Lecture8 of Geiger & Itai's slide brochure
www.cs.technion.ac.il/~dang/courseDS

קבוצות זרות – בעיית Union/Find

חומר קריאה לשיעור זה:

Chapter 22- Data Structures for Disjoint Sets (440 - 462)

©Geiger & Itai, 2001

קבוצות זרות – בעיית Union/Find

נתון עולם של איברים U, |U|=n. לצורך הנוחות נניח U = {1,2,…,n}.

מבנה נתונים לשמירת קבוצות זרות תומך בפעולות הבאות:

1. Makeset(i) – מחזיר קבוצה חדשה בעלת איבר בודד i.
2. Find(i) – מחזיר את הקבוצה לה שייך האיבר i.
3. Union(p,q) – מאחד את הקבוצות p ו- q, כלומר מחזיר קבוצה חדשה המכילה את איחוד האיברים בקבוצות p,q. (הקבוצות p,q המקוריות חדלות מלהתקיים.)

לדוגמא, נניח ברגע נתון הגענו למצב: {1,3} {5} {2,4,6,7}

p = Find (6)
q = Find (5)
r = Union (p,q)
s = Find (6)

הקבוצה שהמשתנה r שומר היא {5,2,4,6,7}

המשתנים p ו-q אינם מחזיקים יותר קבוצות.
והמשתנה s מחזיק את אותה קבוצה כמו r.

©cs,Technion

דוגמא היפותטית לשימוש

נתונות n ערים {1,…,n}. בתחילה כל הערים מנותקות. כל זמן מה בונים כביש ישיר בין שתי ערים. יש לאפשר את הפעולות הבאות:

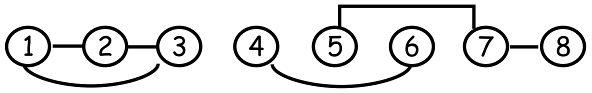
Add-Road(x,y) -- הוסף כביש ישיר בין שתי ערים x ו- y.

Check-Connectivity(x,y) -- קבע האם שתי הערים x ו- y מחוברות במסלול כלשהו.

<u>פתרון</u>: ניצור n קבוצות: {1}, {2},…,{n} ע"י: for (i=1; i ≤ n; i++) Makeset(i) ;

Add-Road(x,y) ממומשת ע"י Union(Find(x),Find(y)).

Check-Connectivity(x,y) מחזירה "אמת" אם ורק אם Find(x)=Find(y)



כביש שהוסף	אוסף הקבוצות הזרות							
	{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
(1,2)	{1,2}		{3}	{4}	{5}	{6}	{7}	{8}
(1,3)	{1,2,3}			{4}	{5}	{6}	{7}	{8}
(2,3)	{1,2,3}			{4}	{5}	{6}	{7}	{8}
(5,7)	{1,2,3}			{4}	{5,7}	{6}		{8}
(4,6)	{1,2,3}			{4,6}	{5,7}			{8}
(7,8)	{1,2,3}			{4,6}	{5,7,8}			

©cs,Technion

פתרון נאיבי ראשון

נשתמש במערך A מטיפוס SET בגודל n ובמונה counter המאותחל ל- 0.
בתא A[i] נשמור את שם הקבוצה אליה שייך האיבר i. לדוגמא: {2,4,6,7} {5} {1,3} מיוצגות ע"י המערך הבא (במימוש זה SET הוא פשוט int):

A

1	2	3	4	5	6	7
4	12	4	12	5	12	12

Makeset(i) ממומשת ע"י A[i]=++counter.

Find(i) ממומשת ע"י A[i].

Union(p,q) ממומשת ע"י הגדלת ה-counter באחד, מעבר על המערך A וכתיבת ערך ה-counter בכל מקום שבו כתוב p או q. הפונקציה מחזירה את הערך של counter.

