Plagiarism-detection software
Clashing intellectual property rights and aggressive vendors yield dismaying results

Cem Kaner, J.D., Ph.D.
Professor of Software Engineering
Florida Institute of Technology

Acknowledgement
This work is partially based on research supported by NSF Grant IIS-0629454: "Learning Units on Law and Ethics in Software Engineering." Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Abstract

Plagiarism is a tedious issue in the academic environment. Many students don't understand why it's such a big deal. Others don't care and just gamble on not getting caught. Many don't understand what it is and find themselves in an unexpected mess. I'll open this talk with a brief discussion of why plagiarism is an important issue, academically and (more important to me) in the outside world. From there, I want to look at the leading plagiarism detection services, what they promise and what they deliver. There is a serious mismatch between the actual effectiveness of these tools and their mythology, and that leads to some counterproductive practices in some schools (maybe even here). Apart from marketing claims that some people might consider too aggressive, one of the problems lies in our professional societies and our journal publishers—and how they restrict rights on the articles that we academics publish. The biggest barrier to us protecting our intellectual property rights (from copyright infringement and plagiarism) might well be the publishers protecting their rights to restrict access to our work.

What is plagiarism?

- Presenting someone else's work as your own:
  - Their words, or
  - Their ideas
- Acknowledgment has to give the reader an understanding of what you actually used:
  - If you take their words, you must use quotes or people will think that this masterful presentation of a concept is yours
    - If you take their words and make minor changes, you are merely covering up the fact that you are copying their words
      - » Some guidelines suggest changing a few words and flipping sentence order to avoid plagiarism (they call this paraphrasing). This is BS. If a reader could recognize where you've taken something from, you've taken too much. If you need their words, use quotes.

What is plagiarism? (2)

- Presenting someone else's work as your own:
  - Their words, or
  - Their ideas
- Acknowledgment of ideas is by inline citation (e.g. Smith, 2003) or by an inline footnote
  - Sometimes you might get the same idea from several sources
    - cite the one(s) you found most useful or the one(s) you think your reader will find most useful
  - Sometimes an idea seems to be floating in the common market of ideas
    - find a source. These ideas didn't appear out of thin air. They cost some people a lot of work, and often, a lot of grief. In addition, the source helps your reader.

What's the big deal?

Academically

- The weakest students (or the most desperate) often plagiarize
  - don't understand the topic
  - out of time
  - For faculty, these are the most important to catch because they are academically dishonest. They ask the professor to reward one student for doing less and worse work than more honest students
- Some students plagiarize to hide a writing skills (or language) problem
- Some students plagiarize because they don't understand the rules
  - big cultural differences
- Impossible for faculty to distinguish fairly and consistently among these, and so most of us adopt a set of rules for plagiaristic behavior, announce them, and then enforce them

Big deal? -- Academic

Academic plagiarism by faculty is shockingly widespread

- JAST -- over 50% of submissions were plagiarized
  - including by some people who **really** should have known better
- "several" submissions to conferences were plagiarized, conference organizers differ on how they deal with this
- It is hard to detect plagiarism and so a fair bit gets through only because it takes time and effort
- Several cases published of 100% plagiarism (X takes an entire paper from Y).
  - This most recently happened to me while we were doing research for this series of papers
- Lose grants, lose jobs, lose reputation

Plagiarism of student work or practitioner work is still plagiarism.

Big deal? -- In industry

Plagiarism is a bigger deal in industry than people realize

- Yes, some people get caught
- Lose degrees, lose jobs, lose grants, lose elections

The bigger issue is that people make enemies

- It can be too socially awkward to accuse a peer of plagiarism publicly and so this is rarely done
- instead people
  - gossip about it
  - form (or are reinforced in their formation) of cliques (I don't want to work with you)

Plagiarism in industry

- Plagiarism of ideas might be more common and is more annoying than plagiarism of words
  - magazines discourage ALL citations
  - book publishers discourage inline citations
  - consultants are afraid to share credit (competition, or loss of aura of infallibility)
  - tracking down the source of your work takes time and skill.
    - time: for consultants, time is money
    - skill: (sigh)
- Many movements struggle because they present innovative ideas, gain a lot of followers, but fail to credit those parts of their history that were developed by others
  - it alienates the old guard (and other innovators) for no good reason

Plagiarism in industry

- Many movements struggle because they present innovative ideas, gain a lot of followers, but fail to credit those parts of their history that were developed by others
  - it alienates the older practitioners (and other innovators) for no good reason
    - you aren't listening to me
    - these issues have been discussed before, you are still talking at the first, simplistic level
    - we made a lot of progress on that and you've missed it
    - you are setting up a straw man
    - you are a thief
  - often, the result is non-discussion, and movements fade away via passive aggression
  - But the undertone is often very angry

What about self-plagiarism?

