Code Plagiarism Detection @ CSE

CSE 501 Presentation

Scott Settembre, TA

September 14, 2010

University at Buffalo
The State University of New York

cse@buffalo
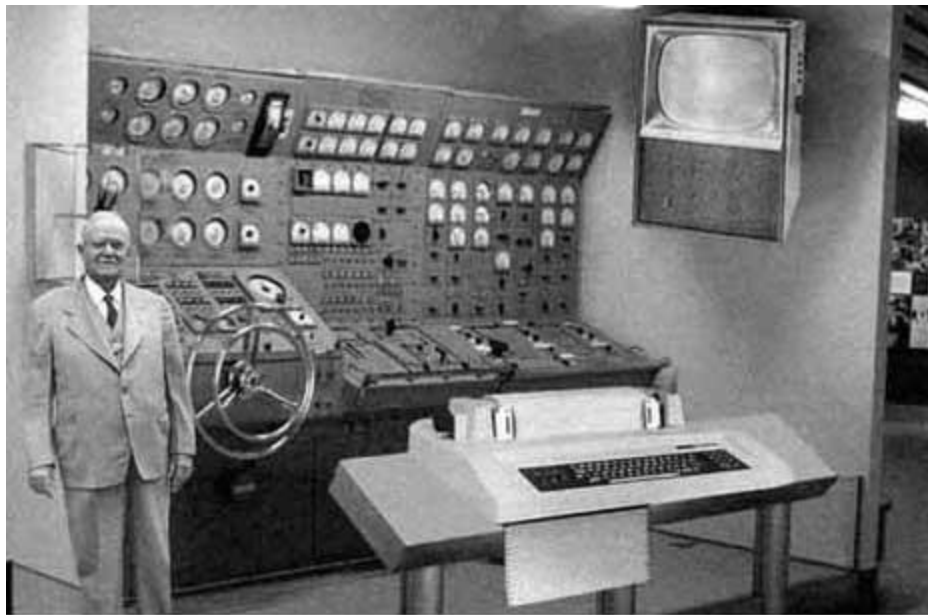
Once Upon a Time…

- Code copying was a problem.

"Nice code Tom, your cut and paste skills are beyond reproach."

Attempted to Solve the Issues…

- Solutions:
  - Honor system – ethical standards
  - Common sense reasoning – "it only hurts you"
  - Threaten to unleash the TA's
  - Severe punishments
- Results
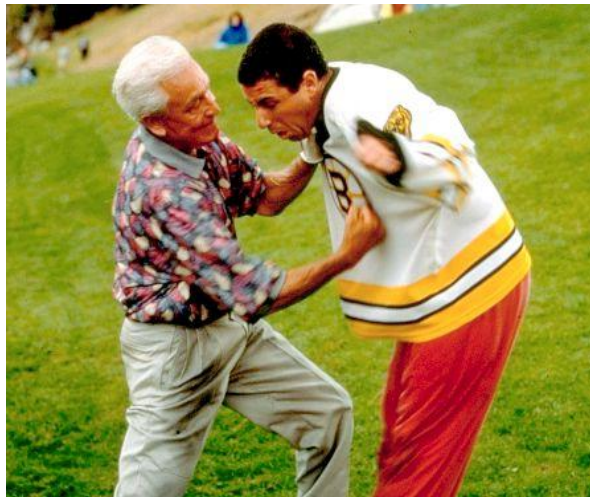 - ??? Unknown ??? – never measured
  - Good enough?

Then Came Along…

- A wonderful invention…

The Computer!!!

Professor Bluffing

- Discovered various PDS or "Plagiarism Detection Software/Services"

- Proved their existence by linking to them in their power-point slides

- Assumed this would deter any cheating...

"Did you copy code? Tell me! Tell me so I can punish you severely!!!"

Students notice…

- Lots of students still cheating…
- Lots of students still getting away with it…

"Psssst! What font are you using?"

"We weren't cheating. We were consulting."

Why is this important?