דוגמא: לאחר Union(p,q) כאשר q=5, p=4, מאוחדות הקבוצות p ו- q ומקבלים:

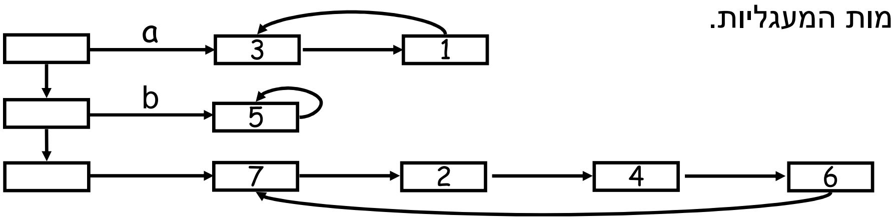
A

1	2	3	4	5	6	7
13	12	13	12	13	12	12

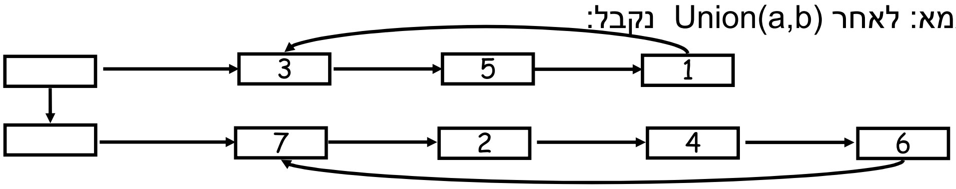
סיבוכיות הזמן של Find ו-Makeset היא O(1) ושל Union היא O(n).

פתרון נאיבי שני

נייצג כל קבוצה כרשימה מעגלית. נחזיק רשימה מקושרת של מצביעים אל הרשימות המעגליות.



Union(p,q) -- ממומשת ע"י איחוד הרשימות המוצבעות ע"י p ו- q. זמן O(1).
במימוש זה SET הוא מצביע לצומת מסוג NODE.
דוגמא: לאחר Union(a,b) נקבל:

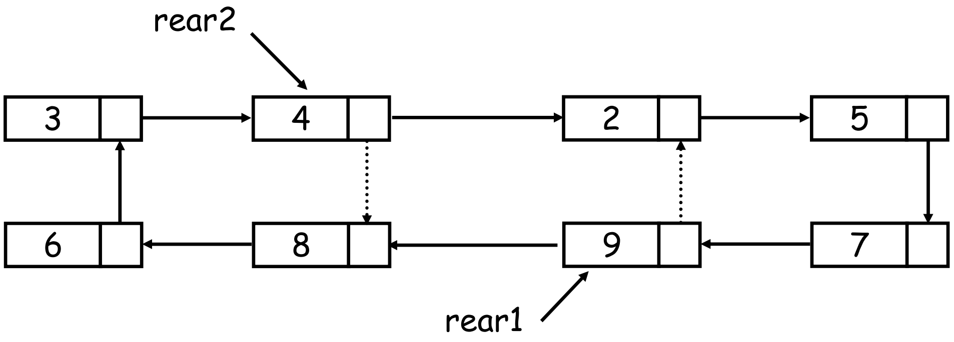


Find(i) ממומשת ע"י מעבר על כל הרשימות עד שנמצא האיבר i. זמן O(n).

Makeset(i) ממומשת ע"י הוספת רשימה עם איבר בודד. זמן O(1).

©cs,Technion

תזכורת: איחוד רשימות מעגליות



©cs,Technion

הערות בהקשר לקידוד בשפת C

כותרות הפונקציות בכל המימושים הניתנים בשיעור זה:

```
SET   Makeset(VALUE  value)
SET   Find(VALUE  value)
SET   Union(SET p,q)

typdef  int  VALUE;
typdef  int  SET;
typdef  NODE  *SET;
```

במימוש הראשון ובמימושים נוספים (typdef int SET;)

במימוש השני ובמימושים נוספים (typdef NODE *SET;)

©cs,Technion

הערות

פתרון	Makeset	Find	Union
נאיבי 1	O(1)	O(1)	O(n)
נאיבי 2	O(1)	O(n)	O(1)

הערה נוספת: בהגדרת הבעיה שמות האיברים נבחרו להיות מספרים שלמים. ניתן להשתמש בשמות כלליים (כלומר מחרוזות תווים). לשם כך ניתן להשתמש במבנה נתונים הממיר ביעילות בין שמות כלליים ומספרים שלמים, למשל Hash table.

