

## Lecture #39: Compression

### Announcements

- HKN surveys Friday in class. **Extra points awarded to those who participate!**
- Changes (simplifications?) to Project 3 spec:
  - Last two parts of "git status" are extra-credit. Leave empty (just the headers) if you don't want to do them.
  - Specified what happens to staging area for checkout and reset.
  - Change error messages for checkout to make more sense.

Last modified: Wed Dec 2 12:26:11 2015

CS61B: Lecture #39 1

## Compression and Git

- Git creates a new object in the repository each time a changed file or directory is committed.
- Things can get crowded as a result.
- To save space, it **compresses** each object.
- Every now and then (such as when sending or receiving from another repository), it packs objects together into a single file: a "packfile."
- Besides just sticking the files together, uses a technique called **delta compression**.

Last modified: Wed Dec 2 12:26:11 2015

CS61B: Lecture #39 2

## Delta Compression

- Typically, there will be many versions of a file in a Git repository: the latest, and previous edits of it, each in different commits.
- Git doesn't keep track explicitly of which file came from where, since that's hard in general:
  - What if a file is split into two, or two are spliced together?
- But, can guess that files with same name and (roughly) same size in two commits are probably versions of the same file.
- When that happens, store one of them as a pointer to the other, plus a list of changes.

Last modified: Wed Dec 2 12:26:11 2015

CS61B: Lecture #39 3

## Delta Compression (II)

- So, store two versions

V1

My eyes are fully open to my awful situation.  
I shall go at once to Roderick and make him an oration. I shall tell him I've recovered my forgotten moral senses,

as

V1

[Fetch 1st 6 lines from V2]

V2

My eyes are fully open to my awful situation.  
I shall go at once to Roderick and make him an oration.  
I shall tell him I've recovered my forgotten moral senses, and don't give twopence halfpenny for any consequences.

V2

My eyes are fully open to my awful situation.  
I shall go at once to Roderick and make him an oration.  
I shall tell him I've recovered my forgotten moral senses, and don't give twopence halfpenny for any consequences.

Last modified: Wed Dec 2 12:26:11 2015

CS61B: Lecture #39 4

## Compression Techniques

Slides from Josh Hug

## LZ77 and DEFLATE

- Git Actually uses a different scheme from LZW for compression: a combination of LZ77 and Huffman coding.
- LZ77 is kind of like delta compression, but within the same text.
- Convert a text such as

One Mississippi, two Mississippi

into something like

One Mississippi, two <11,7>

where the <11,7> is intended to mean "the next 11 characters come from the text that ends 7 characters before this point."

- We add new symbols to the alphabet to represent these (length, distance) inclusions.
- When done, Huffman encode the result.