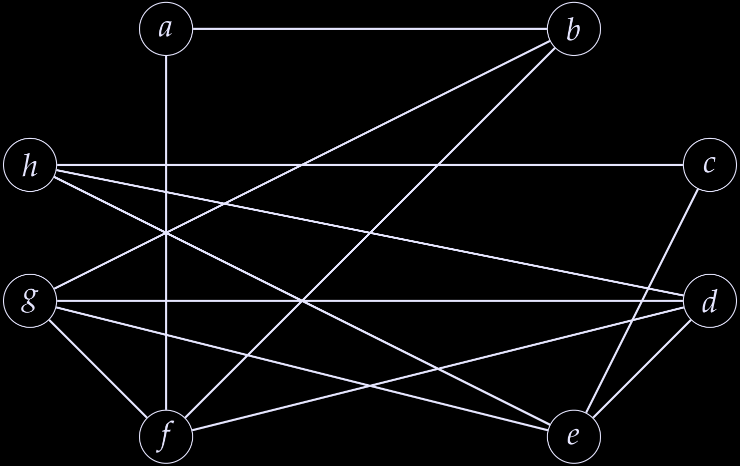
Lossless Graph Compression



Lossless Graph Compression

Main ideas

- Exploit repetitions
- Exploit distribution of values

Huffman coding

Data: Set O of objects, each object o_i has probability p_i

if $|O| > 2$ **then**

  Pick two objects o_i, o_j with smallest probability;

  $x \leftarrow$ new object with probability $p_i + p_j$;

  $h \leftarrow \text{Huffman}(O \setminus \{o_i, o_j\} \cup \{x\})$;

  $h(o_i) \leftarrow h(x)0$;

  $h(o_j) \leftarrow h(x)1$;

  Remove $h(x)$;

else

  $h(o_1) \leftarrow 0$;

  $h(o_2) \leftarrow 1$;

return h;

Elias γ code

binary code for $x \geq 1$

1. $N = \lfloor \log_2 x \rfloor$
2. N zeroes · 1 one · binary representation of x, omitting the leading bit
3. uses $2\lfloor \log_2 x \rfloor + 1$ bits

Elias δ code

binary code for $x \geq 1$

1. $N = \lfloor \log_2 x \rfloor$
2. $\gamma(N+1) \cdot$ binary representation of x, omitting the leading bit
3. uses $\lfloor \log_2 x \rfloor + 2\lfloor \log_2 (\lfloor \log_2 x \rfloor + 1) \rfloor + 1$ bits

Variable-length nibble code

binary code for $x \geq 1$

1. $p \leftarrow$ the binary representation of n, left-padded with zeroes, so that its length is a multiple of 3
2. Split p into 3-bit blocks
3. prepend each block with a zero, replace the leading 0 of the last block with a one.

Minimal binary code

binary code for $0 \le x \le z-1$

1. $s = \lceil \log_2 z \rceil$
2. $p \leftarrow 2^s - z$
3. If $x < p$ then output the x-th binary word of length $s-1$
4. If $x \ge p$ then output the $(x - z + 2^s)$-th binary word of length s

ζ_k code

binary code for $2^{hk} \leq x \leq 2^{(h+1)k} - 1$

1. k: shrinking factor
2. h+1 in unary · minbincode of $x - 2^{hk}$, with $z = 2^{(h+1)k} - 2^{hk} - 1$

Move-to-front transform

1. Maintain the list L of recently used objects
2. Encode an object as its index in L
3. Move the object to the head of L

Run-length encoding

AAAAAABBBB $\longrightarrow$ (A,6)(B,4)

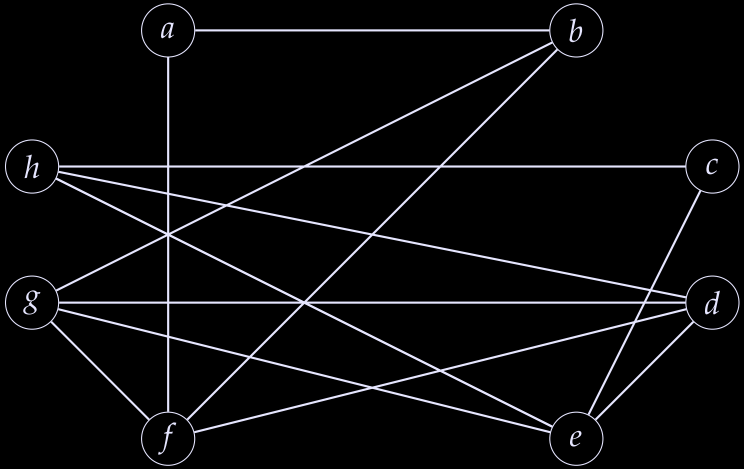
Run-length encoding

AAAAAABBBB $\longrightarrow$ (A,6)(B,4)