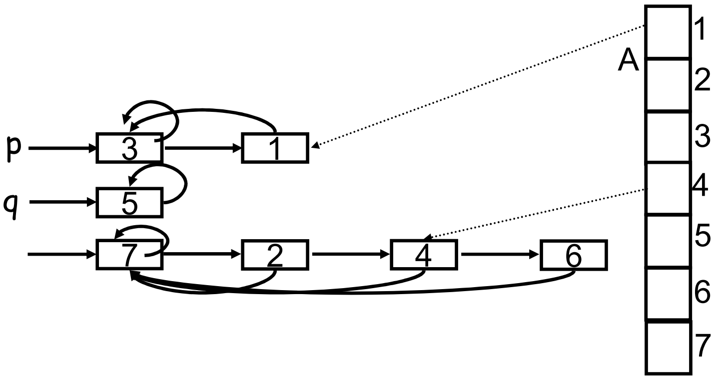
נפתח כעת פתרון ובו הזמן של פעולת Find הוא O(1) והזמן המשוערך (יוגדר בהמשך) של פעולת Union הוא O(log n) .

ולסיום נפתח מבנה נתונים בו הזמן המשוערך לפעולת Find הוא "כמעט קבוע" וזמן פעולת Union הוא O(1).

©cs,Technion

פתרון שלישי

נייצג כל קבוצה כרשימה. כל איבר ברשימה מצביע גם לראש הרשימה וגם לאיבר הבא ברשימה. נחזיק מערך מיפוי איברים A[i] ובו מצביע לאיבר i. קבוצה מוחזרת כמצביע לראש הרשימה המתאימה.



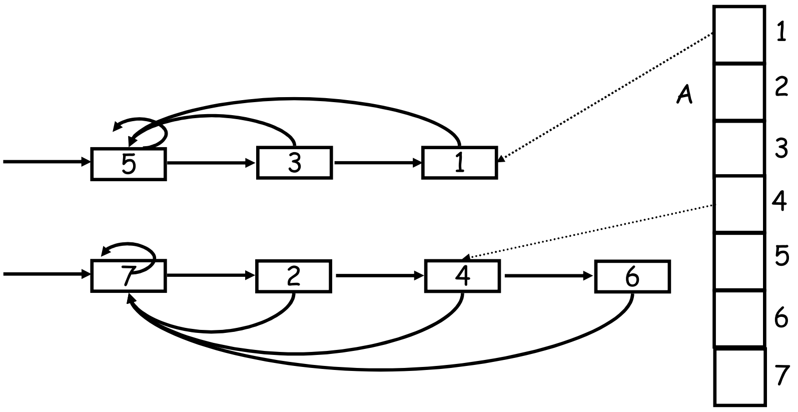
Find(i) -- ממומשת ע"י מעבר על המצביע A[i] והחזרת המצביע לראש הרשימה. זמן O(1).

Union(p,q) -- ממומשת ע"י אחוד הרשימות המוצבעות ע"י p,q לרשימה אחת וכן עדכון מצביעי האיברים לראש הרשימה החדשה. הפונקציה מחזירה את ראש הרשימה החדשה. זמן O(n).

©cs,Technion

דוגמא

דוגמא: לאחר Union(p,q) נקבל:



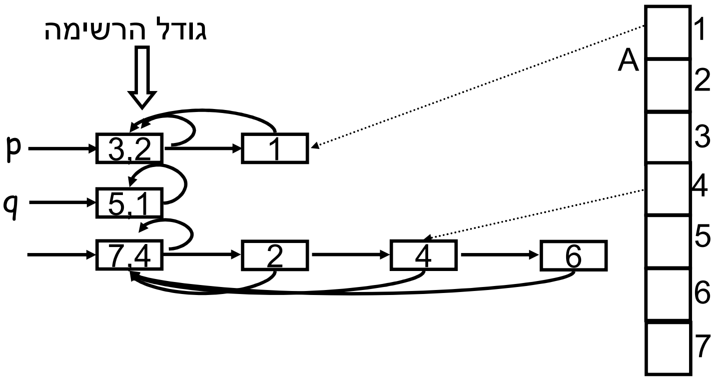
כאשר מאחדים שתי קבוצות יש לשנות באחת הקבוצות את כל המצביעים לכותרת החדשה. ברור שאת שינוי המצביעים עדיף לעשות בקבוצה הקטנה מבין השתיים (בניגוד לדרך שפעלנו בדוגמא זו).