There is a lot of discussion among publishers about "self-plagiarism"

- Republishing work of yours in another journal
- Seen by some as dishonest because it influences statistics that are used by some people as indicators of academic quality (quality = quantity?)
- I don't think this is plagiarism.
  - under some circumstances, it might be in some ways academically dishonest, but we should give this a different name.

Plagiarism-detection services -- Turnitin.com

Turnitin

TAG http://www.turnitin.com/static/plagiarism.html Google

Read Later

A Complete Search

Every paper submitted is returned in the form of a customized Originality Report. Results are based on exhaustive **searches of billions** of pages from both current and archived instances of the internet, millions of student papers previously submitted to Turnitin, and commercial databases of journal articles and periodicals.

Plagiarism "prevention"

Recommended by the services:

- Let the student submit the paper to the service
- get their own plagiarism check, privately
- modify the paper, until the plagiarism checker says it is OK

Plagiarism-detection services -- Turnitin.com

Turnitin

http://www.turnitin.com/static/plagiarism.html

Google

Read Later

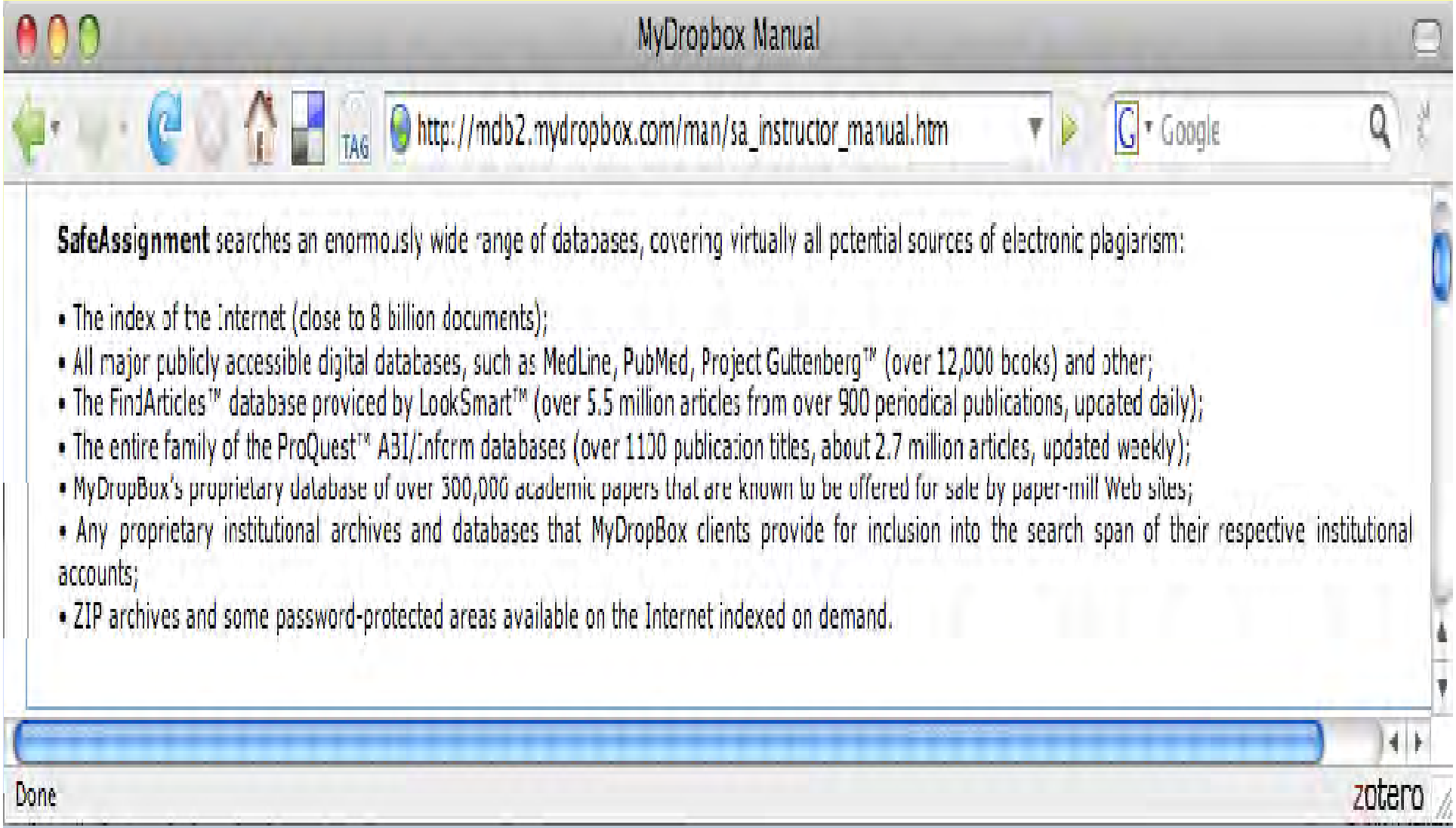
A Detailed Report

Any matches uncovered between submitted papers and source material is detailed in an intuitive and unambiguous format, allowing educators to spend time addressing plagiarism's causes rather than searching for it.

