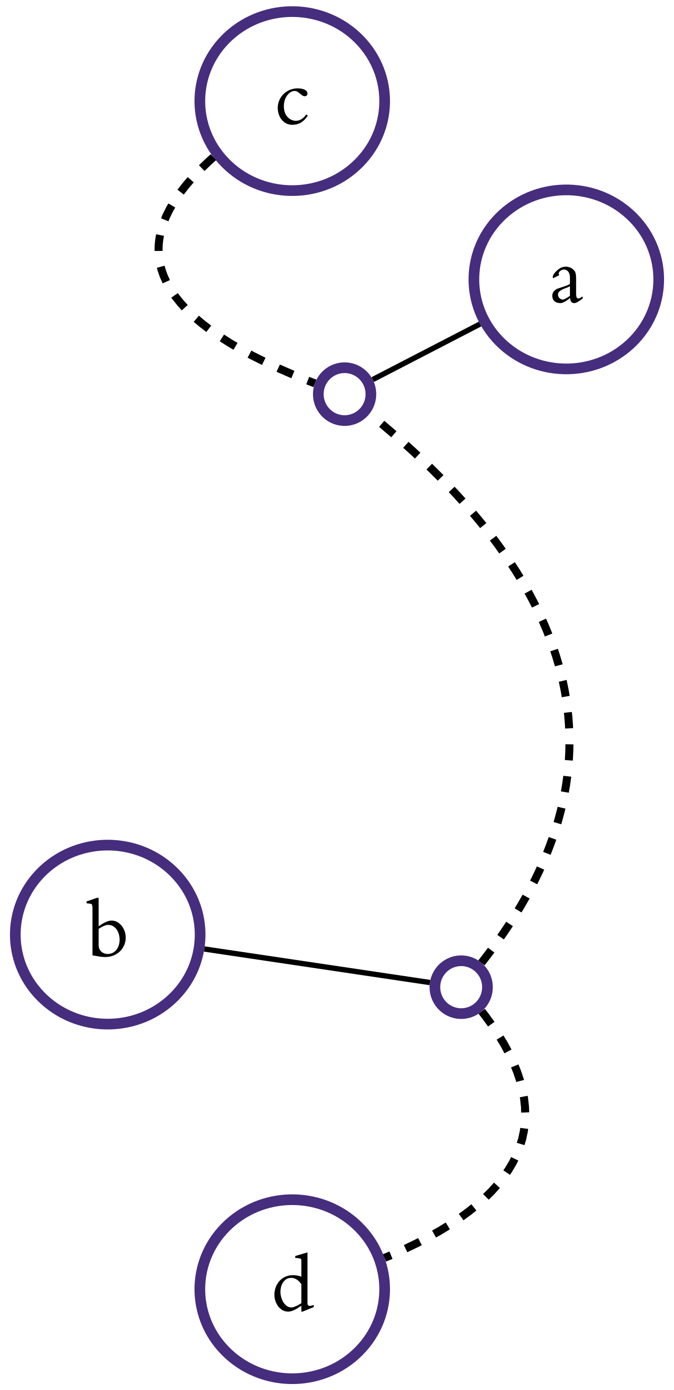
$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$



- شرط چهار نقطه => فاصله جمعی
  - T: درخت S-a
  - T’: درخت S-b
  - S-a-b یک درخت است، پس درخت T + T’ یک درخت یکتای کاندید برای درخت نهایی

ماتریس M، فاصله جمعی است برای درخت یکتای T اگر و فقط اگر
به ازای هر چهارگونه i و j و k و l، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

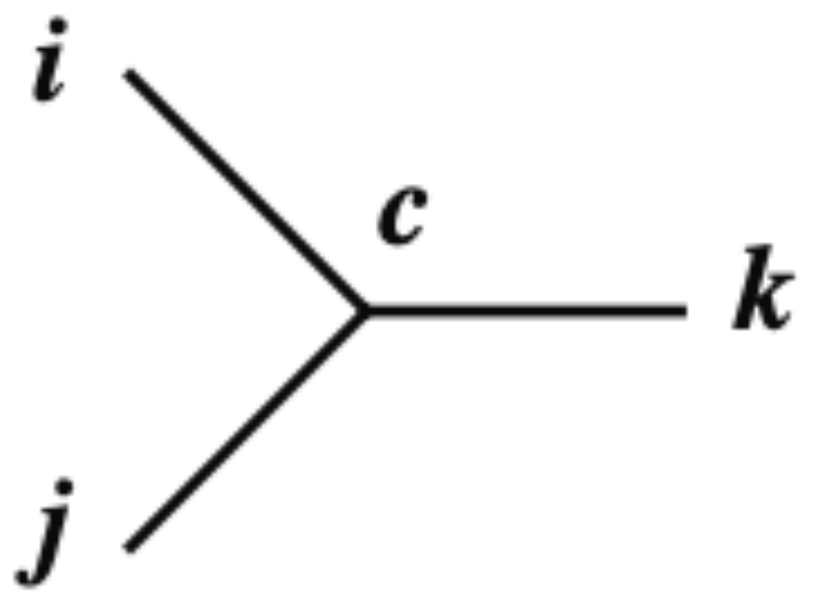
$$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$$



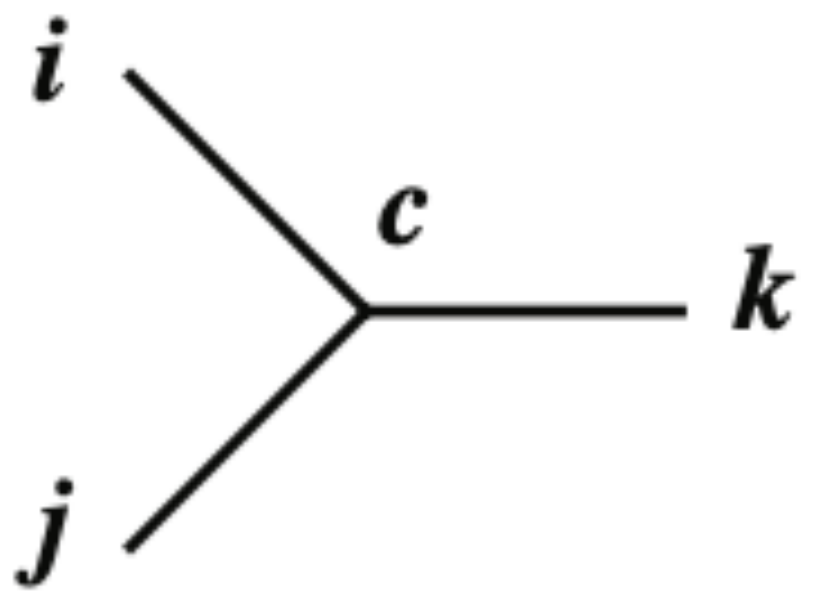
- شرط چهار نقطه <= فاصله جمعی
  - T: درخت S-a
  - T’: درخت S-b
  - S-a-b یک درخت است، پس درخت T + T’ یک درخت یکتای کاندید برای درخت نهایی
  - فاصله‌ها
    - $M_{ad}+M_{bc}-M_{cd} = M_{ab}$

