Catching Plagiarists

Baker Franke
The University of Chicago Laboratory Schools

THE UNIVERSITY OF

CHICAGO

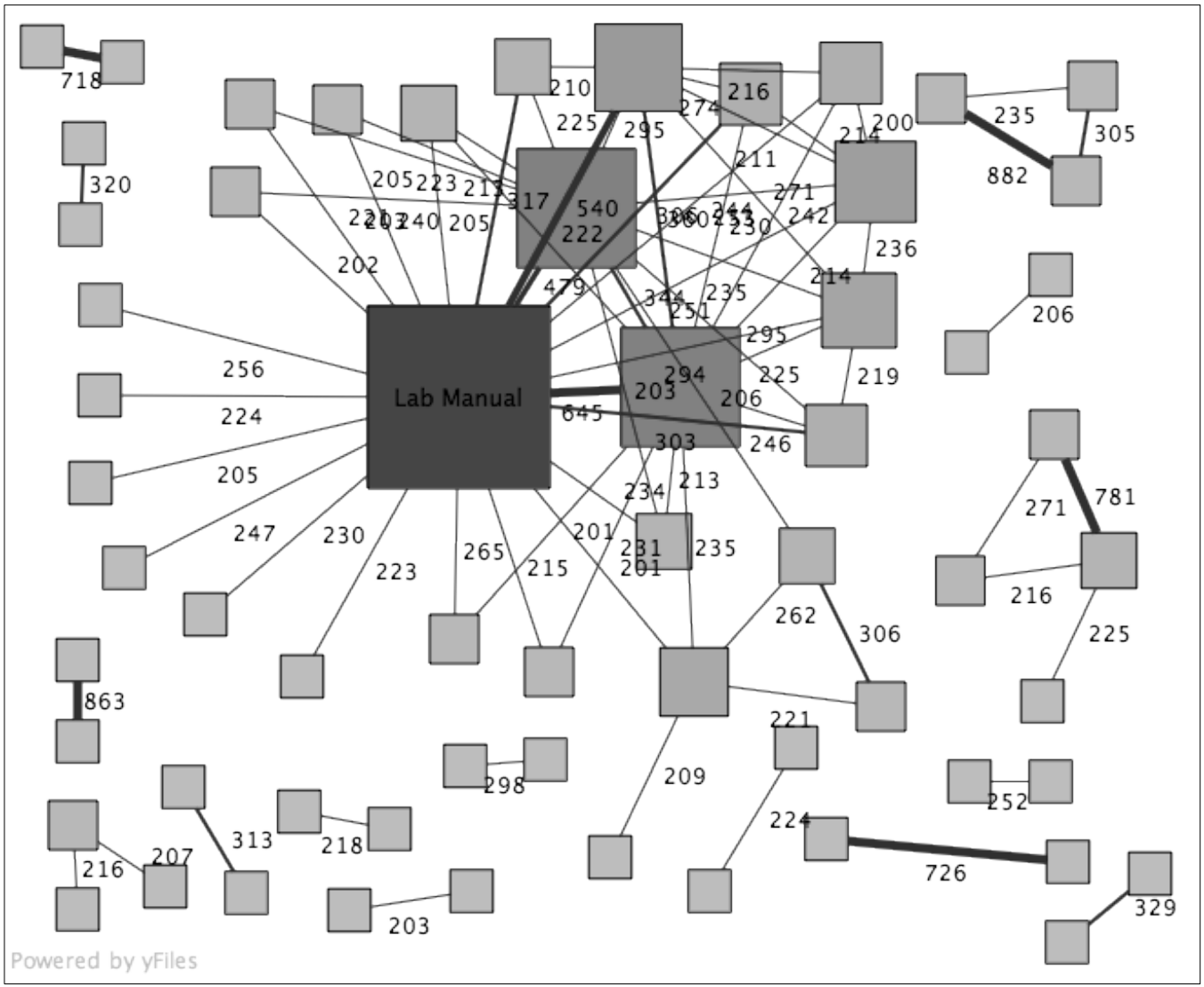
LABORATORY SCHOOLS

Problem Motivation

- e.g. Physics 101 at a large university
- One student helping another by giving him a copy of his assignment only to have the other turn it in (or large portions of it) as his own.

Problem Statement

- Given a set of text documents, find if any have been plagiarized in full, or in part, from some other document in the set.


