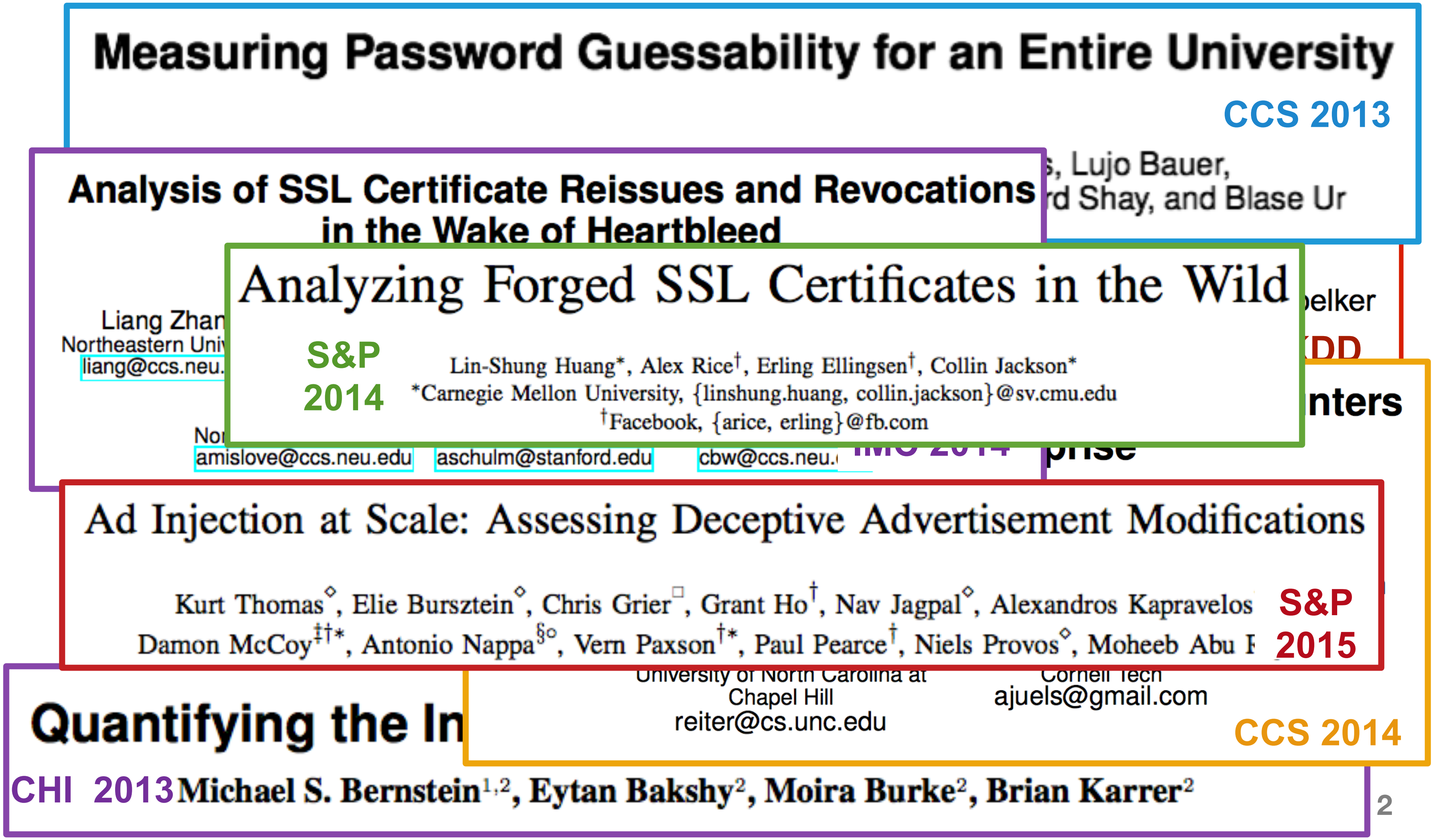
Don't Show Me Yours, I Won't Show You Mine: Security Research with Non-Public Data

Michelle Mazurek and
Tudor Dumitras

University of Maryland

http://ter.ps/hotsec

Non-public data increasingly prevalent



Measuring Password Guessability for an Entire University

CCS 2013

…, Lujo Bauer,
…rd Shay, and Blase Ur

Analysis of SSL Certificate Reissues and Revocations in the Wake of Heartbleed

Liang Zhan…
Northeastern Uni…
liang@ccs.neu.…

Nor…
amislove@ccs.neu.edu aschulm@stanford.edu cbw@ccs.neu.…

IMC 2014

…pelker

…KDD

…nters
…prise

Analyzing Forged SSL Certificates in the Wild

S&P 2014

Lin-Shung Huang*, Alex Rice†, Erling Ellingsen†, Collin Jackson*
*Carnegie Mellon University, {linshung.huang, collin.jackson}@sv.cmu.edu
†Facebook, {arice, erling}@fb.com