الگوریتم برای ماتریس جمعی

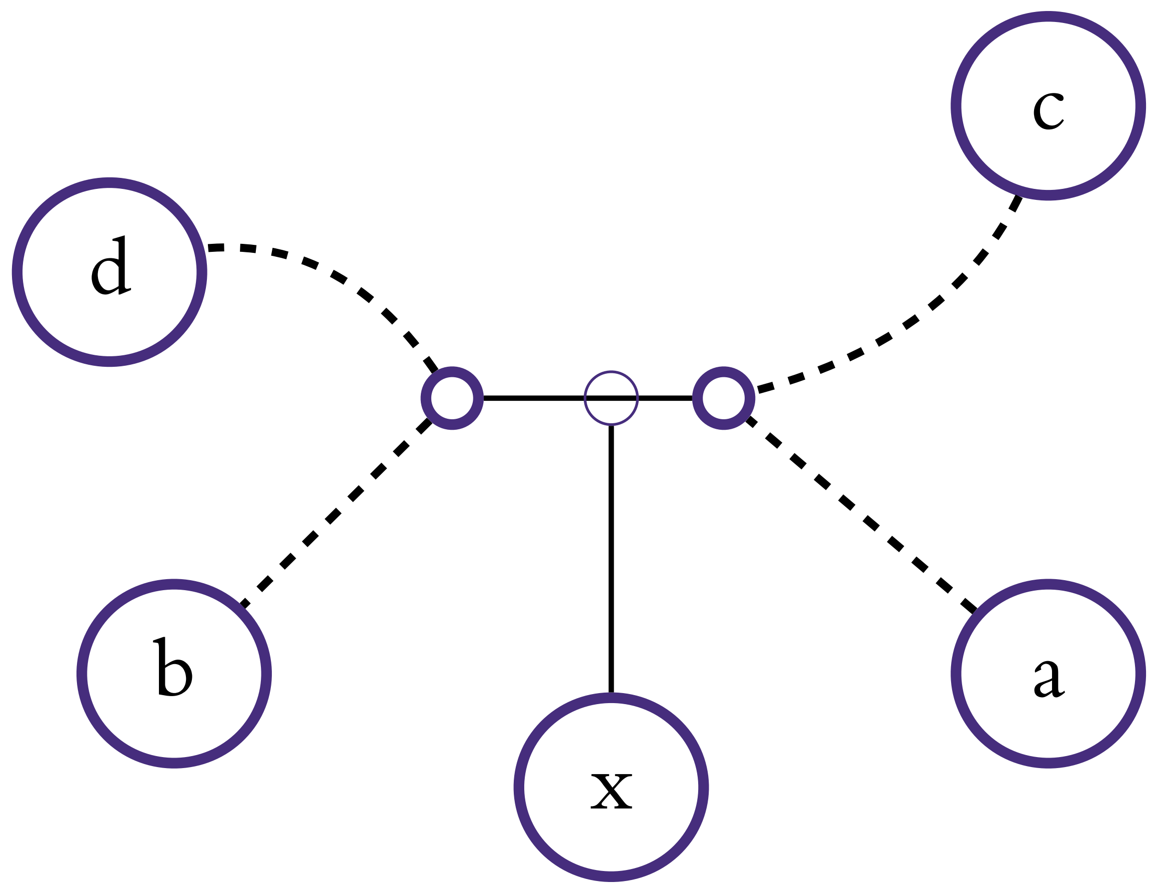
- یک سه‌تایی
  - گونه چهارم دقیقا به یکی‌شان متصل می‌شود.



الگوریتم برای ماتریس جمعی



- یک سه‌تایی
  - گونه چهارم دقیقا به یکی‌شان متصل می‌شود.



- برای راس جدید،
  - درخت یکتاست،
  - یک یال هست که مکان اتصال x به سه‌تا تغییر می‌کند

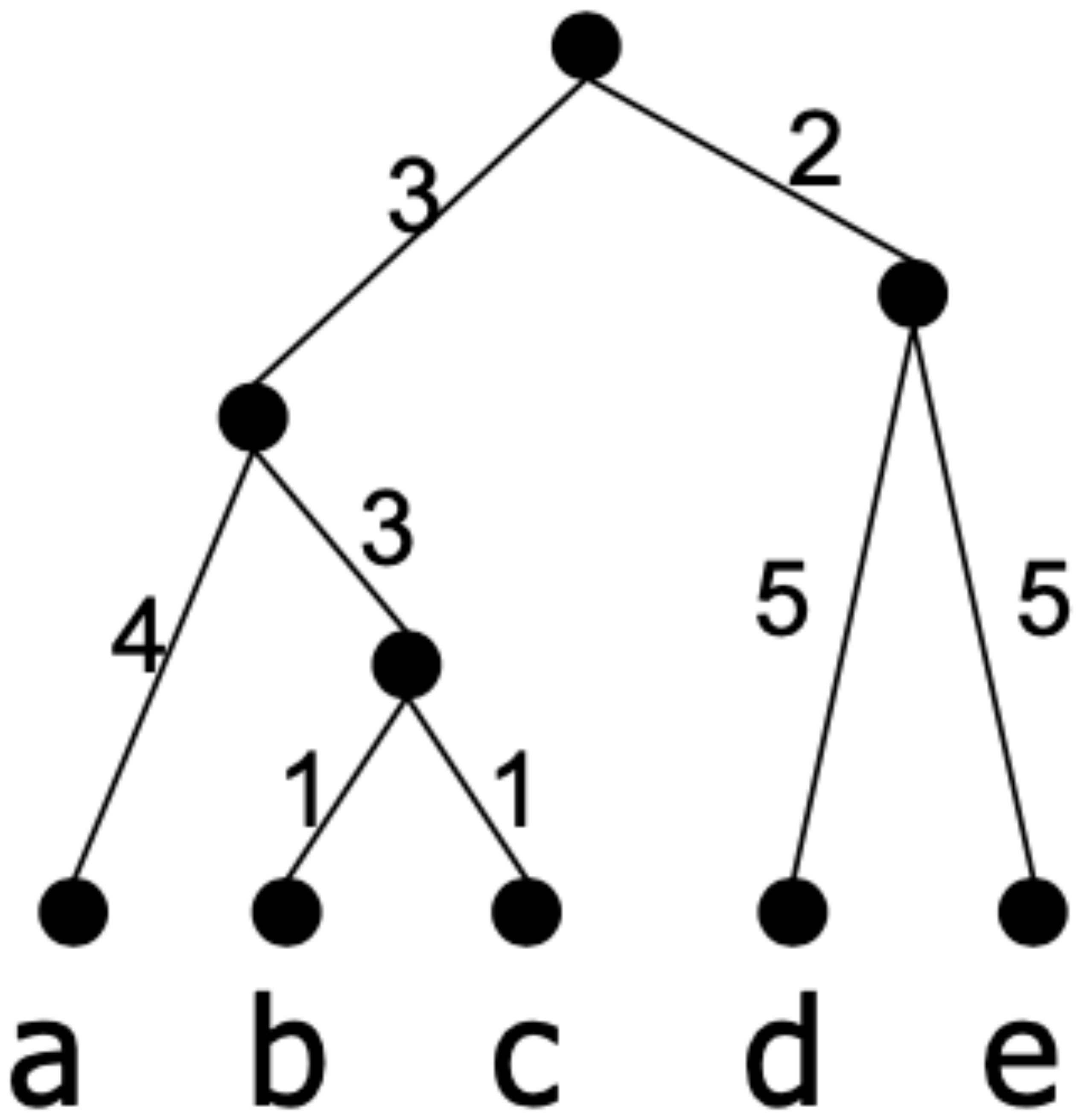
حالت خاص: ماتریس ابرمتریک

- M جمعی باشد
- درخت ریشه‌دار T با وزن‌های مثبت
  - فاصله همه برگ‌ها با ریشه برابر باشد.