- Professors can and do *"curve"* grades. This means they bump a lower grade up when many of the grades for a project are low.
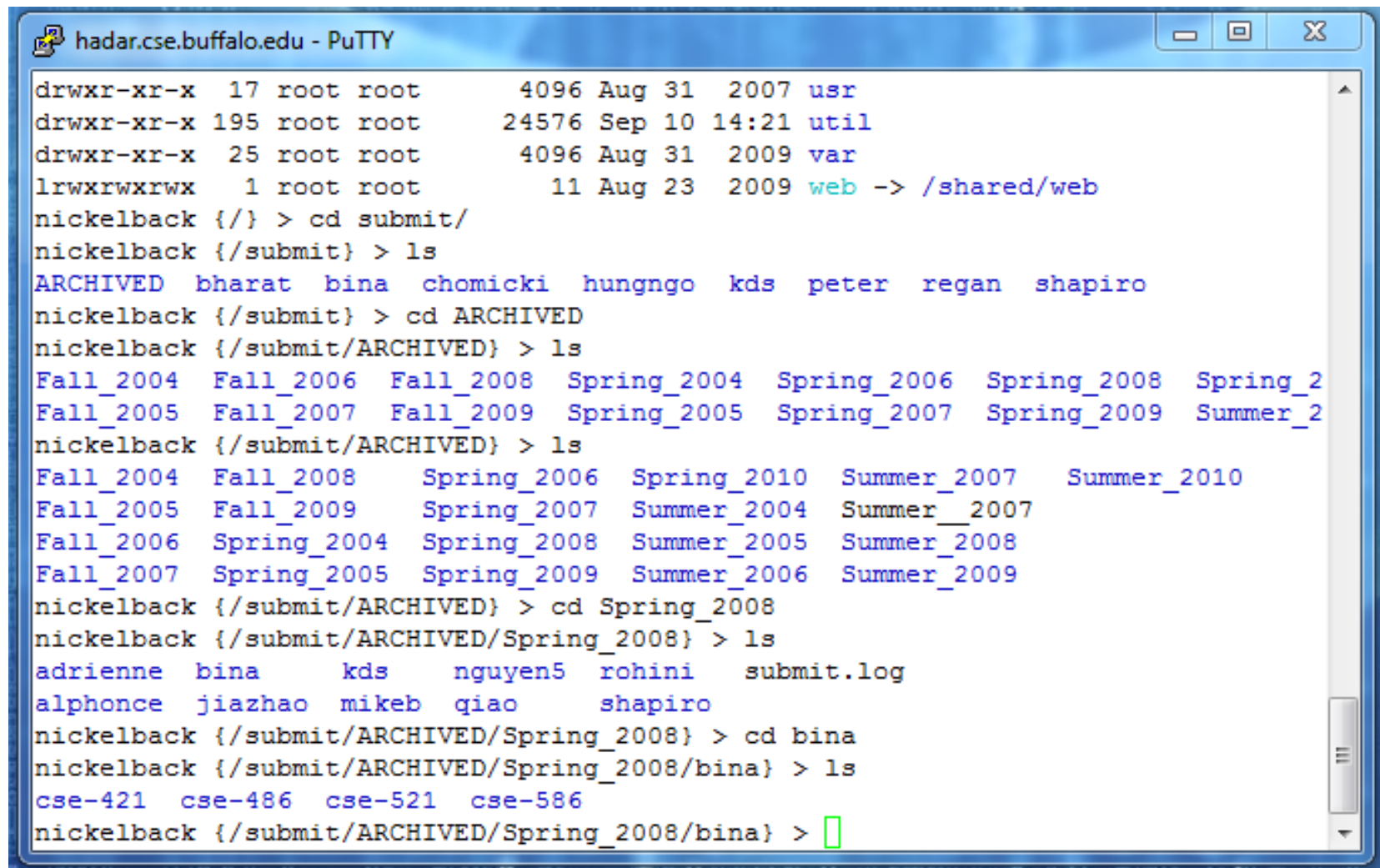
- Cheaters rob good, hardworking students of that "bump"