Lab Manual
718
320
210
225
295
274
216
200
214
211
235
305
882
205
223
215
317
540
222
271
242
230
236
202
479
344
235
251
214
206
256
294
203
206
225
219
224
645
303
246
205
234
213
271
781
247
230
223
265
215
201
231
235
201
262
306
216
225
863
221
298
209
252
313
218
207
216
224
726
329
203
Powered by yFiles

What's Nifty?

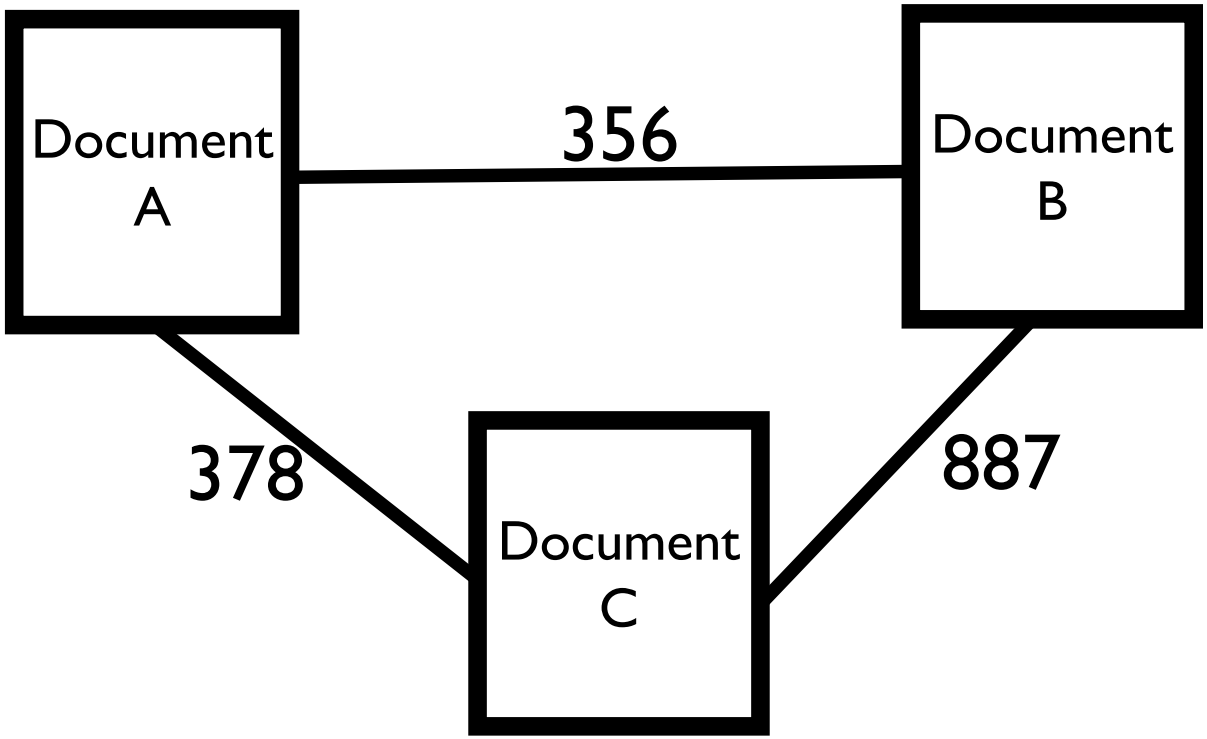
- Virtually zero prep...or a lot.
- Covers *many* areas of CS1/CS2
- Real Big-Oh implications
- Accessible by/challenging to all levels of students
- No one minds it being a console application!
- Relevance and motivation

Nift[yi]est Thing of All

- This problem REQUIRES a computational solution - no other way to do it.

Suggest a Strategy?

- Compare n-word chunks between documents.



Nifty Parts

1. Harder text processing
2. Real data structure choices/consequences
3. “Real World” issues

Part 1 : Text Processing

- Processing many documents in a non-trivial way.
- Option: just compare two documents to find a degree of similarity.
- Interesting Question: *what* do I need to compare?
- Did someone say regular expressions?

Part II : Data Structures

- Comparing even a small set of documents requires some structures
- *Many* possibilities
- Real Big-Oh concerns / analysis

Part III : Real World Issues

- Large document set quickly becomes unmanageable if poor algorithm chosen.
- Memory limits
- Output representation
- Legal issues?

Resources

- Web Site
- Assignment Sheet

“Through the duration of Heathcliff’s life, he encounters many tumultuous events that affects him as a person and transforms his rage deeper into his soul, for which he is unable to escape his nature.”

--bwa93.txt