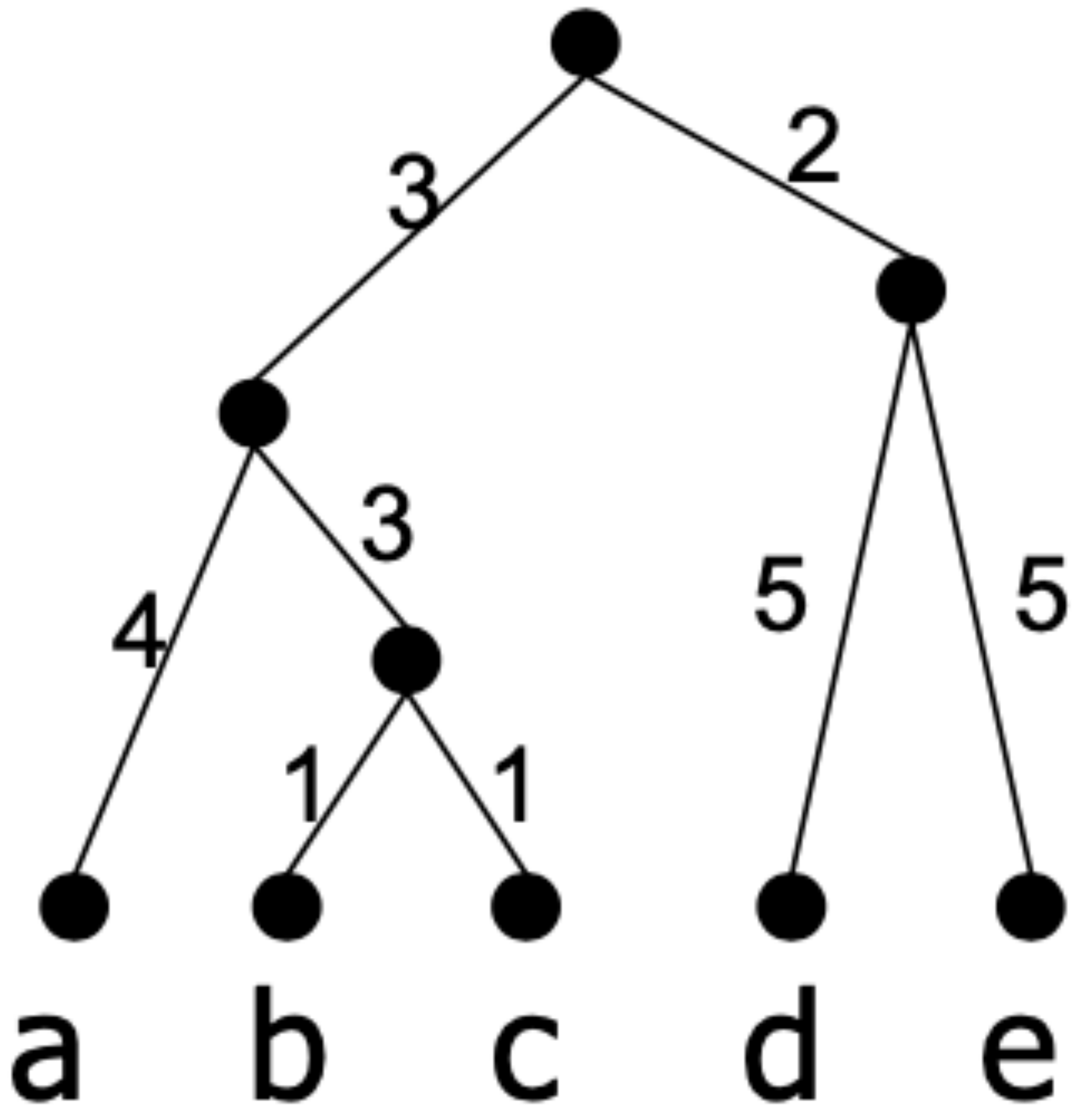
حالت خاص: ماتریس ابرمتریک

- M جمعی باشد
- درخت ریشه‌دار T با وزن‌های مثبت
  - فاصله همه برگ‌ها با ریشه برابر باشد.

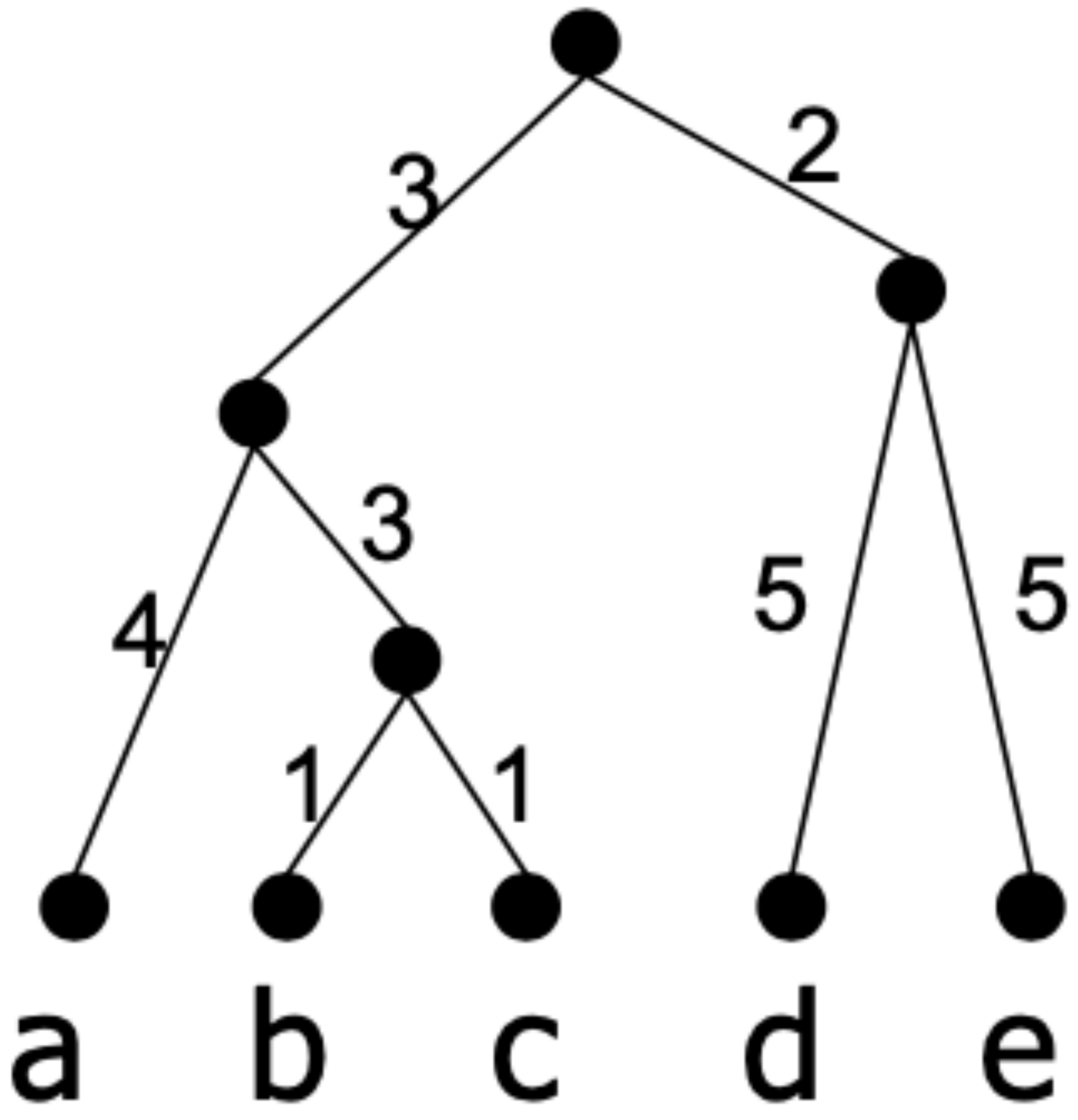
M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