©cs,Technion

פתרון שלישי משופר

שוב נייצג כל קבוצה כרשימה. בנוסף, בראש הרשימה נשמור את גודלה.

Find יתבצע כמו קודם ו- Union יתבצע ע"י הוספת הקבוצה הקטנה לתוך הקבוצה הגדולה. כלומר ראש הרשימה החדשה יהיה ראש הרשימה הגדולה יותר, וכל המצביעים יצביעו עליו.

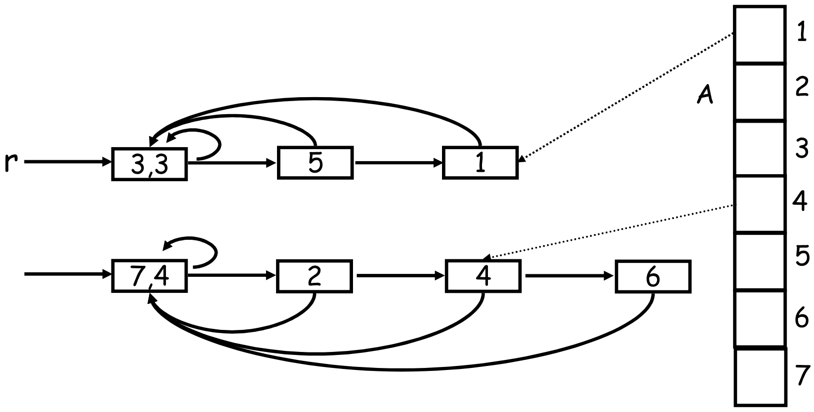


כיצד נבצע r = Union(p,q) ?

©cs,Technion

דוגמא

לאחר r = Union(p,q) נקבל:



שימו לב שהקבוצה {5} שהייתה הקטנה מבין שתי הקבוצות הוספה לתוך הקבוצה הגדולה יותר {3,1} והקבוצה החדשה שנוצרה מוצבעת ע"י r.

©cs,Technion

ניתוח השיפור

טענה א: אם בכל איחוד מוסיפים את הקבוצה הקטנה לגדולה אזי כל איבר x משנה את הקבוצה אליה הוא שייך לכל היותר $\log_2 n$ פעמים כאשר n הוא מספר האיברים במבנה.

הוכחה: בכל פעם שאיבר x משנה את הקבוצה אליה הוא שייך הוא עובר לקבוצה חדשה הגדולה לפחות פי 2 מקבוצתו העכשווית. לאחר לכל היותר $\log_2 n$ מעברים הקבוצה הנוצרת מכילה את כל n האיברים, ולפיכך, שייוכו של איבר x לא ישתנה יותר.

מסקנה: אם בכל איחוד מוסיפים את הקבוצה הקטנה לגדולה אזי הזמן הכולל לביצוע סדרה כלשהי של m פעולות Union/Find/Makeset הוא לכל היותר $O(m \log n)$.

הוכחה: n הוא מספר האיברים ב- U עבורם בוצע Makeset. לאור טענה א, סה"כ הזמן הנדרש לכל פעולות ה- Union הוא $O(n \log n)$ לכל היותר. מתקיים $n \le m$. שאר הפעולות הן פעולות Find/Makeset שזמן כל אחת מהן הוא $O(1)$. לפיכך, סיבוכיות הזמן היא $O(m + n \log n)$. סיבוכיות זמן זו חסומה ע" $O(m \log n)$.

לפיכך כל פעולה לוקחת זמן משוערך (Amortized time)של $O(\log n)$.

©cs,Technion

זמן משוערך

הגדרה: אם m פעולות מתבצעות בתוך זמן M, אזי הזמן המשוערך לכל פעולה מוגדר להיות M/m.

נימוקים להגדרה: כאשר ישנה סדרה ארוכה של פעולות שרובן קלות לביצוע (כגון find במימוש האחרון) וחלקן קשה לביצוע (כגון union באותו מימוש), אז ההשפעה של הפעולות הקשות מתחלקת על סדרת הפעולות כולה. יתכן גם מצב שזמן בצוע הפעולה משתנה כתוצאה מ"עזרה" הניתנת בזמן פעולה קודמת ואז הזמן המשוערך הוא מדד המתחשב בתרומה של פעולה אחת לרעותה.