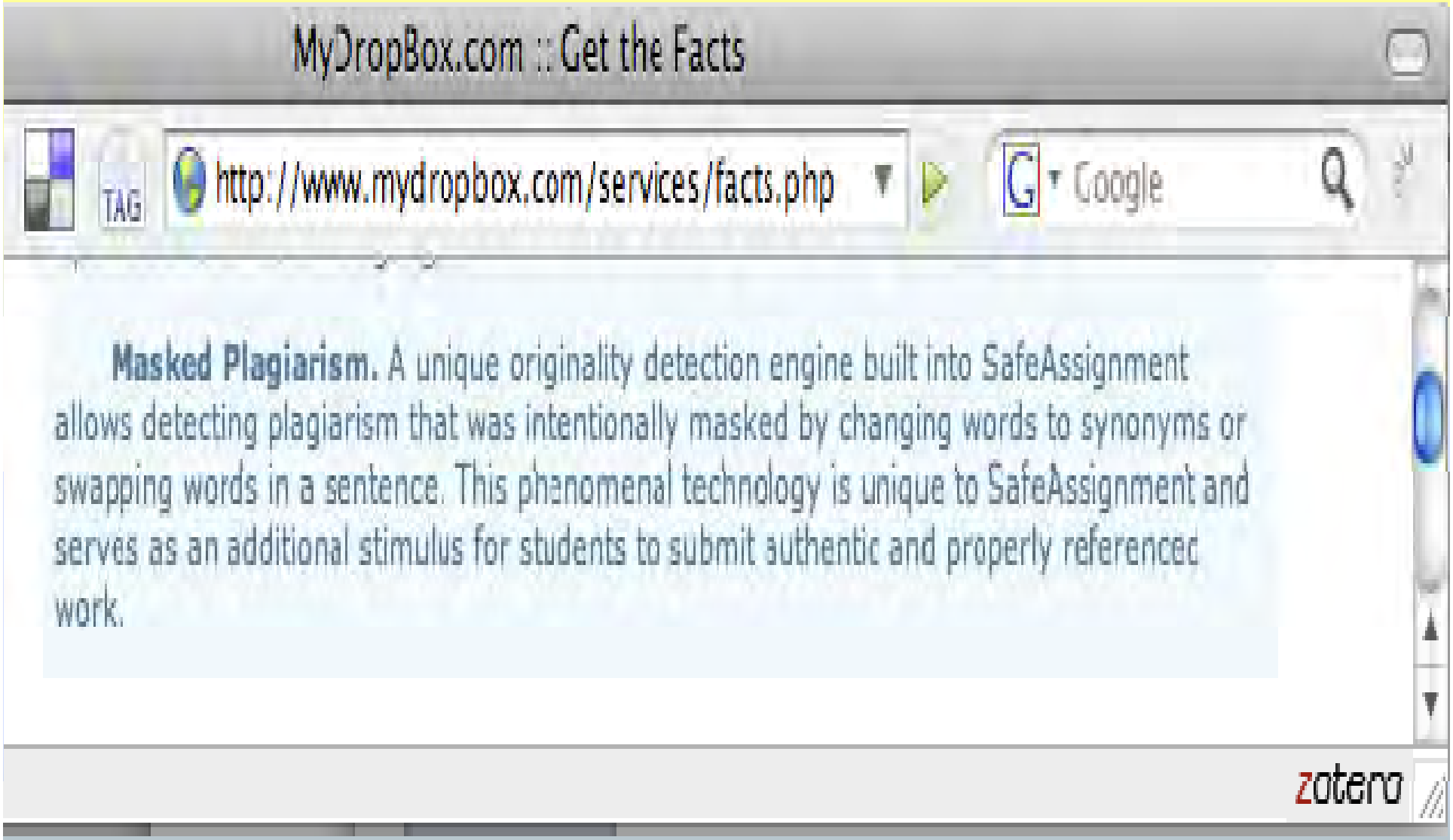
Sample Report

Turnitin's plagiarism prevention is often so successful that institutions using our system on a large scale see measurable **rates of plagiarism drop to almost zero.**

MyDropBox (SafeAssignment) (in BlackBoard)



MyDropBox (SafeAssignment) (in BlackBoard)