اگر v نزدیک‌ترین جد مشترک برگ‌های i و j باشد، $d(i,v) = d(j,v) = M_{ij}/2$



اگر v نزدیک‌ترین جد مشترک برگ‌های i و j باشد، $d(i,v) = d(j,v) = M_{ij}/2$



- فاصله بین ریشه و v برای هر دو مشترک است:
- $d(i,v) = d(j,v)$

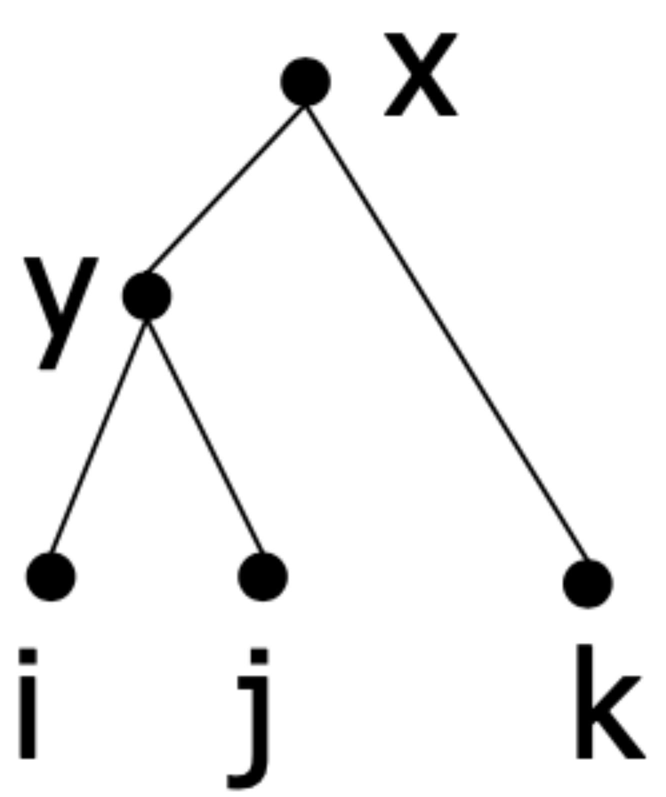
شرط لازم و کافی برای ابرمتریک بودن

ماتریس M، ابرمتریک است اگر و فقط اگر

به ازای هر سه گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه



شرط لازم و کافی برای ابرمتریک بودن

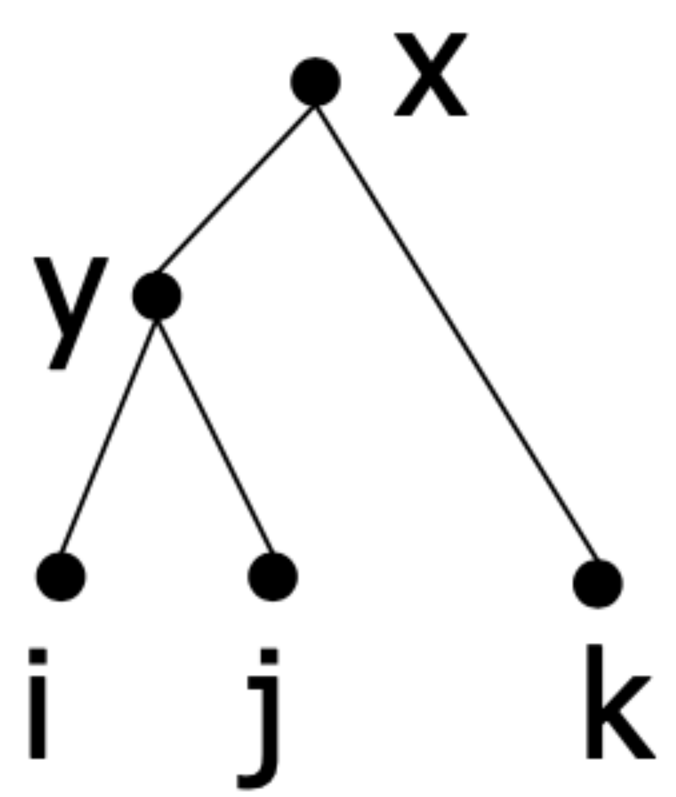
ماتریس M، ابرمتریک است اگر و فقط اگر

به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- ابرمتریک => شرط سه نقطه



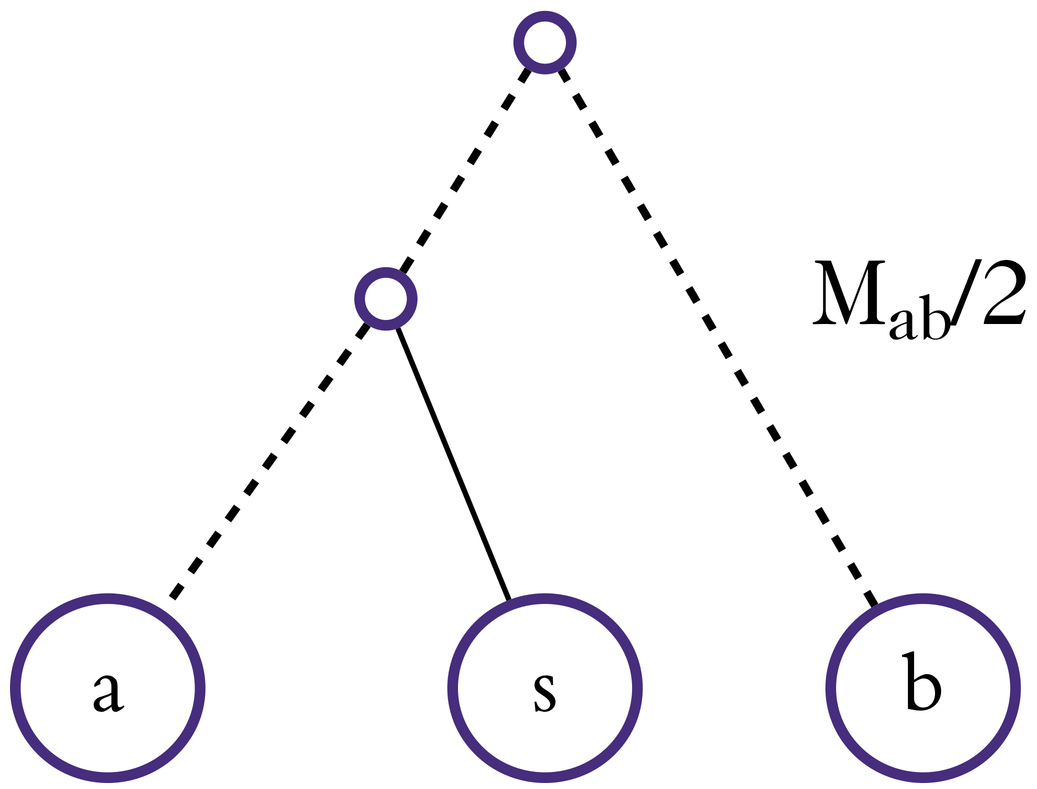
ماتریس M، ابرمتریک است اگر و فقط اگر

