Disjoint sets data type

**Goal.** Support two operations on a set of elements:
- **MAKE-SET(x).** Create a new set containing only element x.
- **FIND(x).** Return a canonical element in the set containing x.
- **UNION(x, y).** Merge the sets containing x and y.

**Dynamic connectivity.** Given an initial empty graph G on n nodes, support the following queries:
- **ADD-EDGE(u, v).** Add an edge between nodes u and v. ← 1 union operation
- **IS-CONNECTED(u, v).** Is there a path between u and v? ← 2 find operations

Disjoint sets data type: applications

**Original motivation.** Compiling **EQUIVALENCE**, **DIMENSION**, and **COMMON** statements in Fortran.

---

**An Improved Equivalence Algorithm**

*Bruno A. Gaal and Michael J. Fischer*

*University of Michigan, Ann Arbor, Michigan*

An algorithm for assigning storage on the basis of **EQUIVALENCE**, **DIMENSION** and **COMMON** declarations is presented. The algorithm is based on a tree structure, and has reduced computation time by 40 percent over a previously published algorithm by identifying all equivalence classes with one scan of the **EQUIVALENCE** declarations. The method is applicable in any problem in which it is necessary to identify equivalence classes, given the element pairs defining the equivalence relation.

---

**Note.** This 1964 paper also introduced key data structure for problem.
**Union-Find**

- **link-by-size**
- **link-by-rank**
- path compression
- **link-by-rank with path compression**
- context

---

**Disjoint-sets data structure**

**Representation.** Represent each set as a tree of elements.
- Each element has a parent pointer in the tree.
- The root serves as the canonical element.
- **FIND(x).** Find the root of the tree containing $x$.
- **UNION(x, y).** Make the root of one tree point to root of other tree.

![Tree Diagram](image)

**Note.** For brevity, we suppress arrows and self loops in figures.

---

**Link-by-size**

**Link-by-size.** Maintain a subtree count for each node, initially 1.
Link root of smaller tree to root of larger tree (breaking ties arbitrarily).

![Tree Diagram](image)

---

**Link-by-size**

**Link-by-size.** Maintain a subtree count for each node, initially 1.
Link root of smaller tree to root of larger tree (breaking ties arbitrarily).

![Tree Diagram](image)
Link-by-size

**Link-by-size.** Maintain a subtree count for each node, initially 1. Link root of smaller tree to root of larger tree (breaking ties arbitrarily).

### MAKE-SET (x)
- `parent(x) ← x.`
- `size(x) ← 1.`

### FIND (x)
- **WHILE** `(x ≠ parent(x))`
  - `x ← parent(x).`
- **RETURN** `x.`

### UNION-BY-SIZE (x, y)
- `r ← FIND (x).`
- `s ← FIND (y).`
- **IF** `(r = s)` **RETURN**.
- **ELSE IF** `(size(r) > size(s))`
  - `parent(s) ← r.`
  - `size(r) ← size(r) + size(s).`
- **ELSE**
  - `parent(r) ← s.`
  - `size(s) ← size(r) + size(s).`

---

### Link-by-size: analysis

**Property.** Using link-by-size, for every root node `r`, `size(r) ≥ 2^{height(r)}.`  
**Pf.** [by induction on number of links]
- **Base case:** singleton tree has size 1 and height 0.
- **Inductive hypothesis:** assume true after first `i` links.
- **Tree rooted at `r`** changes only when a smaller tree rooted at `s` is linked into `r`.
- **Case 1.** `[height(r) > height(s)]`
  - `size'(r) ≥ size(r) ≥ 2^{height(r)} `
  - **→** `size'(r) ≥ 2^{height(r)}` **→** inductive hypothesis

---

**Theorem.** Using link-by-size, any UNION or FIND operations takes `O(log n)` time in the worst case, where `n` is the number of elements.  
**Pf.**
- The running time of each operation is bounded by the tree height.
- By the previous property, the height is `≤ [lg n].`
  - `lg n = log_2 n`
A matching lower bound

**Theorem.** Using link-by-size, a tree with $n$ nodes can have height $= \lg n$.

**Pf.**
- Arrange $2^k - 1$ calls to UNION to form a binomial tree of order $k$.
- An order-$k$ binomial tree has $2^k$ nodes and height $k$. 