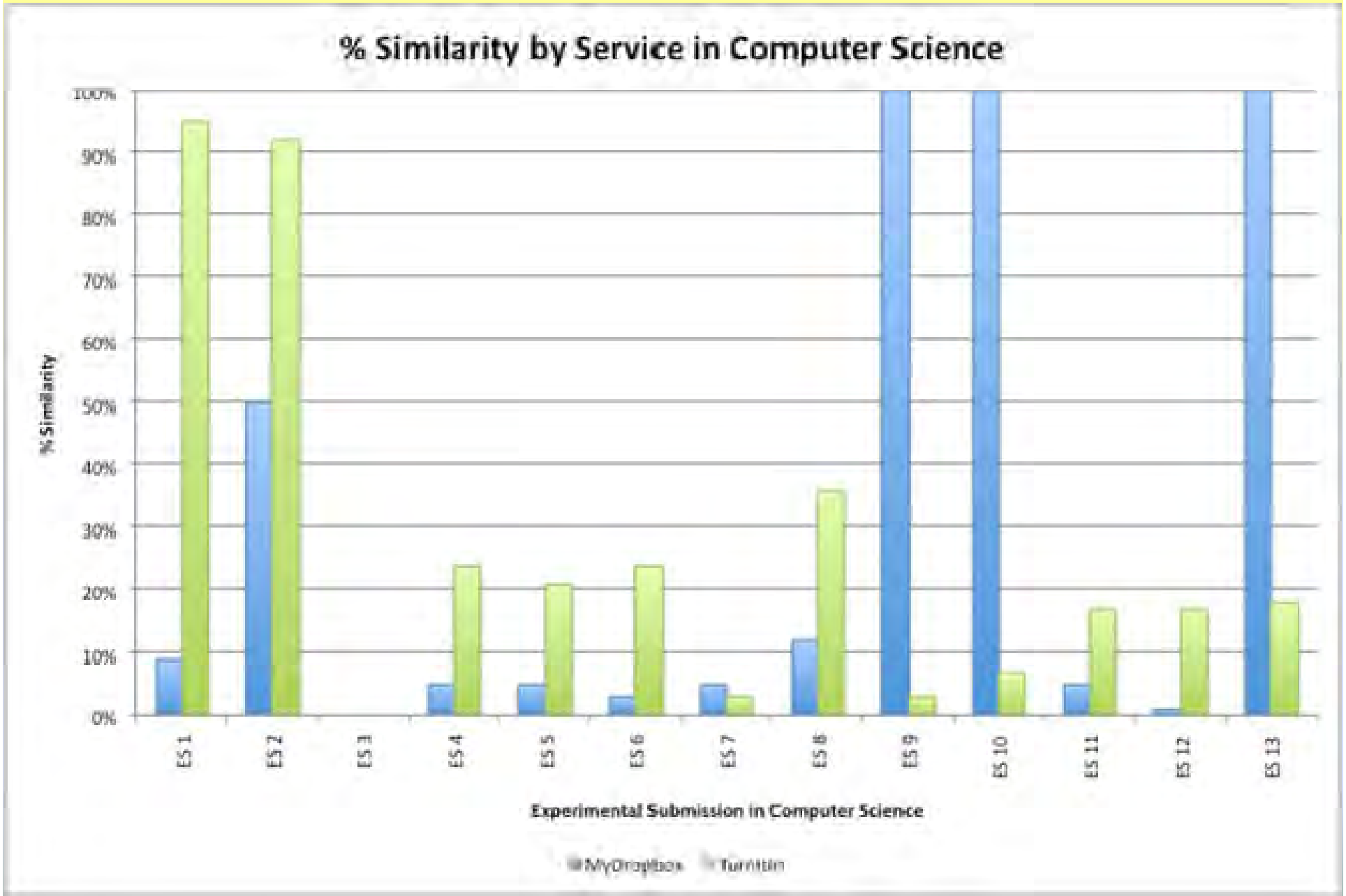


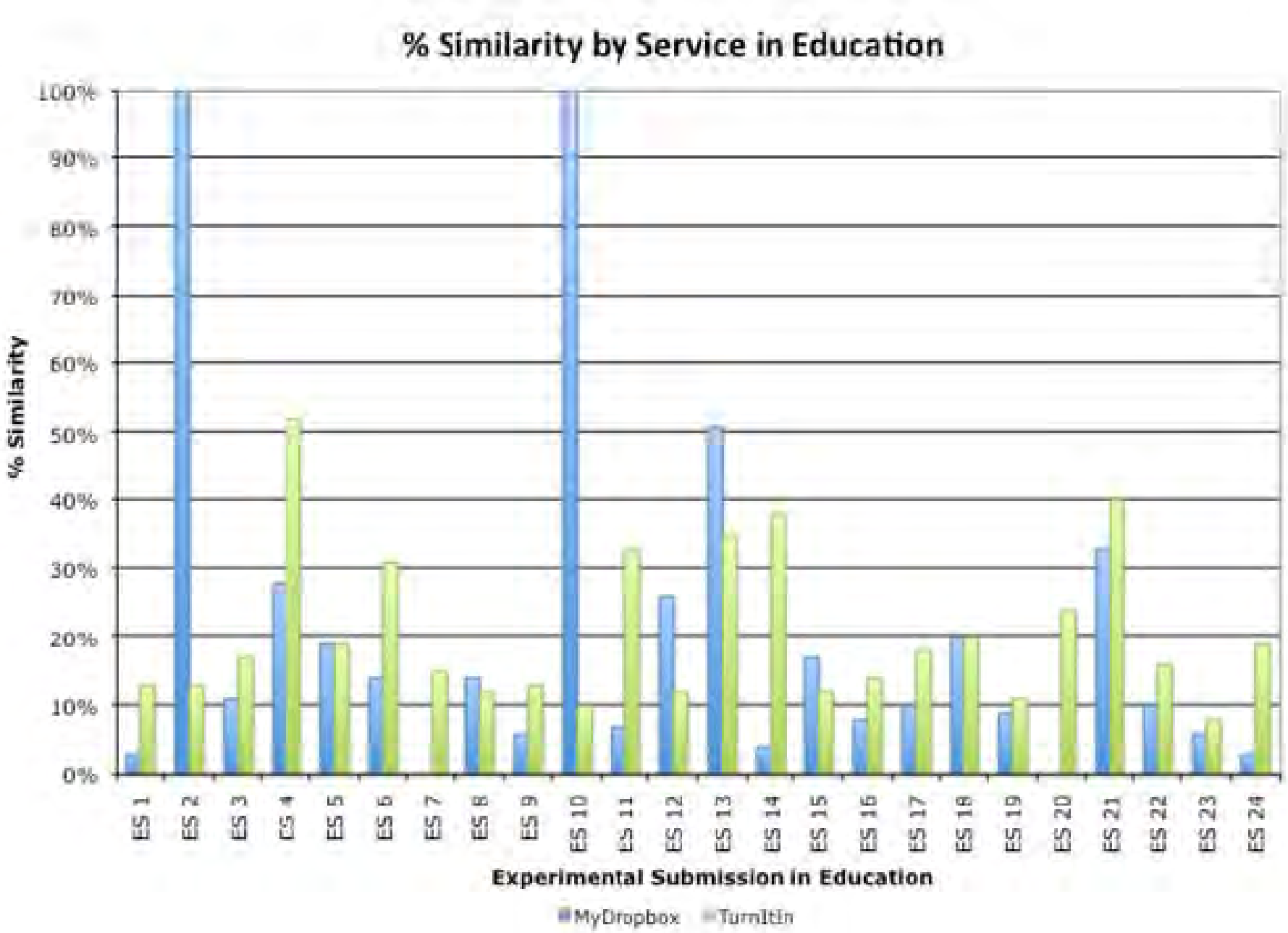
Research questions

1. Will services identify plagiarized submissions?
2. Do these services have blind spots?
3. How widely used are these services?
4. What are user perceptions regarding effectiveness?