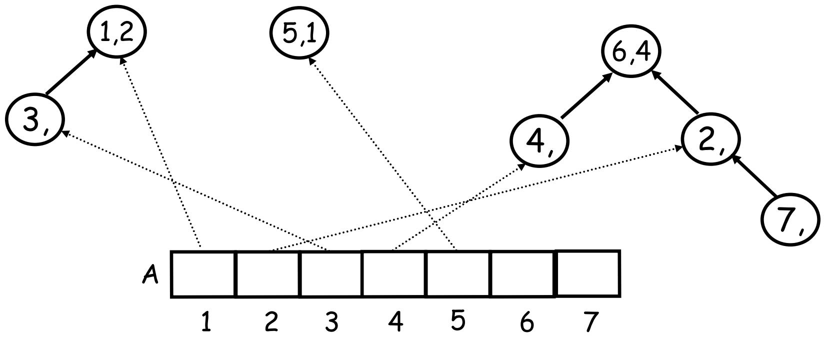
©cs,Technion

מימוש נוסף: עצים הפוכים (Up trees)

לכל קבוצה ניצור עץ הפוך (בנים מצביעים להורה) ובו צומת לכל איבר בקבוצה. שורש העץ יכיל גם את מספר איברי הקבוצה. בנוסף נחזיק מערך גישה לאיברים כמו במימוש השלישי.

בזמן פעולת Union תולים את שורש העץ הקטן יותר מתחת לשורש העץ הגדול יותר ומעדכנים את גודל הקבוצה המאוחדת.

לדוגמא הקבוצות {2,4,6,7} {5} {1,3} מיוצגות ע"י:

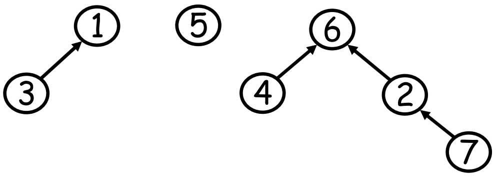


©cs,Technion

מימוש עצים הפוכים בעזרת מערכים

בהנחה שמספר האיברים ידוע מראש, ניתן לייצג את כל הרשימות במערכים ללא שימוש במצביעים. ניתן לעשות זאת גם בכל המימושים הקודמים.

תאור גרפי:



מימוש:

האיבר בשורש משמש מציין לקבוצה

size	2				1	4	
parent	--	6	1	6	--	--	2
elements	1	2	3	4	5	6	7

Find(i) -- ממומש ע"י סיור במעלה העץ. זמן O(h) כאשר h הוא גובה העץ.

Union(p,q) -- ממומש ע"י אחוד העצים ששורשיהן p ו- q כך ששורש העץ הקטן יופנה לשורש העץ הגדול. זמן O(1).

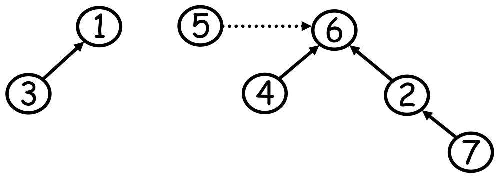
©cs,Technion

דוגמא

מצב התחלתי (כמו בשקף הקודם):

size	2				1	4	
parent	--	6	1	6	--	--	2
elements	1	2	3	4	5	6	7

לאחר Union(5,6):



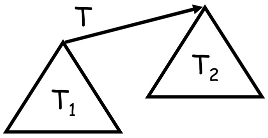
size	2					5	
parent	--	6	1	6	6	--	2
elements	1	2	3	4	5	6	7

זמן לפעולת Union הוא O(1).

©cs,Technion

ניתוח גובה העץ

טענה: עבור עץ T בעל גובה h שנבנה מאיחודים של קבוצות קטנות לתוך גדולות, מתקיים $|T| \geq 2^h$.



הוכחה: באינדוקציה על h. עבור $h = 0$ ישנם 2^0 צמתים.