به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه => ابرمتریک



ماتریس M، ابرمتریک است اگر و فقط اگر

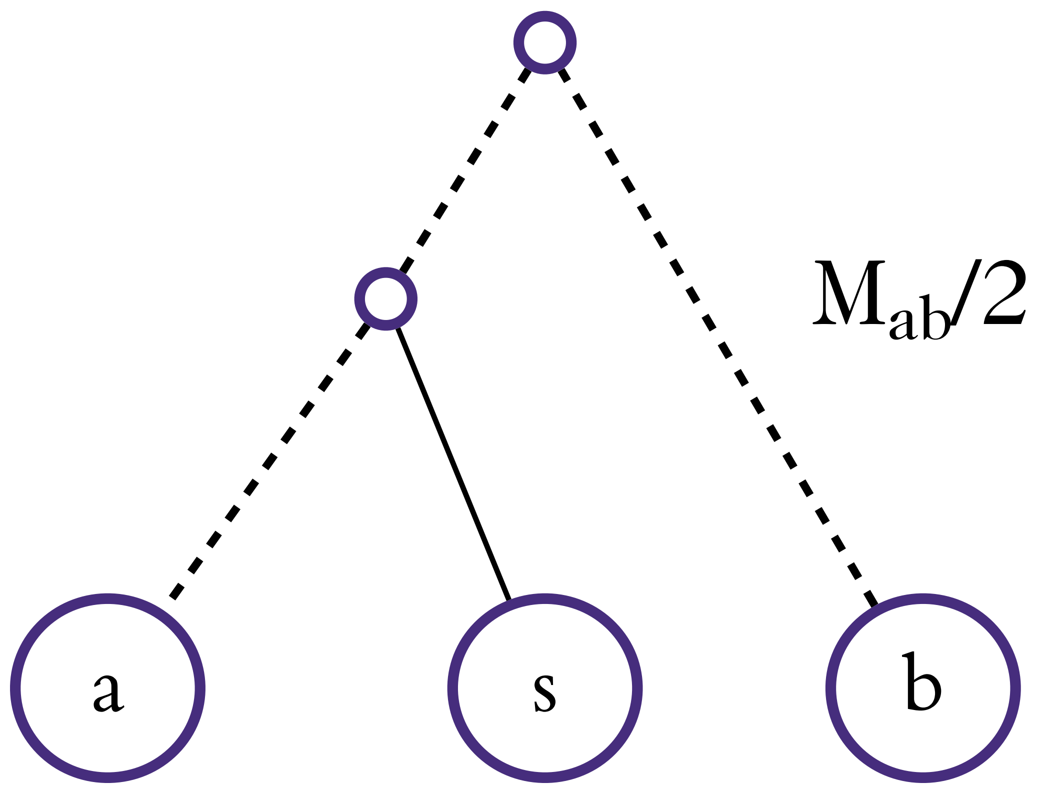
به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه => ابرمتریک
- شرط سه نقطه => شرط چهار نقطه

تمرین!



ماتریس M، ابرمتریک است اگر و فقط اگر

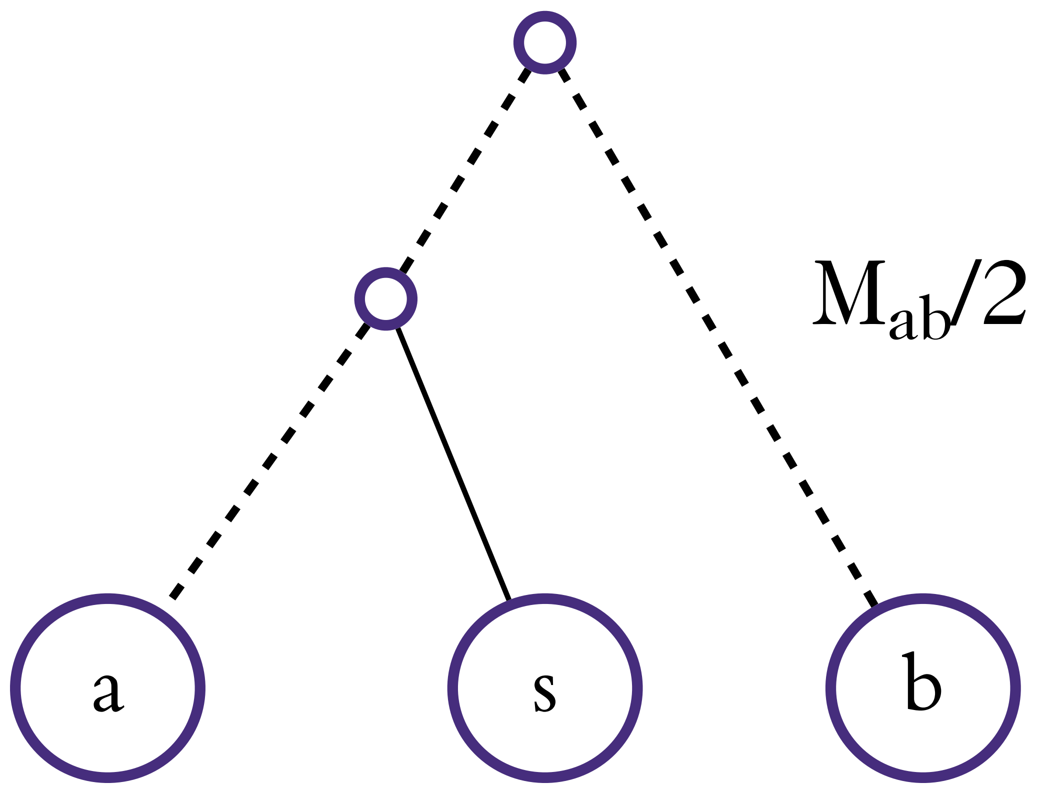
به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه => ابرمتریک
- شرط سه نقطه => شرط چهار نقطه
- درخت بی‌جهت از روی ماتریس M،

تمرین!



ماتریس M، ابرمتریک است اگر و فقط اگر

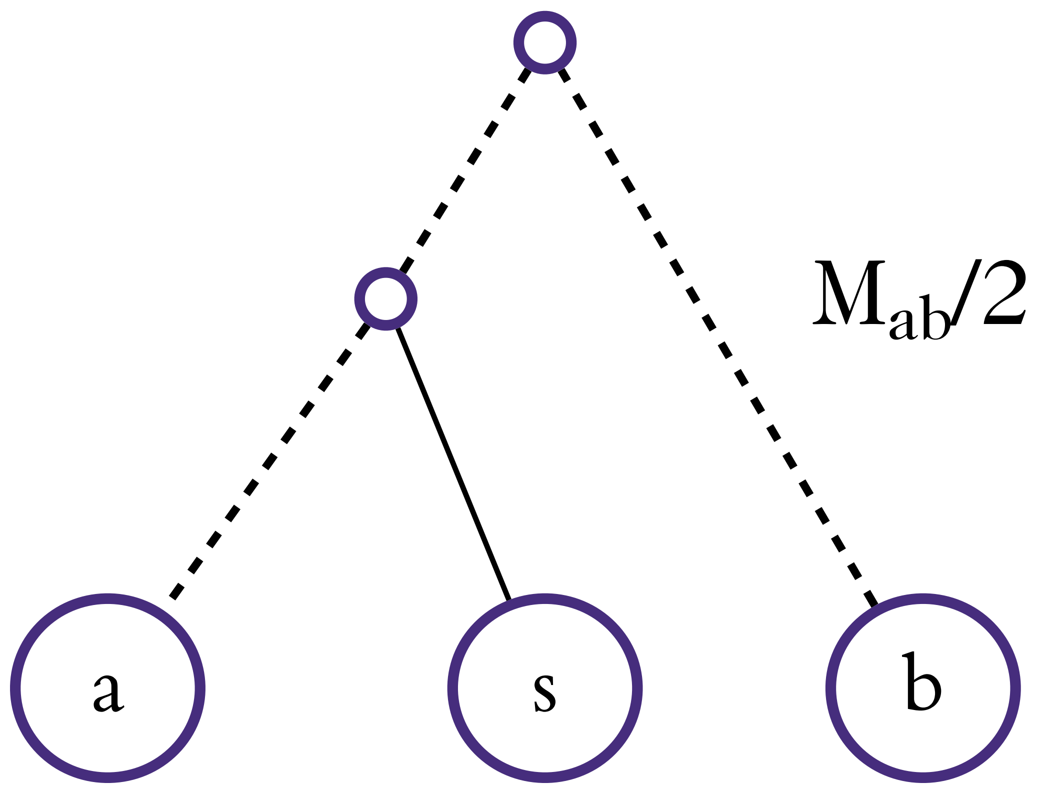
به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه => ابرمتریک
- شرط سه نقطه => شرط چهار نقطه
- درخت بی‌جهت از روی ماتریس M،
- ریشه = r وسط بزرگ‌ترین فاصله

تمرین!