New World Order – no more bluffing

PDS

- Uses several years of back projects from several different classes
- Matches both structure, content, comments
- Works for many languages
  - Java, C, C++
  - Matlab, LISP (similar results in %cheating)
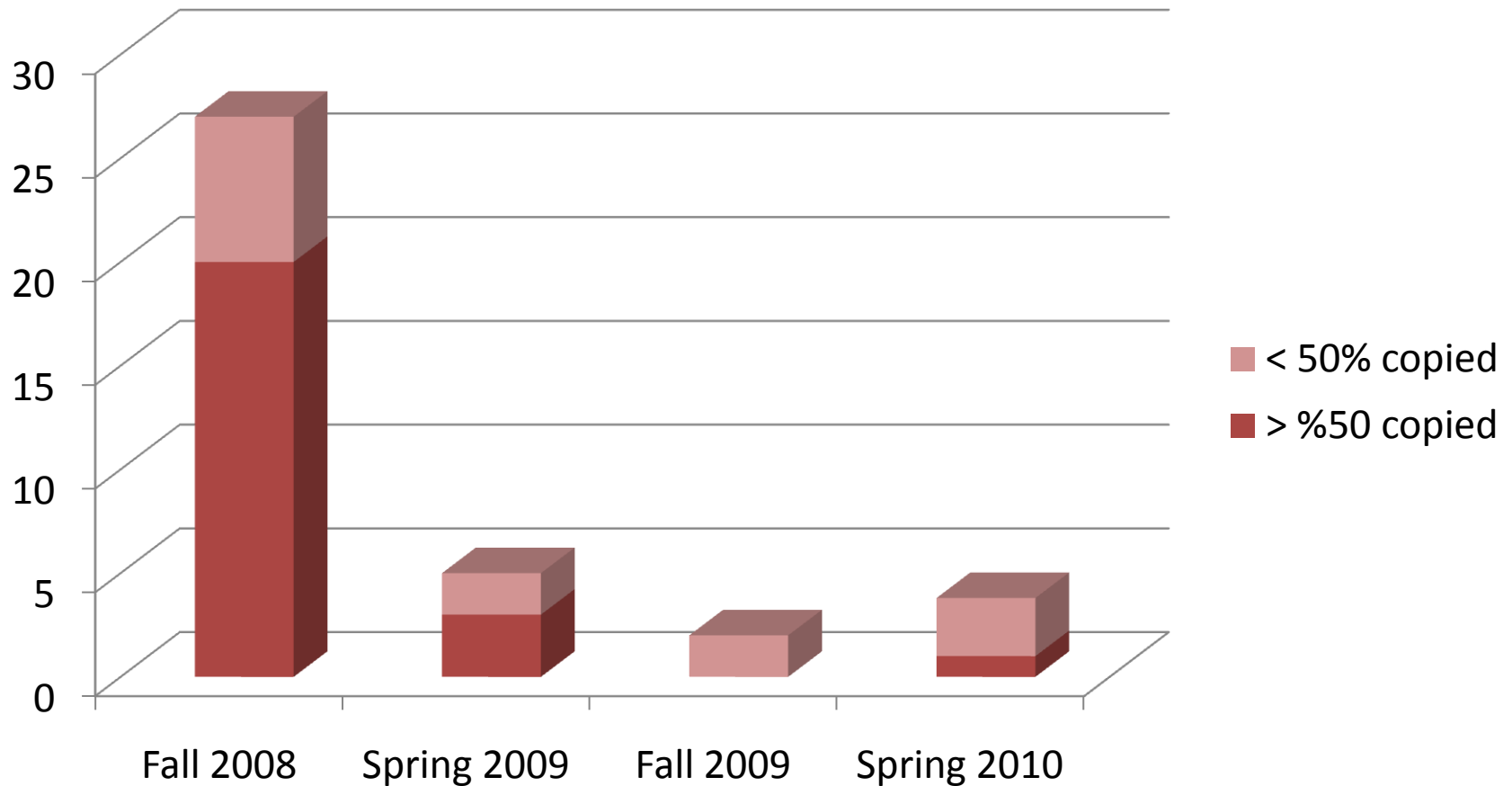 - Perl, Python, Scheme, Prolog