---

**Union-Find**

- link-by-size
- link-by-rank
- path compression
- link-by-rank with path compression
- context

---

**Link-by-rank**

**Link-by-rank.** Maintain an integer rank for each node, initially 0. Link root of smaller rank to root of larger rank; if tie, increase rank of new root by 1.

**union(7, 3)**

<table>
<thead>
<tr>
<th>Rank = 1</th>
<th>Rank = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>3 8 9</td>
<td>0 2 5</td>
</tr>
<tr>
<td>1 7</td>
<td>0 2 5</td>
</tr>
</tbody>
</table>

**Note.** For now, rank = height.
Link-by-rank

Link-by-rank. Maintain an integer rank for each node, initially 0. Link root of smaller rank to root of larger rank; if tie, increase rank of new root by 1.

**MAKE-SET** (x)

\[
\begin{align*}
\text{parent}(x) & \leftarrow x. \\
\text{rank}(x) & \leftarrow 0.
\end{align*}
\]

**UNION-BY-RANK** (x, y)

\[
\begin{align*}
r & \leftarrow \text{FIND}(x). \\
s & \leftarrow \text{FIND}(y). \\
\text{If } (r = s) & \text{ RETURN.} \\
\text{ELSE IF } & \text{ rank}(r) > \text{ rank}(s) \\
\text{parent}(r) & \leftarrow r. \\
\text{ELSE IF } & \text{ rank}(r) < \text{ rank}(s) \\
\text{parent}(s) & \leftarrow s. \\
\text{ELSE } \\
\text{parent}(r) & \leftarrow s. \\
\text{rank}(s) & \leftarrow \text{rank}(s) + 1.
\end{align*}
\]

**FIND** (x)

\[
\begin{align*}
\text{WHILE } x \neq \text{parent}(x) & \\
\quad x & \leftarrow \text{parent}(x). \\
\text{RETURN } x.
\end{align*}
\]

**Link-by-rank: properties**

**Property 1.** If x is not a root node, then \( \text{rank}(x) < \text{rank}(\text{parent}(x)) \).

*Pf.* A node of rank \( k \) is created only by merging two roots of rank \( k - 1 \).

**Property 2.** If \( x \) is not a root, then \( \text{rank}(x) \) will never change again.

*Pf.* Rank changes only for roots; a nonroot never becomes a root.

**Property 3.** If \( \text{parent}(x) \) changes, then \( \text{rank}(\text{parent}(x)) \) strictly increases.

*Pf.* The parent can change only for a root, so before linking \( \text{parent}(x) = x \); after \( x \) is linked-by-rank to new root \( r \) we have \( \text{rank}(r) > \text{rank}(x) \).

**Property 4.** Any root node of rank \( k \) has \( \geq 2^k \) nodes in its tree.

*Pf.* [by induction on \( k \)]

- Base case: true for \( k = 0 \).
- Inductive hypothesis: assume true for \( k - 1 \).
- A node of rank \( k \) is created only by merging two roots of rank \( k - 1 \).
- By inductive hypothesis, each subtree has \( \geq 2^{k-1} \) nodes.
- \( \Rightarrow \) resulting tree has \( \geq 2^k \) nodes.

**Property 5.** The highest rank of a node is \( \leq \lfloor \log n \rfloor \).

*Pf.* Immediate from **Property 1** and **Property 4**.

**Property 6.** For any integer \( r \geq 0 \), there are \( \leq n / 2^r \) nodes with rank \( r \).

*Pf.*

- Any root node of rank \( k \) has \( \geq 2^k \) descendants. [**PROPERTY 4**]
- Any nonroot node of rank \( k \) has \( \geq 2^k \) descendants because:
  - it had this property just before it became a nonroot [**PROPERTY 4**]
  - its rank doesn’t change once it becomes a nonroot [**PROPERTY 2**]
  - its set of descendants doesn’t change once it became a nonroot
- Different nodes of rank \( k \) can’t have common descendants. [**PROPERTY 1**]

**Link-by-rank: properties**

**Property 1.** If \( x \) is not a root node, then \( \text{rank}(x) < \text{rank}(\text{parent}(x)) \).

*Pf.* A node of rank \( k \) is created only by merging two roots of rank \( k - 1 \).

**Property 2.** If \( x \) is not a root, then \( \text{rank}(x) \) will never change again.

*Pf.* Rank changes only for roots; a nonroot never becomes a root.

