Pigeonhole Principle

Robert Y. Lewis

CS 0220 2024

April 1, 2024
Overview

1. Refresher: where are we?

2. The Pigeonhole Principle (14.8)
   - Hairs on Heads (14.8.1)
   - Subsets with the Same Sum (14.8.2)

3. Graphs primer for Wed
What’s coming up

- We’ve just started talking about *combinatorics*, the art of counting things.
  - Our move: get good at counting certain kinds of things, like sequences. Then count other things by finding bijections to sets of sequences.
- \( \binom{n}{k} \): the number of ways to pick \( k \) items from a set of \( n \) options.

- Leads into *probability*.
  - In the finite case, can be thought of as “weighted counting.”

- A general framework (probability spaces, random variables), on which we’ll define many operations in ways that generalize.
Puzzle

A drawer in a dark room contains red socks, green socks, and blue socks. You are going to grab $k$ socks, then go someplace light. How many socks must you take with you to be sure that you have at least one matching pair?

Two might be enough. But it’s not guaranteed. After all, one might be one red, one green. Three is not enough. Might grab one of each. At four, one of the colors must repeat—you have a matching pair!
Pigeonhole principle

*If there are more pigeons than holes they occupy, then at least two pigeons must be in the same hole.*


We will use it as a tool for reasoning about the relationships between the sizes of sets.
Pigeonhole principle as a relation

Pigeons

Polly
Pippin
Percival
Priscilla

Pigeonholes

1
2
3
Sock example

Socks

Colors

1

2

3

4

red

blue

green
Pigeonhole principle

**Rule:** If $|A| > |B|$, then, for every function $f : A \to B$, there exist two different elements of $A$ that are mapped by $f$ to the same element of $B$.

Recall: If a function from $A$ to $B$ is injective (no collisions), then $|A| \leq |B|$.

By the contrapositive of this statement, if $|A| > |B|$ (it’s not the case that $|A| \leq |B|$), then a function from $A$ to $B$ has collisions (is not injective).
Australian sheep

There’s a pair of sheep in Australia that have precisely the same number of hairs.

Sheep have up to 60 million wool follicles. Australia has 70 million sheep. Pigeonhole principle says there must be (at least!) two that match up.
All the sheep in China

China has 185 million sheep, each with up to 60 million follicles. Same argument as Australia applies. But, we can go further.

There must be a *foursome* of sheep with precisely the same number of hairs on their body. Argument: Even if we spread things out maximally, there’s some follicle count with 4 sheep in it.
Generalized Pigeonhole Principle

**Rule:** If $|A| > k|B|$, then every function $f : A \rightarrow B$ maps at least $k + 1$ different elements of $A$ to the same element of $B$. 
Lossless compression

An example from CS!

Let \( B \) be the set of bit sequences of length at most \( n \). A *compression algorithm* is a function \( f : B \rightarrow B \) that intends to map “interesting” sequences to shorter sequences. A *lossless* compression algorithm is one that is invertible.

By the pigeonhole principle: every lossless compression algorithm that compresses at least one input must expand some other input.

Why isn’t this a problem?
Hash functions

Another classic CS application: hash maps.

Idea: store elements of $A$ in an array of length $n$. Compute the index of element $a$ by $h(a)$, where $h : A \rightarrow \mathbb{N}$ is a hash function.

Pigeonhole principle tells you when you’re guaranteed to have a collision: there exist two distinct elements with the same hash value exactly when $|A| > n$.

But this isn’t the analysis we really want to do. How likely is a collision? How big should $n$ be, in general?
Subsets with the Same Sum (14.8.2)

25 2-digit numbers

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>34</td>
<td>41</td>
<td>89</td>
<td>99</td>
</tr>
<tr>
<td>0</td>
<td>95</td>
<td>56</td>
<td>90</td>
<td>43</td>
</tr>
<tr>
<td>41</td>
<td>37</td>
<td>94</td>
<td>94</td>
<td>69</td>
</tr>
<tr>
<td>15</td>
<td>96</td>
<td>85</td>
<td>76</td>
<td>30</td>
</tr>
<tr>
<td>27</td>
<td>24</td>
<td>75</td>
<td>63</td>
<td>1</td>
</tr>
</tbody>
</table>

Two pairs with the same sum?

\[76 + 30 = 106 = 43 + 63\]
\[37 + 94 = 131 = 41 + 90\]

Lots of others.
Analysis

List of 25 2-digit numbers. Must two different pairs have the same sum?

A: How many possible pairs? \( \binom{25}{2} = 25 \times 24/2 = 300 \). In general, \( n \) numbers can be paired in \( \binom{n}{2} \) ways.

B: How many possible pairwise sums? Smallest is \( 0 + 0 = 0 \), largest is \( 99 + 99 = 198 \). So, 199 possible values. In general, pairs of \( d \)-digit numbers can sum to \( 2 \times 10^d - 1 \) possible values.

Since \( 300 > 199 \), by the pigeonhole principle, some possible sum must be the result of two different pairs.
40 10-digit numbers

372995585  2234938293  7149708291  7060913492
2931952606  3111391181  951202341  735405394
9217967649  824907338  5014746657  1237286631
522671716  8407467172  8707242896  683510550
4879527699  5950653401  5161166931  816710053
1132327190  5154516759  2790970967  2560087807
8425382298  4088307493  3641040020  2603650330
3153589691  8144289385  4677056137  4058664392
9202688143  144564604  6259750302  2210904697
6037630341  498877221  186119657  71015872

Two subsets with the same sum?
Analysis

List of 40 10-digit numbers. Must two different subsets have the same sum?

A: How many possible subsets? $2^{40} = 1.1e + 12$. In general, $n$ numbers can produce $2^n$ subsets.

B: How many possible subset sums? Smallest is $0 + 0 = 0$, largest is $40 \times 9999999999 = 4.0e + 11$. In general, up to $n$ $d$-digit numbers can sum to $n \times 10^d$ possible values.

Since $1.1e + 12 > 4.0e + 11$, by the pigeonhole principle, some possible sum must be the result of two different subsets.
Graphs and polyhedra

**Definition:** A graph $G$ consists of a set of vertices $V(G)$ and edges $E(G)$.

This terminology comes from descriptions of flat-sided solid shapes—polyhedra.

\[
\{a, b, c, d, e\} \text{ are the vertices, } \{\{a, b\}, \{a, c\}, \{c, d\}, \{b, d\}, \{a, e\}, \{b, e\}, \{c, e\}, \{d, e\}\} \text{ are the edges.}
\]
Adjacency relation

\[ E(G) \subseteq \{\{u, v\} \mid u, v \in V(G), u \neq v\}. \]

\( u \) adjacent to \( v \) in \( G \) iff \( \{u, v\} \in E(G) \).

Is the “adjacent to” relation reflexive? No. Symmetric? Yes! Transitive? Depends. Empty is graph is! Otherwise, no, because symmetry and non-reflexive.

A simple path between vertices \( v_0 \) and \( v_k \) is a sequence of vertices \( (v_0, v_1, \ldots, v_k) \in V(G)^{k+1} \) such that

- \( v_0, \ldots, v_k \) are all distinct;
- for each \( i < k \), \( v_i \) and \( v_{i+1} \) are adjacent.

A graph is connected if there is a simple path between every pair of vertices.
Trees

A *simple cycle* is a simple path between vertices $v_0$ and $v_k$ with the added constraint that $v_0$ and $v_k$ are adjacent.

A *tree* is a connected graph that has no cycles.