Project Archives go back to Fall 2004 for ALL classes for ALL students

```
hadar.cse.buffalo.edu - PuTTY

drwxr-xr-x  17 root  root        4096 Aug 31  2007 usr
drwxr-xr-x 195 root  root       24576 Sep 10 14:21 util
drwxr-xr-x  25 root  root        4096 Aug 31  2009 var
lrwxrwxrwx   1 root  root          11 Aug 23  2009 web -> /shared/web
nickelback {/} > cd submit/
nickelback {/submit} > ls
ARCHIVED  bharat  bina  chomicki  hungngo  kds  peter  regan  shapiro
nickelback {/submit} > cd ARCHIVED
nickelback {/submit/ARCHIVED} > ls
Fall_2004  Fall_2006  Fall_2008  Spring_2004  Spring_2006  Spring_2008  Spring_2
Fall_2005  Fall_2007  Fall_2009  Spring_2005  Spring_2007  Spring_2009  Summer_2
nickelback {/submit/ARCHIVED} > ls
Fall_2004  Fall_2008    Spring_2006  Spring_2010  Summer_2007   Summer_2010
Fall_2005  Fall_2009    Spring_2007  Summer_2004  Summer__2007
Fall_2006  Spring_2004  Spring_2008  Summer_2005  Summer_2008
Fall_2007  Spring_2005  Spring_2009  Summer_2006  Summer_2009
nickelback {/submit/ARCHIVED} > cd Spring_2008
nickelback {/submit/ARCHIVED/Spring_2008} > ls
adrienne  bina      kds      nguyen5  rohini    submit.log
alphonce  jiazhao  mikeb    qiao      shapiro
nickelback {/submit/ARCHIVED/Spring_2008} > cd bina
nickelback {/submit/ARCHIVED/Spring_2008/bina} > ls
cse-421  cse-486  cse-521  cse-586
nickelback {/submit/ARCHIVED/Spring_2008/bina} >
```