**Property 3.** If \( \text{parent}(x) \) changes, then \( \text{rank}(\text{parent}(x)) \) strictly increases.

*Pf.* The parent can change only for a root, so before linking \( \text{parent}(x) = x \); after \( x \) is linked-by-rank to new root \( r \) we have \( \text{rank}(r) > \text{rank}(x) \).

**Property 4.** Any root node of rank \( k \) has \( \geq 2^k \) nodes in its tree.

*Pf.* [by induction on \( k \)]

- Base case: true for \( k = 0 \).
- Inductive hypothesis: assume true for \( k - 1 \).
- A node of rank \( k \) is created only by merging two roots of rank \( k - 1 \).
- By inductive hypothesis, each subtree has \( \geq 2^{k-1} \) nodes.
- \( \Rightarrow \) resulting tree has \( \geq 2^k \) nodes.

**Property 5.** The highest rank of a node is \( \leq \lfloor \log n \rfloor \).

*Pf.* Immediate from **Property 1** and **Property 4**.

**Property 6.** For any integer \( r \geq 0 \), there are \( \leq n / 2^r \) nodes with rank \( r \).

*Pf.*

- Any root node of rank \( k \) has \( \geq 2^k \) descendants. [**PROPERTY 4**]
- Any nonroot node of rank \( k \) has \( \geq 2^k \) descendants because:
  - it had this property just before it became a nonroot [**PROPERTY 4**]
  - its rank doesn’t change once it becomes a nonroot [**PROPERTY 2**]
  - its set of descendants doesn’t change once it became a nonroot
- Different nodes of rank \( k \) can’t have common descendants. [**PROPERTY 1**]
Link-by-rank: analysis

**Theorem.** Using link-by-rank, any UNION or FIND operations takes $O(\log n)$ time in the worst case, where $n$ is the number of elements.

**Pf.**
- The running time of each operation is bounded by the tree height.
- By the PROPERTY 5, the height is $\leq \lceil \log n \rceil$.

### UNION-FIND

- link-by-size
- link-by-rank
- path compression
- link-by-rank with path compression
- context

#### SECTION 5.1.4

**Path compression.** After finding the root $r$ of the tree containing $x$, change the parent pointer of all nodes along the path to point directly to $r$. 

![Diagram of path compression before and after](image-url)
Path compression. After finding the root $r$ of the tree containing $x$, change the parent pointer of all nodes along the path to point directly to $r$. 
Path compression. After finding the root $r$ of the tree containing $x$, change the parent pointer of all nodes along the path to point directly to $r$.

```python
FIND (x)
If $x \neq \text{parent}(x)$
    parent($x$) ← FIND (parent($x$)).
RETURN parent($x$).
```

**Note.** Path compression does not change the rank of a node; so $\text{height}(x) \leq \text{rank}(x)$ but they are not necessarily equal.

### Theorem. [Tarjan-van Leeuwen 1984] Starting from an empty data structure, path compression (with naive linking) performs any intermixed sequence of $m \geq n$ find and $n – 1$ union operations in $O(m \log n)$ time.

**Pf.** Nontrivial but omitted.

---

**Link-by-rank with path compression: properties**

**Property.** The tree roots, node ranks, and elements within a tree are the same with or without path compression.

**Pf.** Path compression does not create new roots, change ranks, or move elements from one tree to another. •
**Link-by-rank with path compression: properties**

**Property.** The tree roots, node ranks, and elements within a tree are the same with or without path compression.

**Corollary.** PROPERTY 2, 4–6 hold for link-by-rank with path compression.

**Property 1.** If \( x \) is not a root node, then \( \text{rank}(x) < \text{rank}(\text{parent}(x)) \).

**Property 2.** If \( x \) is not a root, then \( \text{rank}(x) \) will never change again.

**Property 3.** If \( \text{parent}(x) \) changes, then \( \text{rank}(\text{parent}(x)) \) strictly increases.

**Property 4.** Any root node of rank \( k \) has \( \geq 2^k \) nodes in its tree.

**Property 5.** The highest rank of a node is \( \leq \lfloor \log n \rfloor \).

**Property 6.** For any integer \( r \geq 0 \), there are \( \leq n / 2^r \) nodes with rank \( r \).

**Bottom line.** PROPERTY 1–6 hold for link-by-rank with path compression. (but we need to recheck PROPERTY 1 and PROPERTY 3)