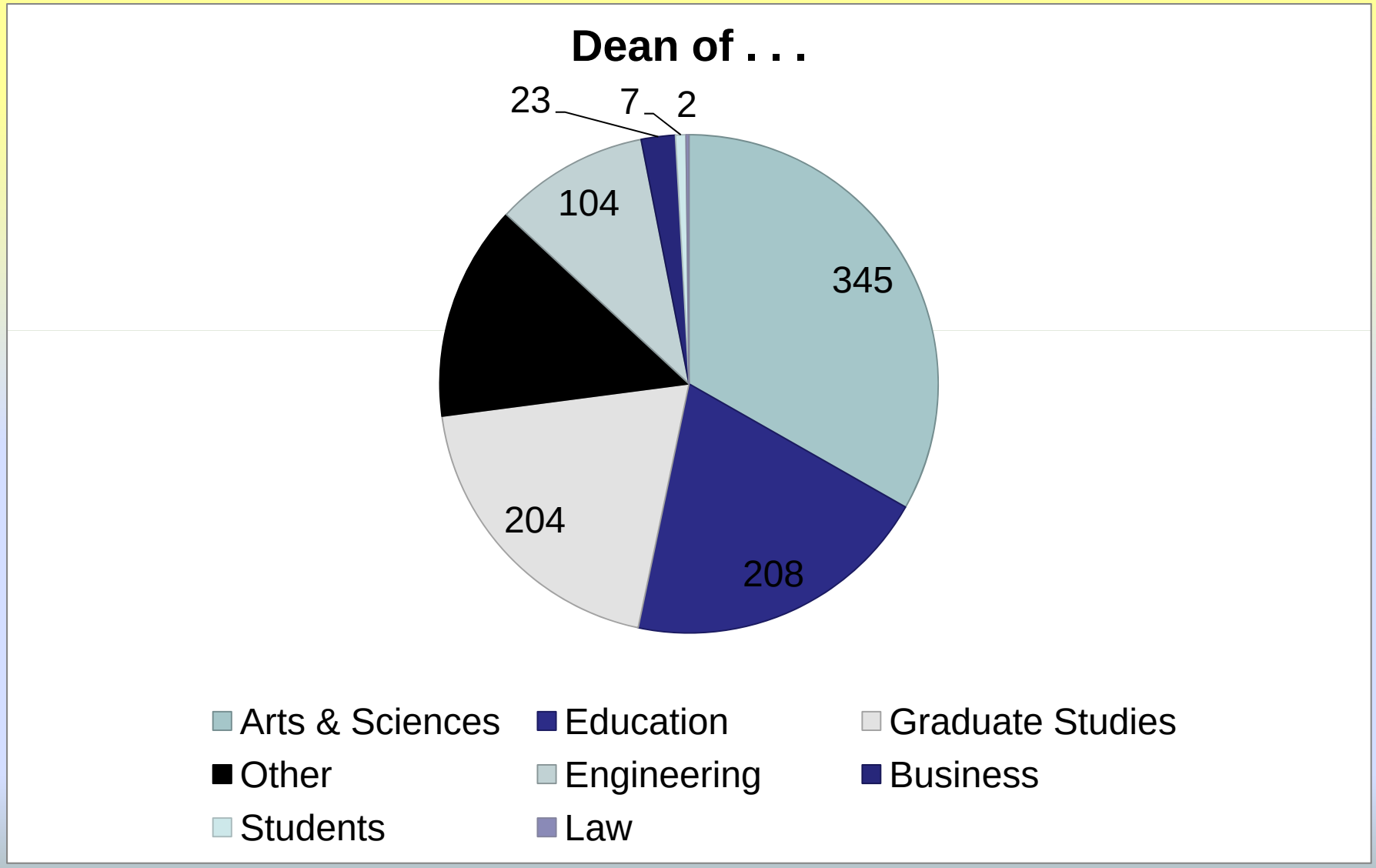
Method

Downloaded and submitted unmodified previously published papers to each service to generate report


% Similarity by Service in Computer Science
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% Similarity
ES 1
ES 2
ES 3
ES 4
ES 5
ES 6
ES 7
ES 8
ES 9
ES 10
ES 11
ES 12
ES 13
Experimental Submission in Computer Science
MyDropBox
Turnitin


% Similarity by Service in Education