יהי T עץ בגובה h בעל מספר צמתים מינימלי. נבחן את פעולת האיחוד האחרונה:

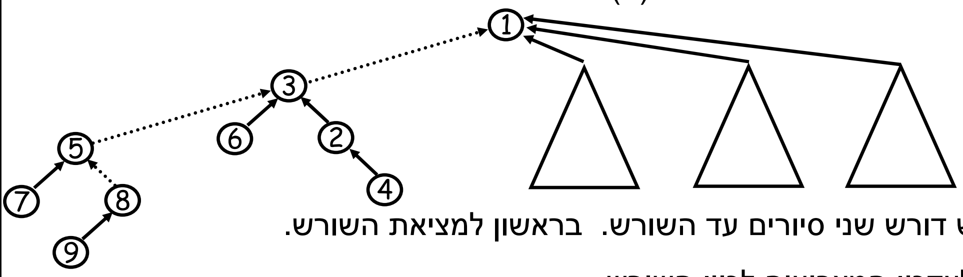
- בפעולת האיחוד האחרונה חובר עץ T_1 לתוך עץ T_2 ולפיכך התקיים $|T_1| \leq |T_2|$.
- נסמן ב- h_1 וב- h_2 את גובה העצים T_1 ו- T_2 בהתאמה.
- מקרה א: $h_2 > h_1$. סתירה שכן אז T הוא עץ בעל גובה זהה ל-T_2 ואינו בעל מספר צמתים מינימלי.
- מקרה ב: $h_2 \leq h_1$. במקרה זה מתקיים $h = h_1 + 1$ ולפיכך
- $|T| = |T_1| + |T_2| \geq 2 \cdot |T_1| \geq 2 \cdot 2^{h-1} = 2^h$ כנדרש.

מסקנה 1

©cs,Technion

שיפור נוסף: כיווץ מסלולים

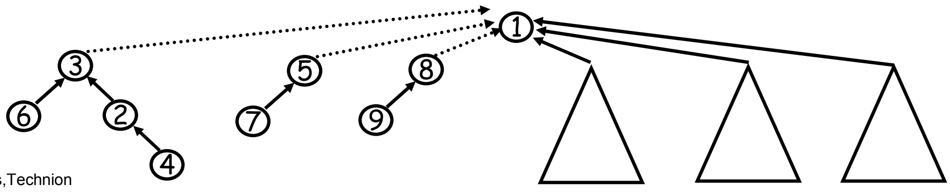
בזמן ביצוע Find(i), עדכן את שדה ה-parent של כל הצמתים מ-i ועד השורש כך שיצביעו ישירות לשורש. דוגמא: Find(8).



המימוש דורש שני סיורים עד השורש. בראשון למציאת השורש.

בשני לעדכון המצביעים לכוון השורש.

<u>תהליך העדכון</u>:



©cs,Technion

ניתוח זמנים

<u>משפט</u>: במימוש באמצעות אוסף עצים של n איברים, סיבוכיות זמן בצוע m פעולות union/find/makeset חסומה ע"י $O(m\ log^*n)$.

לפיכך הזמן המשוערך (amortized time) לכל פעולה חסום ע"י $O(log^*n)$,

כאשר הפונקציה log*n מוגדרת כדלהלן:

$$\log^{(i)}(n) = \underbrace{\log_2 \log_2 \ldots \log_2}_{i} n$$

$$\log^*(n) = \min\{i \geq 0 \mid \log^{(i)}(n) \leq 1\}$$

ובמילים, log*n הוא מספר הפעמים שצריך לקחת log כדי לקבל מספר קטן או שווה לאחד. זוהי פונקציה מונוטונית העולה בקצב מאד איטי. לכל מספר n פרקטי $log^*n \leq 5$. כלומר פעולת find לוקחת זמן משוערך שהוא קבוע מבחינה מעשית.

$$\log^*(1) = 0 \quad \log^*(2) = 1 \quad \log^*(2^2) = 2 \quad \log^*(2^{2^2}) = 3 \quad \log^*(2^{2^{2^2}}) = 4$$

$$\log^*(2^{2^{2^{2^2}}}) = \log^*(2^{65,536}) = 5$$

הוכחת המשפט ותאור חסם הדוק עוד יותר באמצעות פונקציות אקרמן ניתן למצוא בספר הלימוד בעמודים 450-453.