Graph of Plagiarists over 4 semesters

Typical average class size: 100 students

Figure 1 - Main Summary Page – accessed by web browser. The top line shows a project that matched 99% since it was from a student that submitted the same code twice in two different directories. All other lines show percentages of matching in an approximated 4-5K line project. (Names are grayed out to protect the guilty…)
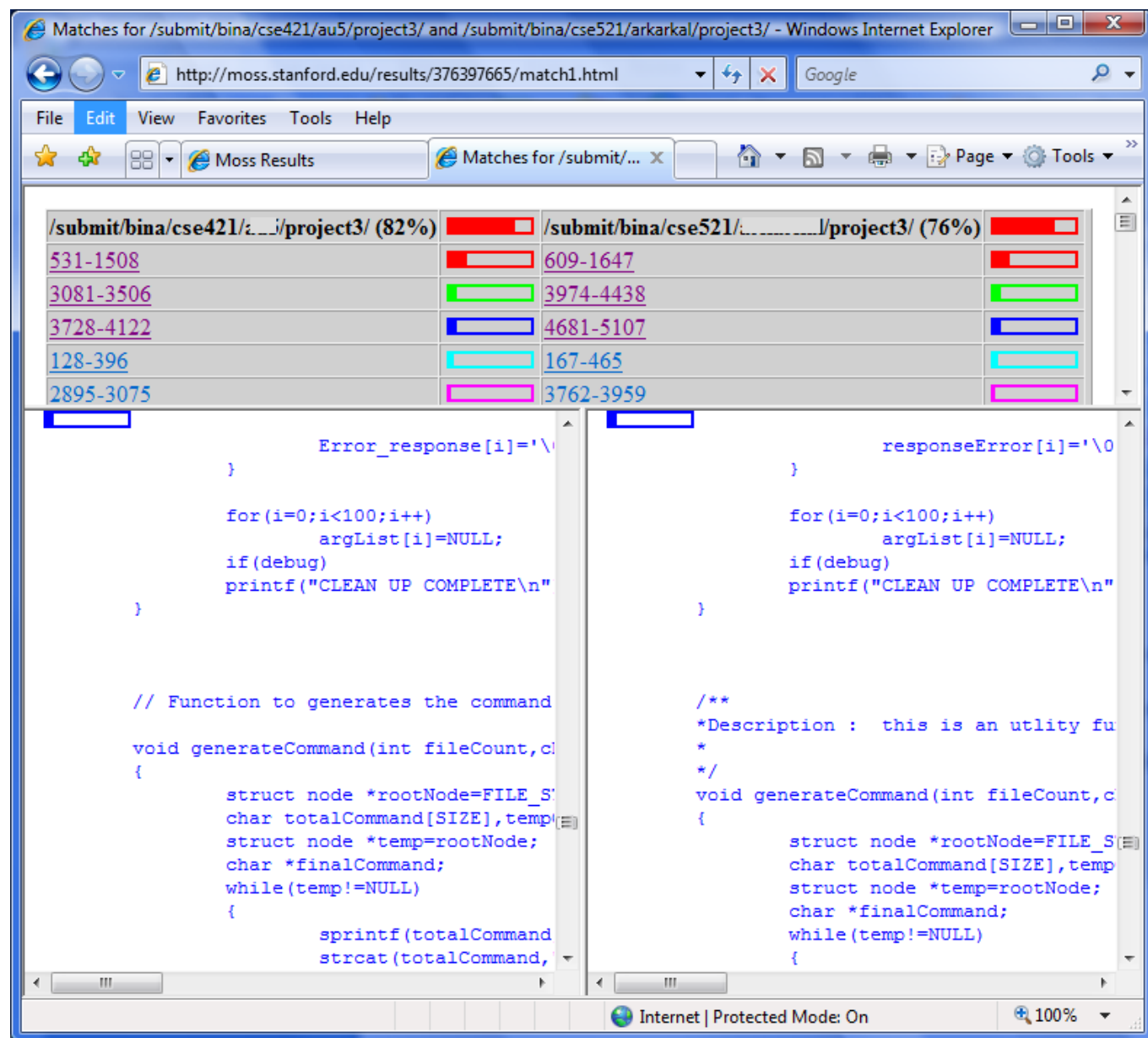
Figure 2. Offending pair code matching window. This shows one student who didn't even attempt to subvert a "diff" command! (They seem to not even call the bluff!) As one student who 'lended' the code out, wrote to me in an email – "3874 lines of code [donno what to say ; I coded so well :) || they copied so bad :( ]"
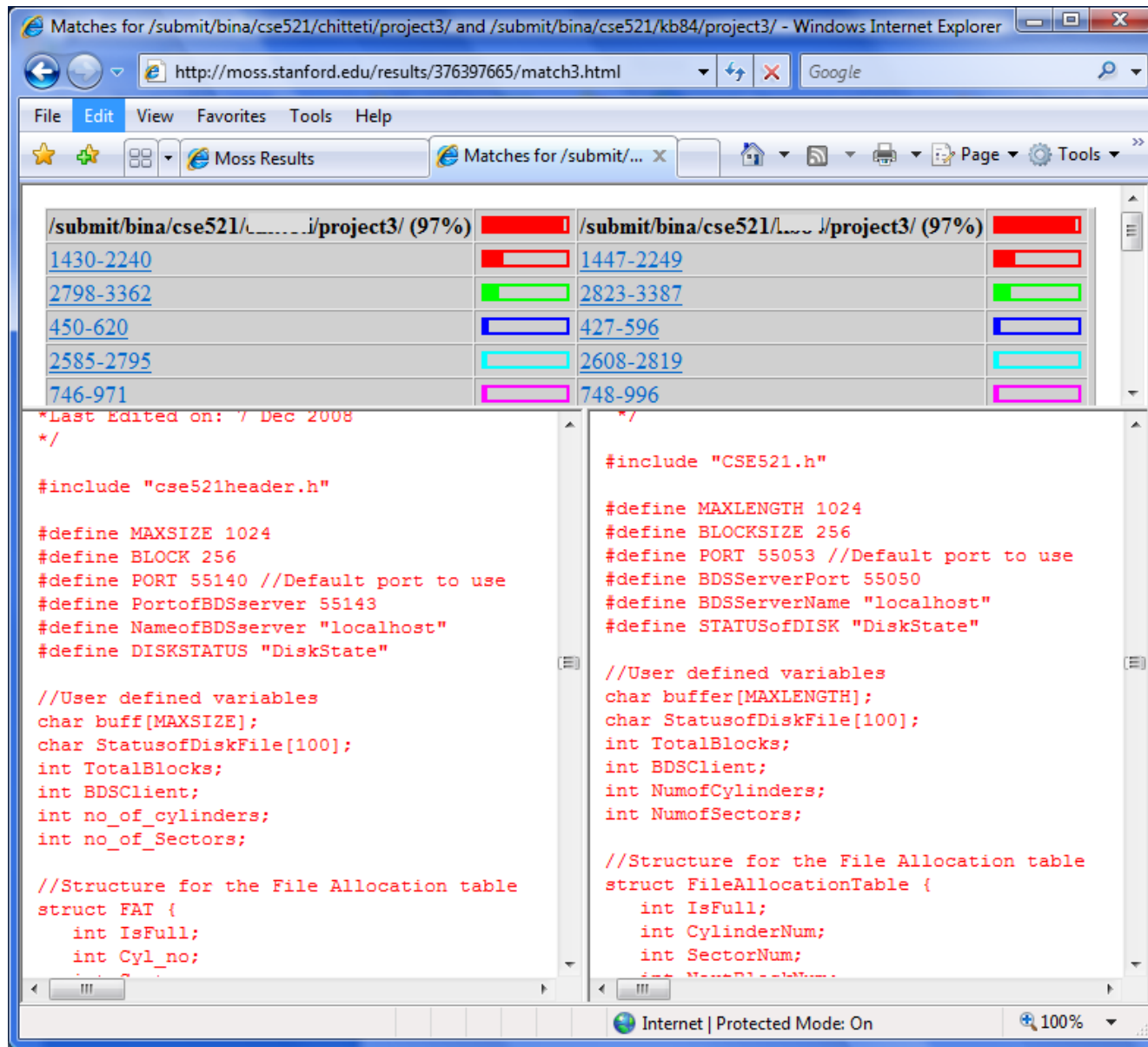