ماتریس M، ابرمتریک است اگر و فقط اگر

به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

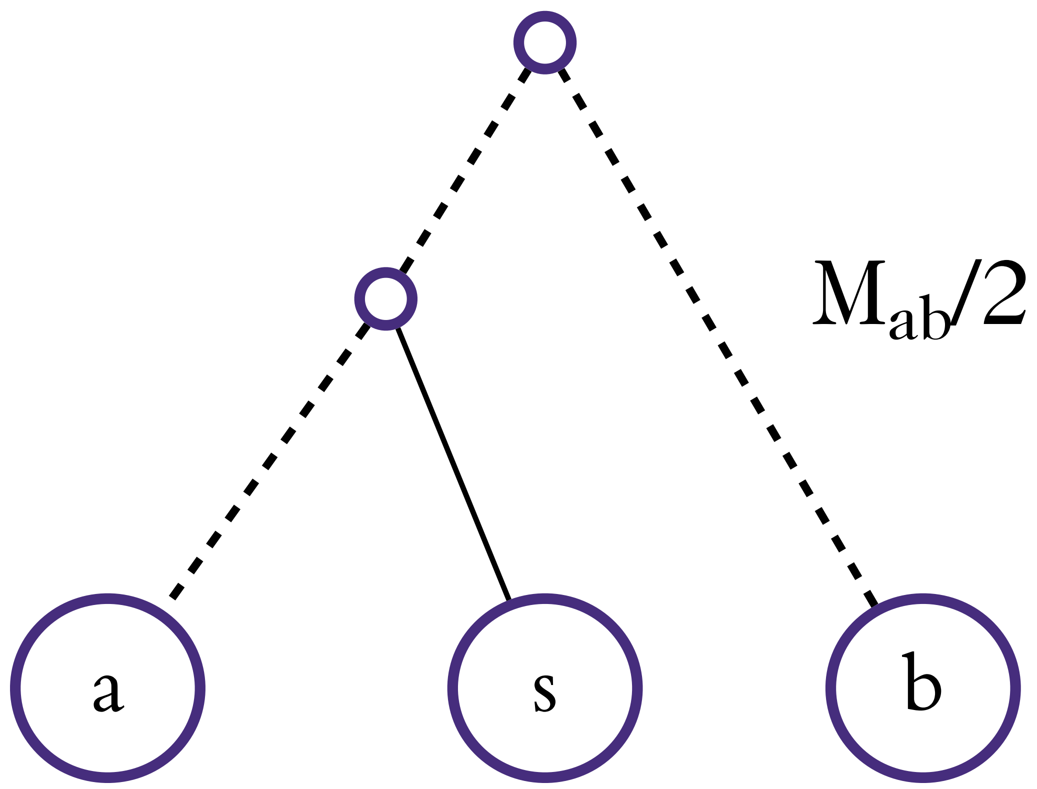
$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه <= ابرمتریک
- شرط سه نقطه <= شرط چهار نقطه

تمرین!

- درخت بی‌جهت از روی ماتریس M،
- ریشه r = وسط بزرگ‌ترین فاصله
- حکم: فاصله ریشه با همه برگ‌ها برابر است



ماتریس M، ابرمتریک است اگر و فقط اگر

به ازای هر سه گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

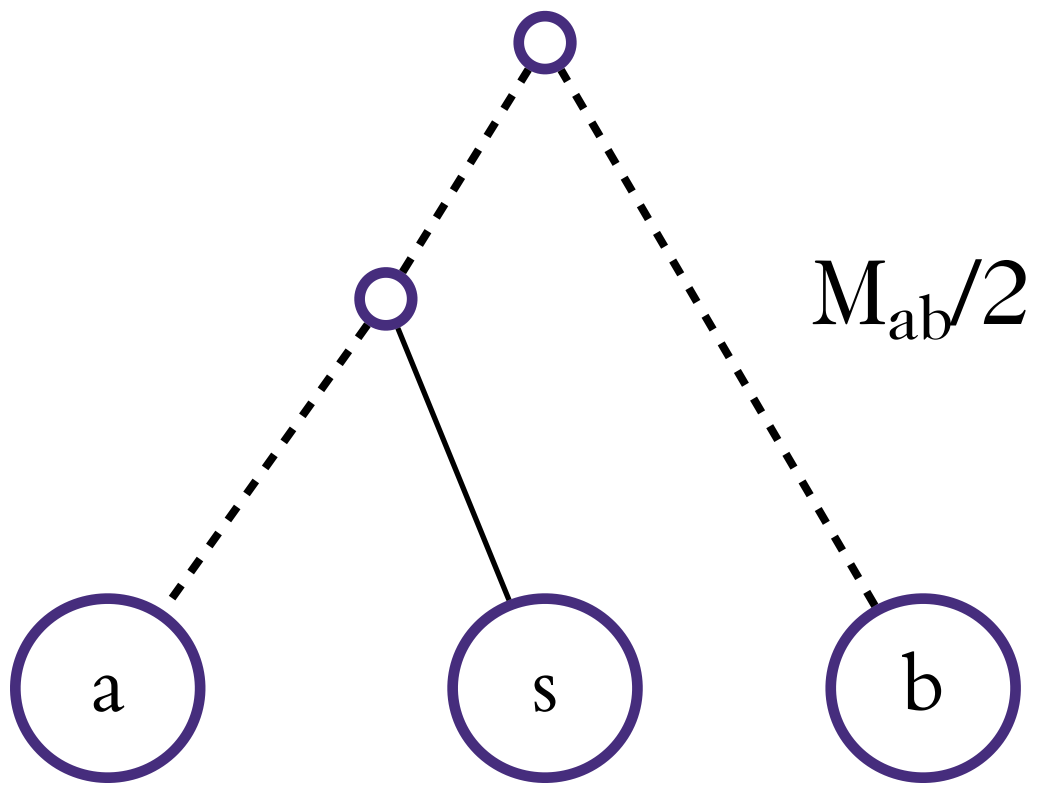
$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه => ابرمتریک
- شرط سه نقطه => شرط چهار نقطه

تمرین!