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% Similarity
ES 1
ES 2
ES 3
ES 4
ES 5
ES 6
ES 7
ES 8
ES 9
ES 10
ES 11
ES 12
ES 13
ES 14
ES 15
ES 16
ES 17
ES 18
ES 19
ES 20
ES 21
ES 22
ES 23
ES 24
Experimental Submission in Education
MyDropbox
TurnItIn

Survey of deans (n = 954, 74% use services)



Usage

79% use Turnitin

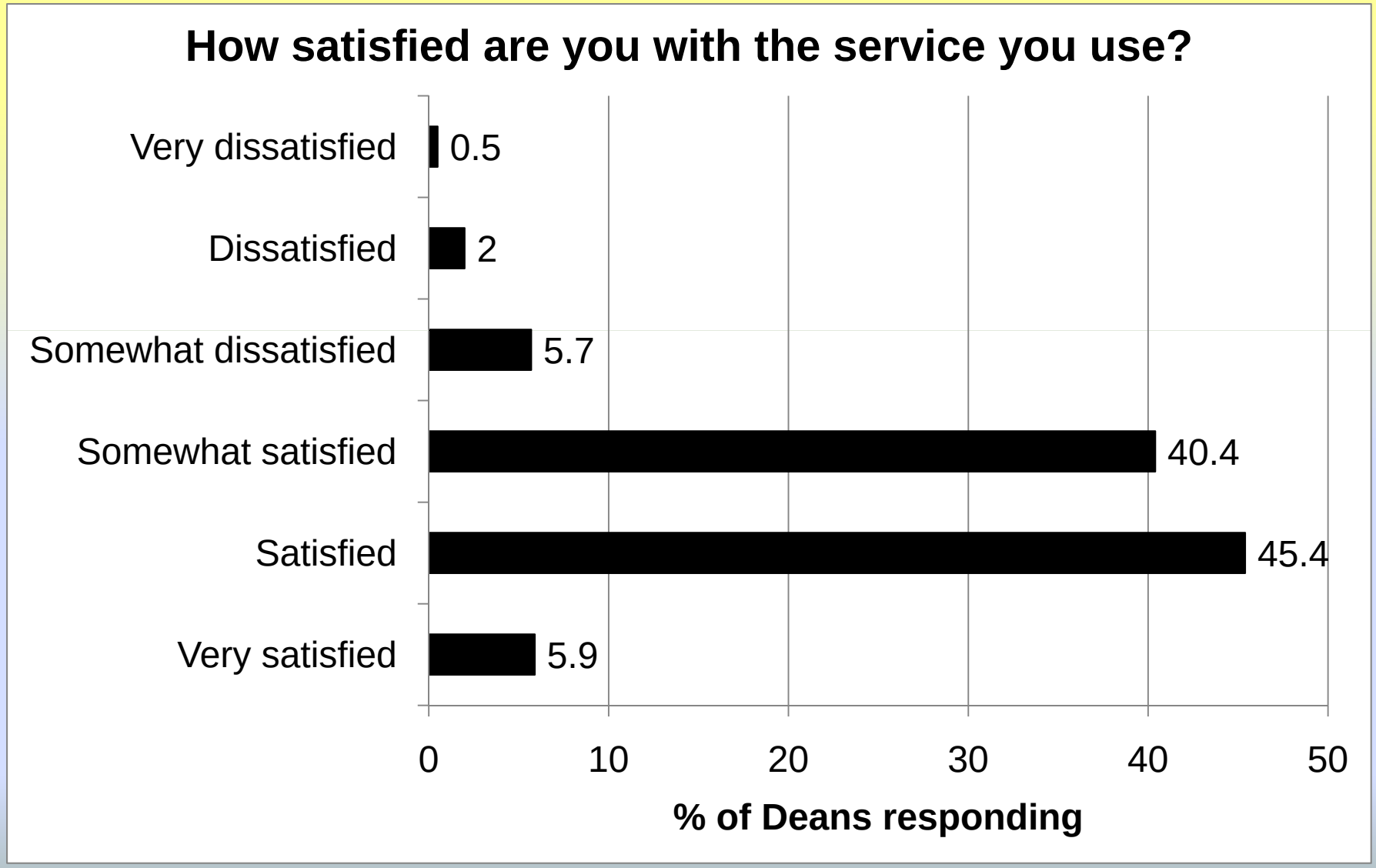
26% use MyDropBox (maybe more now...)