Figure 3. Here is a pair of students that actually both copied from ANOTHER student nearly their entire project. They spent several hours searching through the code and renaming each and every variable and structure field. When first questioned, one of the students claimed she didn't cut and paste or use anyone else's code, she just "collaborated" on the computer next to them.
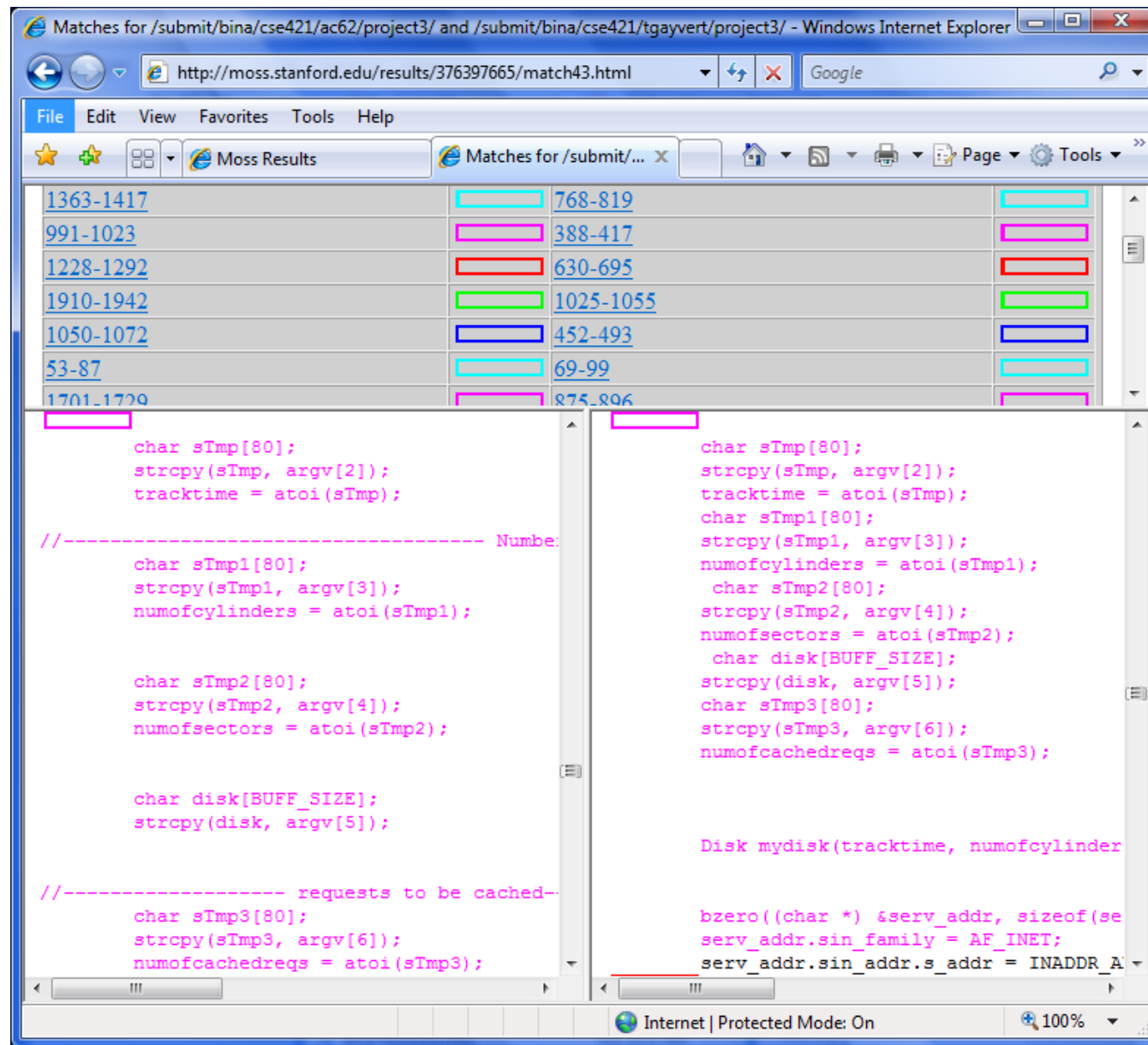
Figure 4. This shows a match for a code block that had different formatting, comments inserted, and even new lines of useful code added. I would never have been able to catch something like this; it's like an NP-Complete problem for TAs, code matches might take a long time to find, but it is much quicker to verify. With MOSS we need only verify!