- درخت بی‌جهت از روی ماتریس M،
- ریشه r = وسط بزرگ‌ترین فاصله
- حکم: فاصله ریشه با همه برگ‌ها برابر است
- فرض: s از بین a و b به a نزدیک‌تر باشد



ماتریس M، ابرمتریک است اگر و فقط اگر

به ازای هر سه‌گونه i و j و k، بتوان نام‌گذاری‌شان را تغییر داد که داشته باشیم

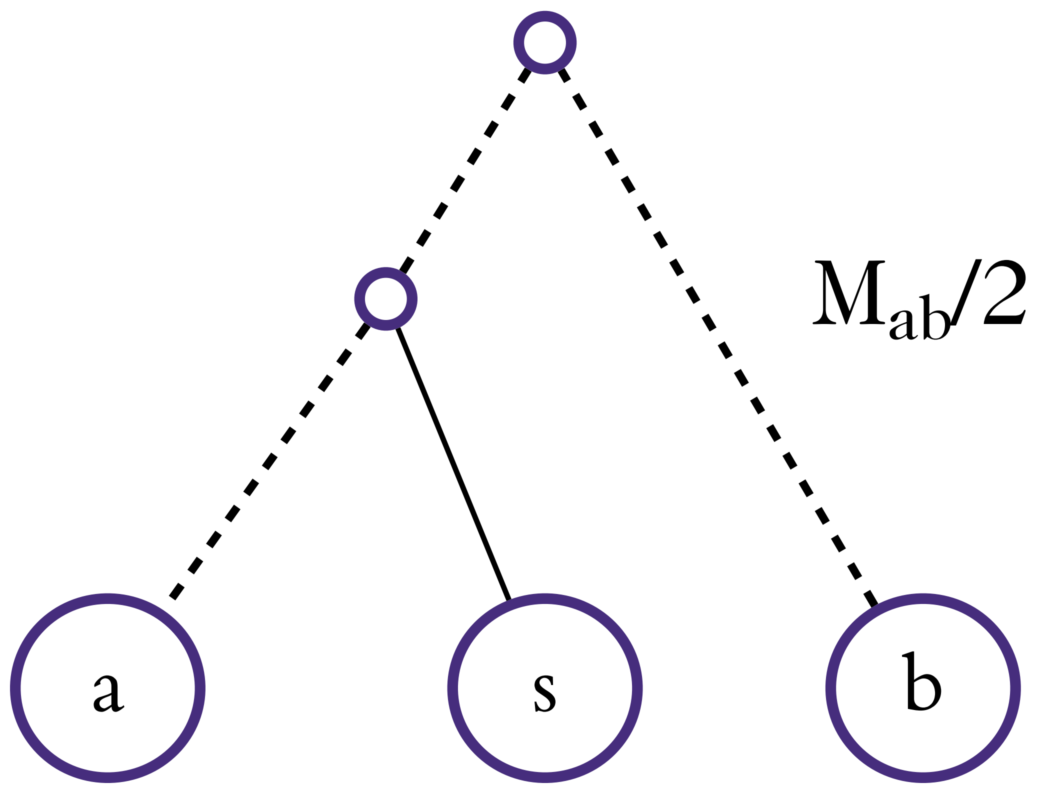
$$M_{ik} = M_{jk} \geq M_{ij}$$

شرط سه نقطه

- شرط سه‌نقطه <= ابرمتریک
- شرط سه نقطه <= شرط چهار نقطه

تمرین!

- درخت بی‌جهت از روی ماتریس M،
- ریشه r = وسط بزرگ‌ترین فاصله
- حکم: فاصله ریشه با همه برگ‌ها برابر است
- فرض: s از بین a و b به a نزدیک‌تر باشد
- $M_{ab} = M_{sb} >= M_{sa}$