---

**Iterated logarithm function**

**Def.** The iterated logarithm function is defined by:

\[
\lg^* n = \begin{cases} 
1 & \text{if } n \leq 1 \\
1 + \lg^*(\lg n) & \text{otherwise}
\end{cases}
\]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \lg^* n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>(3, 4)</td>
<td>2</td>
</tr>
<tr>
<td>[5, 16]</td>
<td>3</td>
</tr>
<tr>
<td>[17, 65536]</td>
<td>4</td>
</tr>
<tr>
<td>[65537, 2^{65536}]</td>
<td>5</td>
</tr>
</tbody>
</table>

**Note.** We have \( \lg^* n \leq 5 \) unless \( n \) exceeds the \# atoms in the universe.

---

**Analysis**

Divide nonzero ranks into the following groups:

- \( \{ 1 \} \)
- \( \{ 2 \} \)
- \( \{ 3, 4 \} \)
- \( \{ 5, 6, ..., 16 \} \)
- \( \{ 17, 18, ..., 2^{16} \} \)
- \( \{ 65537, 65538, ..., 2^{65536} \} \)
- \( \ldots \)

**Property 7.** Every nonzero rank falls within one of the first \( \lg^* n \) groups.

**Pf.** The rank is between 0 and \( \lfloor \log n \rfloor \). [PROPERTY 5]
Creative accounting

Credits. A node receives credits as soon as it ceases to be a root. If its rank is in the interval \( \{ k+1, k+2, \ldots, 2^k \} \), we give it \( 2^k \) credits.

Proposition. Number of credits disbursed to all nodes is \( \leq n \lg^* n \).

\[ \text{Pf.} \]

- By \textit{PROPERTY 6}, the number of nodes with rank \( \geq k+1 \) is at most \( \frac{n}{2^k+1} + \frac{n}{2^k+2} + \cdots \leq \frac{n}{2^k} \)
- Thus, nodes in group \( k \) need at most \( n \) credits in total.
- There are \( \leq \lg^* n \) groups. \[
\text{[PROPERTY 7]} \] •

Running time of find

Running time of find. Bounded by number of parent pointers followed.

- Recall: the rank strictly increases as you go up a tree. \[
\text{[PROPERTY 1]} \]
- Case 0: \( \text{parent}(x) \) is a root \( \Rightarrow \) only happens for one link per \text{FIND}.
- Case 1: \( \text{rank(parent)}(x) \) is in a higher group than \( \text{rank}(x) \).
- Case 2: \( \text{rank(parent)}(x) \) is in the same group as \( \text{rank}(x) \).

\[ \text{Case 1. At most } \lg^* n \text{ nodes on path can be in a higher group.} \] \[
\text{[PROPERTY 7]} \]

\[ \text{Case 2. These nodes are charged } 1 \text{ credit to follow parent pointer.} \]

- Each time \( x \) pays 1 credit, \( \text{rank(parent)}(x) \) strictly increases. \[
\text{[PROPERTY 1]} \]
- Therefore, if \( \text{rank}(x) \) is in the group \( \{ k+1, \ldots, 2^k \} \), the rank of its parent will be in a higher group before \( x \) pays \( 2^k \) credits.
- Once \( \text{rank(parent)}(x) \) is in a higher group than \( \text{rank}(x) \), it remains so because:
  - \( \text{rank}(x) \) does not change once it ceases to be a root. \[
\text{[PROPERTY 2]} \]
  - \( \text{rank(parent)}(x) \) does not decrease. \[
\text{[PROPERTY 3]} \]
  - thus, \( x \) has enough credits to pay until it becomes a Case 1 node. •

Link-by-rank with path compression

\textbf{Theorem.} Starting from an empty data structure, link-by-size with path compression performs any intermixed sequence of \( m \geq n \) \text{FIND} and \( n-1 \) \text{UNION} operations in \( O(m \log^* n) \) time.

\textbf{UNION-FIND}

- link-by-size
- link-by-rank
- path compression
- link-by-rank with path compression
- context
Theorem. [Fischer 1972] Link-by-size with path compression performs any intermixed sequence of $m \geq n$ FIND and $n - 1$ UNION operations in $O(m \log \log n)$ time.