Initial Conclusion

- Actually *using* the PDS works better than just bluffing about using the PDS

- Why is there still some cheating then?
  - Students didn't attend recitation when we tell them about the PDS
  - Students believe that they can trick the PDS
  - Students believe they can talk their way out of the factual results

Fact #1: You can't trick the PDS

- Initially, everyone denies copying
  - "There must be some mistake…"
  - "This is impossible…"
  - "I was using someone else's computer, maybe I accidentally pasted what was in their clip-board…"

- When faced with side by side, color coded, matched code
  - "Well, I did look at her code a bit…"
  - "I was sick, so my roommate typed for me…"
  - "I used online code, but there was no copyright for it…"
  - Most of them just say: "OK, I did it. I'm sorry. But, I've never done this before!!!"

Fact #2: You can't deny the results

- Coincidence is highly unlikely

"Hey, Bard, I'm the one who wrote all those plays you've been taking credit for, chimpanzee #158,634,765,234,897,555, 345,876,999,..."

See: http://theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf

What to do if you are in a bind?

- Best chance is to communicate with the TA or Professor as early as possible
  - Example #1 : having a child is a good reason for an extension

- Do as much as you can, submit your assignment, ask for more time
 - Example #2 : An "A" TA/student panics!

What not to do!

- Do not steal your roommate's password
- Do not submit your friend's project accidentally!!!
- Do not take someone else's project
  - Rename variables
  - Move code around
  - Insert NOP code
  - Copy small sections from various code bases

Good Rules to use @ CSE

- Do not read someone else's code

- Do not use someone else's code

- Do not download code from the web

- Do not ask someone else to write your code

- Do not collaborate with another student (unless your professor says you can)

- Do not GIVE your code to someone else to help them (you aren't helping them)

…and they all lived happily ever after!

So remember…
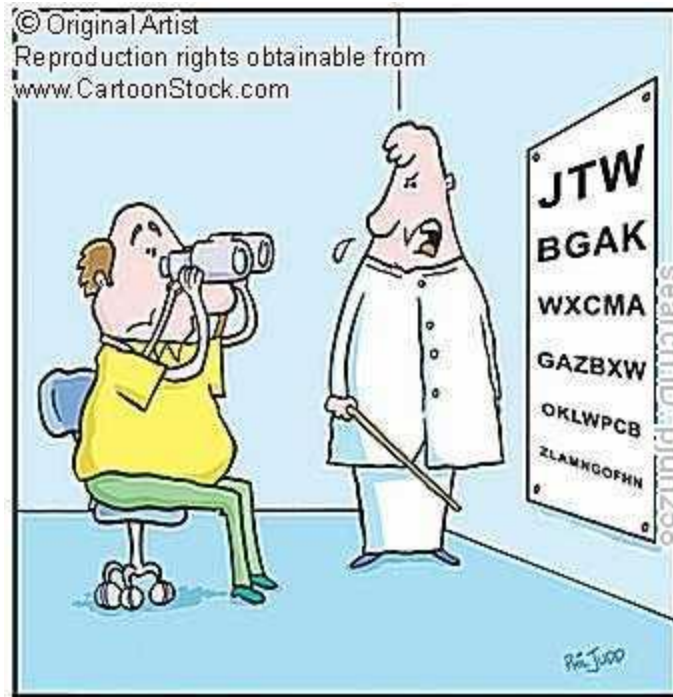
"No cheating!"

…and don't help anyone cheat, it doesn't help them!