©cs,Technion

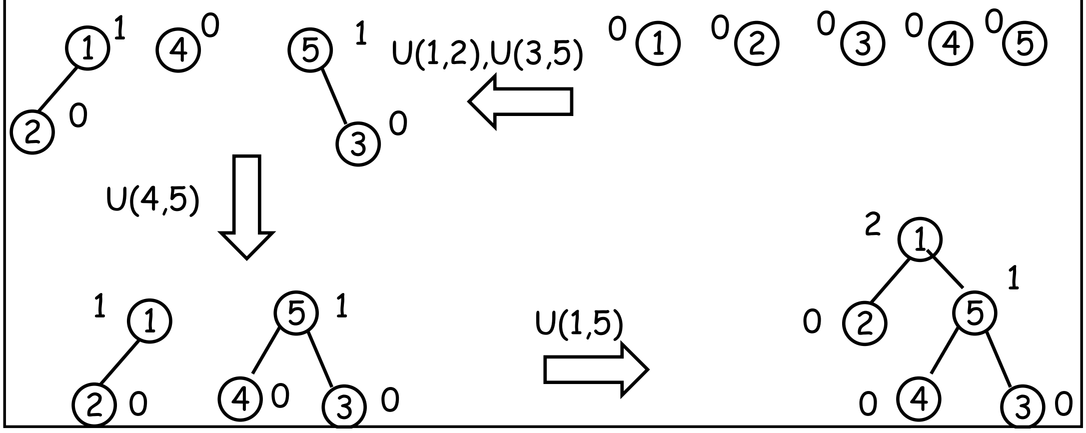
Union-find

וריאנט של האלגוריתם – שימוש ב-rank

לכל איבר נוסיף שדה rank, כך שהשורש של קבוצת יחידה מקבל rank(x) = 0 .

בזמן Union(x,y): אם rank(x)‹ rank(y) הוסף את x כבן נוסף של y .

אם יש שוויון, הוסף שרירותית את x כבן נוסף של y , אך הגדל את rank(y) באחד .



Union-find

וריאנט של האלגוריתם – שימוש ב-rank

לכל איבר נוסיף שדה rank, כך שהשורש של קבוצת יחידה מקבל rank(x) = 0 .

בזמן Union(x,y): אם rank(x)‹ rank(y) הוסף את x כבן נוסף של y .

אם יש שוויון, הוסף שרירותית את x כבן נוסף של y , אך הגדל את rank(y) באחד

טענה 1: rank(x) ‹ rank(parent(x))

<u>הוכחה</u>: מתקיים כאשר x הופך לבן של parent(x).

אחרי זמן זה rank(x) לא גדל, אך rank(parent(x)) לא קטן (יכול אפילו לגדול).

מתי גדל?

- בזמן Unionכאשר לאבא מצטרף תת-עץ בעל אותו rank
- בזמן Find כאשר הבן מתנתק מהאבא והופך להיות בן של השורש

טענה 2:

בעץ שלשורשו יש $rank(r) = k$ יש לכל הפחות 2^k צמתים.

<u>הוכחה</u>: אינדוקציה על $rank(r)$.

מסקנה $rank(x) \le \log n$

טענה 3 יש לכל היותר $n/2^r$ צמתים x בעלי $rank(x) = r$

<u>הוכחה</u>: בזמן ש- x קיבל $rank(x) = r$ היו בעץ שלו לפחות 2^r צמתים.

נסמן ב- x את כל הצמתים בעץ הנ"ל.

במהלך האלגוריתם כל צומת מסומן פעם אחת לכל ערך של r.

כיווץ מסלולים לא משנה את rank.

©cs,Technion

חישוב זמן ה-Find:

נחלק את הצמתים לבלוקים, כך שבלוק i יכיל את כל צמתים v כך ש- $\log^* rank(v) = i$.

תהי g הפונקציה ההפוכה ל-$\log^*$ $g(n) = 2^{2^{\cdot^{\cdot^{2}}}}\}n$

$$Block(j) = \{v : \log^* rank(v) = j\} = \{v : g(j-1) < rank(v) \le g(j)\}$$

חלק מהמחיר של ה-Find נזקוף לבלוק (יסומן B),

וחלק למסלול (יסומן P). נחסום כל חלק בנפרד.