Theorem. [Hopcroft-Ullman 1973] Link-by-size with path compression performs any intermixed sequence of $m \geq n$ FIND and $n - 1$ UNION operations in $O(m \log^2 n)$ time.

Theorem. [Tarjan 1975] Link-by-size with path compression performs any intermixed sequence of $m \geq n$ FIND and $n - 1$ UNION operations in $O(m \alpha(m, n))$ time, where $\alpha(m, n)$ is a functional inverse of the Ackermann function.

Ackermann function. A computable function that is not primitive recursive.

$$A(m, n) = \begin{cases} 
  n + 1 & \text{if } m = 0 \\
  A(m, n - 1) & \text{if } m > 0 \text{ and } n = 0 \\
  A(m - 1, A(m, n - 1)) & \text{if } m > 0 \text{ and } n > 0 
\end{cases}$$

Zum Hilbertschen Aufbau der reellen Zahlen.

von

Wilhelm Ackermann in Göttingen.

Um den Beweis für die vorerst unbestimmte Voraussetzung zu erbringen, daß sich die Menge der reellen Zahlen, d. h. der unendlichen Dezimalbrachen zusammensetzen läßt, braucht Hilbert einen speziellen Ausbau der sogenannten infiniten Zahlen. Wesentlich bei diesem Beweis ist der Begriff des Typs einer Funktion. Eine Funktion vom Typ 1 ist eine solche, deren Argumente und Werte ganze Zahlen sind, also eine gewöhnliche Funktion in der Mathematik. Eine Funktion vom Typ 2 ist die Funktion der Funktionen. Ein solcher Begriff erfordert die Begriffe von Typ 3 usw. Die Definitionen der Typen hängen sich auch von Bedeutung ab, die für den Einzelfall dieser Arbeit ist das aber nicht von Belang.

Note. There are many inequivalent definitions.
Ackermann function

A computable function that is not primitive recursive.

\[ A(m, n) = \begin{cases} 
  n + 1 & \text{if } m = 0 \\
  A(m-1, 1) & \text{if } m > 0 \text{ and } n = 0 \\
  A(m-1, A(m, n-1)) & \text{if } m > 0 \text{ and } n > 0 
\end{cases} \]

Inverse Ackermann function.

\[ \alpha(m, n) = \min\{i \geq 1 : A(i, \lfloor m/n \rfloor) \geq \log_2 n\} \]

“I am not smart enough to understand this easily.”

— Raymond Seidel

Inverse Ackermann function

Definition.

\[ \alpha_k(n) = \begin{cases} 
  1 & \text{if } n = 1 \\
  \lfloor n/2 \rfloor & \text{if } k = 1 \\
  1 + \alpha_k(\alpha_{k-1}(n)) & \text{otherwise} 
\end{cases} \]

Ex.

• \( \alpha_0(n) = \lfloor n/2 \rfloor \).
• \( \alpha_2(n) = \lceil \log n \rceil = \# \text{ of times we divide } n \text{ by two, until we reach 1} \).
• \( \alpha_3(n) = \lceil \log^* n \rceil = \# \text{ of times we apply the } \log \text{ function to } n, \text{ until we reach 1} \).
• \( \alpha_d(n) = \# \text{ of times we apply the iterated } \log \text{ function to } n, \text{ until we reach 1} \).

A matching lower bound

Theorem. [Fredman-Saks 1989] Any CPROBE(log n) algorithm for the disjoint set union problem requires \( \Omega(m \alpha(m, n)) \) time to perform an intermixed sequence of \( m \geq n \) FIND and \( n-1 \) UNION operations in the worst case.

Cell probe model. [Yao 1981] Count only number of words of memory accessed; all other operations are free.
Path compaction variants

**Path splitting.** Make every node on path point to its grandparent.

**Path halving.** Make every other node on path point to its grandparent.

Linking variants

**Link-by-size.** Number of nodes in tree.

**Link-by-rank.** Rank of tree.

**Link-by-random.** Label each element with a random real number between 0.0 and 1.0. Link root with smaller label into root with larger label.

Disjoint set union algorithms

**Theorem.** [Tarjan-van Leeuwen 1984] Link-by \{ size, rank \} combined with \{ path compression, path splitting, path halving \} performs any intermixed sequence of \( m \geq n \) find and \( n - 1 \) union operations in \( O(m \alpha(m, n)) \) time.