21% use other

What commonly triggers use?

- Routine submission of assignments 60.4%
- Routine checks of theses or dissertations 22.3%
- Routine self-evaluation by students 14.%1
- Suspicion of plagiarism 89.7%

How satisfied are you with the service you use?



How well or poorly does the service you use cover . . .	Very Well	Moderately Well	Somewhat Well	Somewhat Poorly	Moderately Poorly	Very Poorly
Popular sources?	14.1%	35.7%	34.3%	11.0%	3.0%	2.0%
Internet?	23.4%	38.2%	26.9%	7.4%	2.9%	1.2%
Academic literature from professional societies?	11.4%	44.0%	28.8%	11.2%	3.4%	1.2%
Academic literature from commercial publishers?	12.7%	43.8%	30.4%	9.3%	3.0%	0.8%
Dissertations & theses?	16.8%	38.0%	29.7%	9.6%	3.4%	2.5%

A critical underlying problem

Access to academic literature

- proprietary publication
  - we get paid by the government to produce research
  - that we give to publishers at no charge
  - and they charge for access
  - so even we can't check for plagiarism, even of our own work.
- several professional societies have been leaders in advocating for MORE digital rights management and greater protection of THEIR databases

New initiative

Collaboration with some academic publishers

- crossref, crosscheck
- service from the same publisher as turnitin, but for commercial and copyright-lawyer use (and maybe for publishers of the databases who allow their articles to be used) (but as a reviewers I've never had access to this)
- allegedly, this is gradually moving into turnitin, but I have seen no evidence of it and I have apparently misunderstood scheduling suggestions from phone calls

IF WE DON'T DEMAND MORE, WE WON'T GET MORE

- I publish at a minimum in the traditional journals, partially out of disgust with the public-turns-proprietary system
- There are open journals in science and education, though most don't yet have the prestige academics prefer (even though it is easier to find the work in them)

About Cem Kaner

- Professor of Software Engineering, Florida Tech
- Research Fellow at Satisfice, Inc.

I've worked in all areas of product development (programmer, tester, writer, teacher, user interface designer, software salesperson, organization development consultant, as a manager of user documentation, software testing, and software development, and as an attorney focusing on the law of software quality.)

Senior author of three books:

- *Lessons Learned in Software Testing* (with James Bach & Bret Pettichord)
- *Bad Software* (with David Pels)
- *Testing Computer Software* (with Jack Falk & Hung Quoc Nguyen).

My doctoral research on psychophysics (perceptual measurement) nurtured my interests in human factors (usable computer systems) and measurement theory.