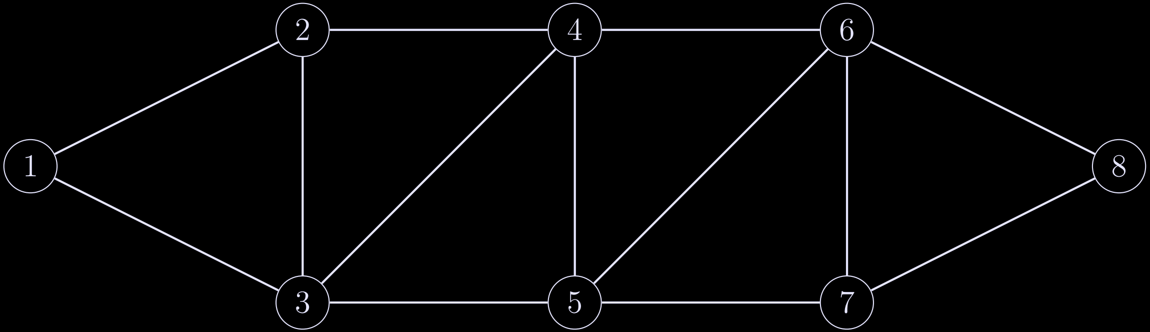
Binary alphabet

- 0-runs and 1-runs alternate
- start with 1-run (0-length if 0-run)
- first, number of runs
- then run-lengths (lengths decremented by 1)

Lossless Graph Compression



Lossless Graph Compression



Gap representation

Delta

1. adjacency list of v
2. store difference with previous vertex
3. store difference with v

Gap representation

Delta

1. adjacency list of v
2. store difference with previous vertex
3. store difference with v

Example adj(3)

-2, 1, 2, 1

Gap representation

Delta

1. adjacency list of v
2. store difference with previous vertex
3. store difference with v

Example adj(3)

-2, 1, 2, 1

Assumptions?

Gap representation

Delta

1. adjacency list of v
2. store difference with previous vertex
3. store difference with v

Example adj(3)

-2, 1, 2, 1

Assumptions?

Neighbors of a vertex are close to the vertex.

Reference compression

$N(x)$ and $N(y)$

1. $N(y)$ is a previous vertex
2. which elements of $N(x)$ are not in $N(y)$?
3. $N(x) \setminus N(y)$

Reference compression

$N(x)$ and $N(y)$

1. $N(y)$ is a previous vertex
2. which elements of $N(x)$ are not in $N(y)$?
3. $N(x) \setminus N(y)$

Example adj(4)

adj(2) = [1, 3, 4]; adj(4) = [2, 3, 5, 6]$\longrightarrow \langle 2, 101_2, [1, 1] \rangle$
using the triple ⟨ previous vertex, characteristic vector of the vertices of $N(y)$ that are not in $N(x)$, the encoding of $N(x) \setminus N(y)$⟩

Reference compression

$N(x)$ and $N(y)$

1. $N(y)$ is a previous vertex
2. which elements of $N(x)$ are not in $N(y)$?
3. $N(x) \setminus N(y)$

Example adj(4)

adj(2) = [1, 3, 4]; adj(4) = [2, 3, 5, 6]$\longrightarrow \langle 2, 101_2, [1, 1] \rangle$
using the triple ⟨ previous vertex, characteristic vector of the vertices of $N(y)$ that are not in $N(x)$, the encoding of $N(x) \setminus N(y)$⟩

Assumptions?

Reference compression

$N(x)$ and $N(y)$

1. $N(y)$ is a previous vertex
2. which elements of $N(x)$ are not in $N(y)$?
3. $N(x) \setminus N(y)$

Example adj(4)

adj(2) = [1, 3, 4]; adj(4) = [2, 3, 5, 6]$\longrightarrow \langle 2, 101_2, [1, 1] \rangle$
using the triple $\langle$ previous vertex, characteristic vector of the vertices of $N(y)$ that are not in $N(x)$, the encoding of $N(x) \setminus N(y)\rangle$

Assumptions?

$N(x)$ and $N(y)$ are almost identical

Interval encoding

Interval encoding

$[b, e]$

- $[b, e - b]$
- if all intervals are longer than a threshold $L \Rightarrow$ decrement all lengths by L

Figures

- David Eppstein, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10261635