©cs,Technion

פעולות B:

נסתכל על מסלול:

את המעבר על הצומת האחרון של כל בלוק נזקוף לחובת הבלוק.

וכן נזקוף לחובת הבלוקים את המעבר על בן של השורש.

כיון ש- $rank(v) \le \log n$,

בכל מסלול מספר פעולות B חסום ע"י $1 + \log^* \log n \le 2\log^* n$ (צמתים + השורש ובנו).

טענה מהרגע שאביו של צומת ואביו שייכים לבלוקים שונים,
לצומת לא יהיה אב ששייך לאותו בלוק כמוהו.

<u>הוכחה</u> כיון שהבן איננו שורש ה-rank שלו לא יגדל, ולכן הבן יישאר באותו בלוק.
אם האב הוא שורש, ה-rank שלו עשוי לגדול, ואז הוא עלול לעבור לבלוק
עם rank עוד יותר גדול, ועוד יותר רחוק מהבלוק של הבן.

©cs,Technion

פעולות P:

כמה צמתים יש בבלוק ?

$$N(j) = \sum_{g(j-1) < i \le g(j)} \#\text{ nodes with rank i}$$

$$\le \sum_{g(j-1) < i \le g(j)} n/2^i = \frac{n}{2^{g(j-1)+1}} \sum_{0 \le k \le g(j)-g(j-1)-1} 1/2^i$$

$$< \frac{n}{2^{g(j-1)+1}} \sum_{0 \le k} 1/2^i$$

$$= \frac{n}{2^{g(j-1)+1}} 2 = \frac{n}{2^{g(j-1)}}$$

$$= \frac{n}{g(j)}$$

©cs,Technion

פעולות P (המשך)

כיון ש-rank(x) < rank(parent(x)) (טענה 1),

בפעולת P על צומת x, parent(x) הופך להיות השורש.

כיון שבפעולת P x אינו שורש או בנו, rank(parent(x)) גדל ממש.

יהי x,parent(x)∈Block(j).

כיון ש-rank(parent(x)) לא יכול לגדול יותר g(j)-g(j-1) פעמים מבלי ש-parent(x) יעבור לבלוק אחר,
אף צומת בבלוק j לא יכול לקבל יותר מאשר g(j)-g(j-1) זקיפות P.

מרגע ש-x ו-parent(x) לא נמצאים באותו הבלוק, לא ייזקפו ל- x פעולות P.

מכאן:

$$\text{Total cost of P} = \sum_{0\le j<\log^* n} N(j)\big(g(j)-g(j-1)\big)$$

$$\le \sum_{0\le j<\log^* n} \frac{n}{g(j)}\big(g(j)-g(j-1)\big) < \sum_{0\le j<\log^* n} n = n\log^* n$$

מש"ל

©cs,Technion

פונקצית אקרמן

נגדיר את פונקצית אקרמן

$$A(1,j) = 2^j$$
$$A(i,1) = A(i-1,2)$$
$$A(i,j) = A\big(i-1, A(i,j-1)\big)$$

	j=1	j=2	j=3
i=1	2^1	2^2	2^3
i=2	2^2	2^{2^2}	$2^{2^{2^2}}$
i=3	2^{2^2}	$2^{\left.2^{\cdot^{\cdot^{2}}}\right\}16}$	$2^{\left.2^{\cdot^{\cdot^{2}}}\right\}\left.2^{\cdot^{\cdot^{2}}}\right\}16}$

©cs,Technion

הפונקציה ההפוכה

$$\alpha(m,n) = \min\left\{ i \geq 1 : A\left(i, \lfloor m/n \rfloor\right) > \log n \right\}$$

אם פונקציה גדלה מהר אז ההופכית שלה גדלה לאט.

$\alpha(m,n) \leq 4$ לכל ערך סביר של m,n.

משפט Tarjan(בלי הוכחה)

הזמן המשוערך של m פעולות Union/Find על קבוצות זרות מעל n איברים הוא

$$\theta\left(m\alpha(m,n)\right)$$

©cs,Technion