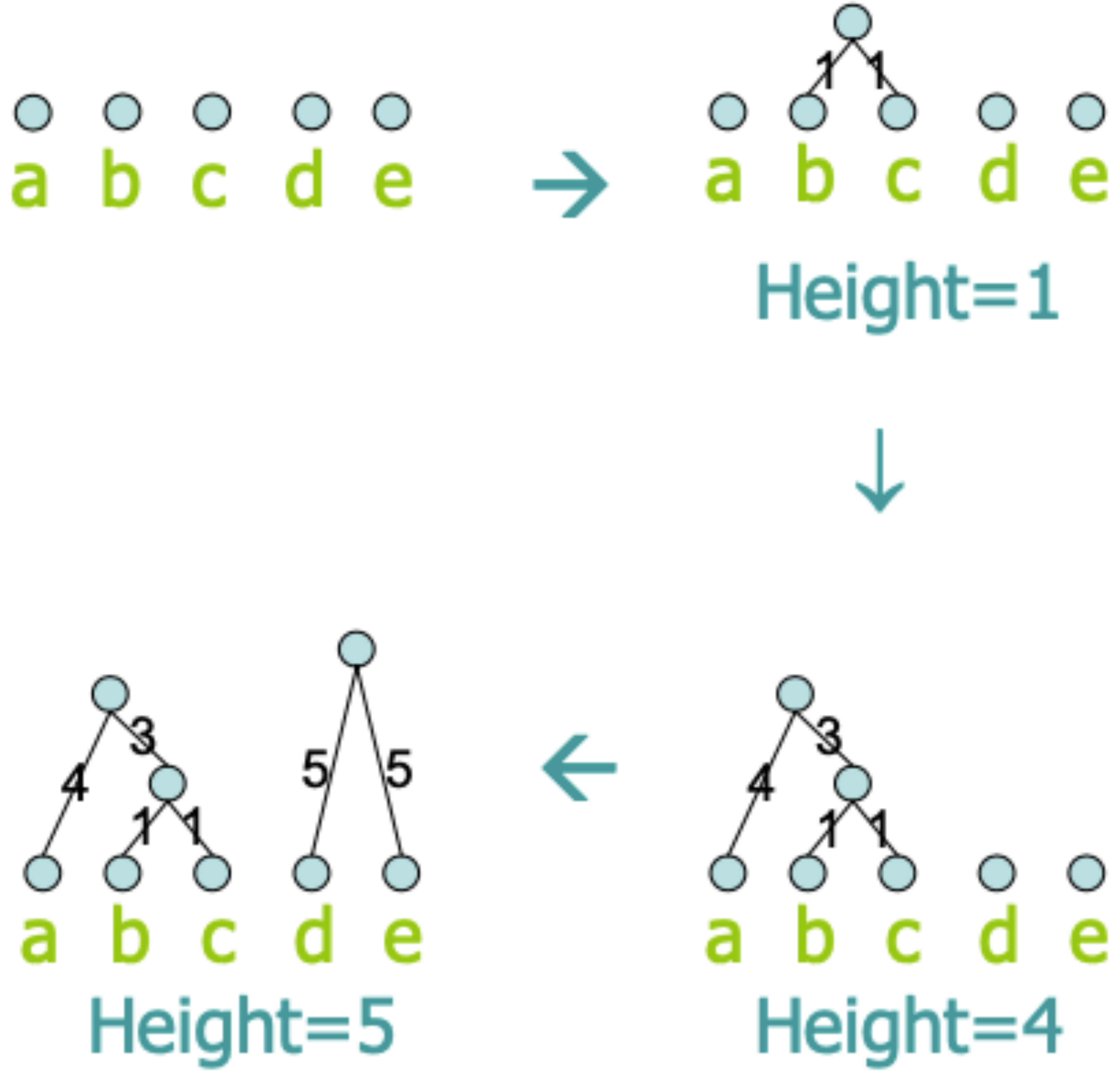


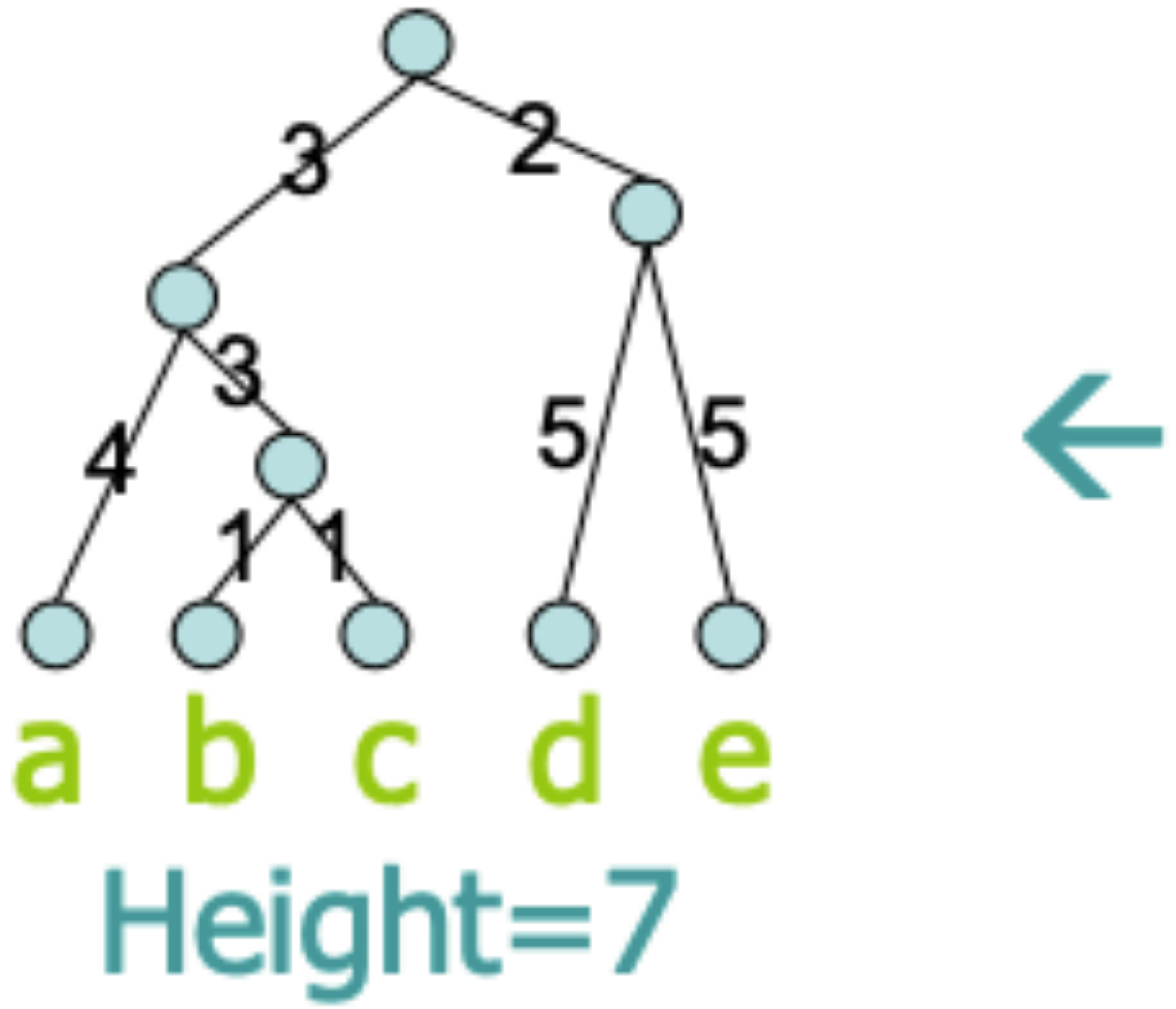
UPGMA algorithm

Set $\mathcal{C} = \{\{c_1\}, \{c_2\}, \ldots, \{c_n\}\}$ where $height(\{c_i\}) = 0$ for $i \in \{1, \ldots, n\}$;
For all $\{c_i\}, \{c_j\} \in \mathcal{C}$, set $dist(\{c_i\}, \{c_j\}) = M_{ij}$;
for $i = 2$ to n **do**
  Determine clusters $C_i, C_j \in \mathcal{C}$ such that $dist(C_i, C_j)$ is minimized;
  Let C_k be a cluster formed by connecting C_i and C_j to the same root;
  Let $height(C_k) = dist(C_i, C_j)/2$;
  Let $d(C_k, C_i) = height(C_k) - height(C_i)$;
  Let $d(C_k, C_j) = height(C_k) - height(C_j)$;
  $\mathcal{C} = \mathcal{C} - \{C_i, C_j\} \cup \{C_k\}$;
  For all $C_x \in C - \{C_k\}$, define $dist(C_x, C_k) = dist(C_k, C_x) = \frac{|C_i| dist(C_i, C_x) + |C_j| dist(C_j, C_x)}{(|C_i| + |C_j|)}$;
end for

$$dist(C_1, C_2) = \frac{\sum_{i \in C_1, j \in C_2} M_{ij}}{|C_1| \times |C_2|}$$

M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0



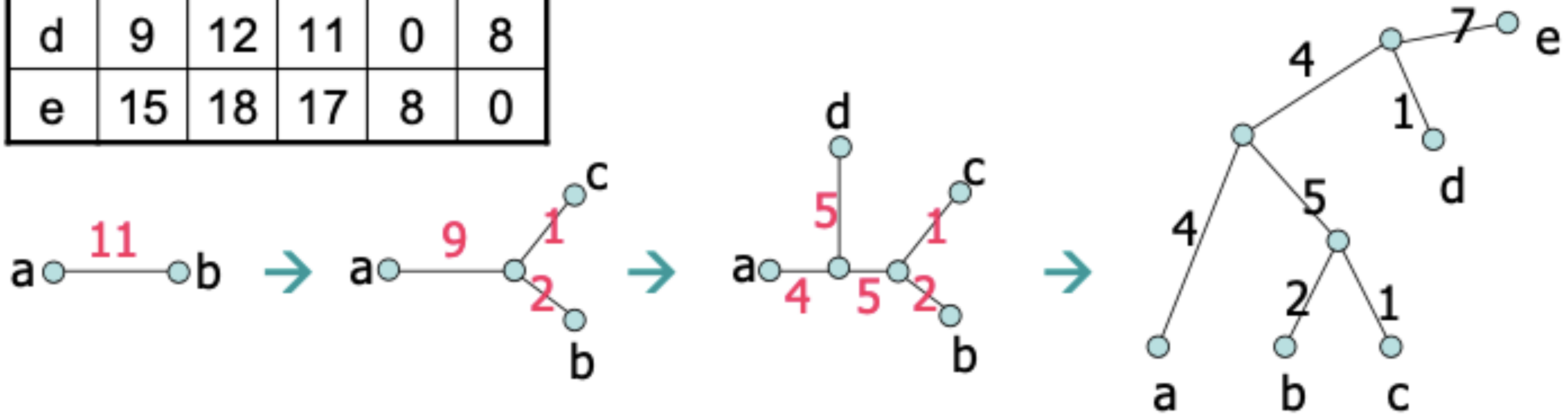


قضیه: UPGMA روی ابرمتریک‌ها درست کار می‌کند!

Neighbor-Joining algorithm

Let $Z = \{\{1\}, \{2\}, \cdots, \{n\}\}$ be the set of initial clusters;
For all $\{i\}, \{j\} \in Z$, set $dist(\{i\}, \{j\}) = M_{ij}$;
for $i = 2$ to n **do**
    For every cluster $A \in Z$, set $u_A = \frac{1}{n-2} \sum_{D \in Z} dist(D, A)$;
    Find two clusters $A, B \in Z$ which minimizes $dist(A, B) - u_A - u_B$;
    Let C be a new cluster formed by connecting A and B to the same root r. Let r_A and r_B be the roots of A and B. The edge weights of (r, r_A) and (r, r_B) are $\frac{1}{2} dist(A, B) + \frac{1}{2}(u_A - u_B)$ and $\frac{1}{2} dist(A, B) + \frac{1}{2}(u_B - u_A)$, respectively;
    Set $Z = Z \cup \{C\} - \{A, B\}$;
    For any $D \in Z - \{C\}$, define $dist(D, C) = dist(C, D) = \frac{1}{2}(dist(A, D) + dist(B, D) - dist(A, B))$;
end for

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



قضیه: NJ روی متریک‌های جمعی درست کار می‌کند!

پشتیبانی از خطا؟