Ad Injection at Scale: Assessing Deceptive Advertisement Modifications

Kurt Thomas◇, Elie Bursztein◇, Chris Grier□, Grant Ho†, Nav Jagpal◇, Alexandros Kapravelos
Damon McCoy‡†*, Antonio Nappa§○, Vern Paxson†*, Paul Pearce†, Niels Provos◇, Moheeb Abu R…

S&P 2015

University of North Carolina at Chapel Hill
reiter@cs.unc.edu

Cornell Tech
ajuels@gmail.com

CCS 2014

Quantifying the In…

CHI 2013

Michael S. Bernstein[1,2], Eytan Bakshy[2], Moira Burke[2], Brian Karrer[2]

Why not make all data public?

- Confidentiality, privacy or security concerns
  - May leak PII (e.g., users of social network)
  - May cause harm (passwords, vuln disclosure, IRB, cars)
  - Source may require confidentiality (e.g., industry data)
- Cost concerns
  - Collection may be expensive (e.g., car hacking, sensor deployments for measuring censorship)
- Practical concerns
  - Data may be too big (e.g., 20+ TB in WINE)
  - May be useless if released (e.g., Cybercrime)

Why care about non-public data?

Reproducibility!

… but what do we mean by reproducible?

What is (or isn’t) reproducible?

- Difficult (time, money, resources)

Difficult to reproduce

- Time, resources, connections
  - Years infiltrating a botnet
  - Buying expensive equipment
  - Relationships with Google, Yahoo!, Facebook, etc.
- What are the incentives?
  - Collector: Amortize collection over several papers
  - Why spend resources reproducing?

What is (or isn't) reproducible?

- Difficult (time, money, resources)
- Precise data source is not available

Reproducing, but differently

- With a different organization
  - ! Passwords not with Yahoo! or CMU
  - ! Political malware with different NGOs [Hardy+ 2014]
  - ! Malware encounters with different enterprise [Yen+ 14]
- With newer data
  - – Measure cybercrime again later
- Analogous to sampling?
  - – New data, hopefully same result
  - – New insights as data changes
  - – What is your data representative of?

What is (or isn’t) reproducible?

- Difficult (time, money, resources)
- Precise data source is not available
- Natural experiment

Natural experiment

- Response to a specific event
  - Can’t be reproduced in a controlled way
  - Heartbleed, Debian low-entropy bug [Zhang+ 2014]
  - Leaked criminal data

What data access is needed?

- Goal 1: Independent verification
  - Must reproduce all steps, including collection
  - Data changes provide insights about threats
    - ! [Sabottke+, USENIX Sec 15] reexamines [Bozorgi+ 2010]

What data access is needed?

- Goal 2: Enable follow-on research
  - Reference benchmarks, detailed comparisons
    - ! DARPA IDS, Android Malware Genome, Malicia
    - ! Patching measurments: [Durumeric+ 2014] vs. [Yilek+ 2009]
  - Incentives against using reference data
    - Datasets age quickly
    - Steer research direction to quirks of data

Value added by non-public data

- Validate other research strategies
  - ! [Fahl+ 2013], [Mazurek+ 2013]
- Scale and coverage
  - – Rare events, large network effects
  - ! FB m-i-t-m [Huang+ 2014], ad injection [Thomas+ 2015], invisible audience [Bernstein+ 2013]
- Insights otherwise unavailable
  - ! Malware encounters, password expiration [Zhang+ 2013], social media bias in hiring [Acquisti+ 2013]

Emerging data sharing models

- Define formal process for access
  - DHS PREDICT: https://www.predict.org/default.aspx
  - Symantec WINE: http://ter.ps/8ga
- Allow queries on restricted data
  - Differential privacy?
- Restrict derived data released
  - WINE: See raw data on-site, only take aggregate out
- Access tiers for different users/needs

http://ter.ps/hotsec to contribute anonymously!

DISCUSSION PROMPTS

What should our standards be?

- What reproducibility is required, encouraged?
  - Require detailed methodology?
  - Require explanation of why data not shared?
- Should we draw a line somewhere
  - Other than ethics?
- How do we assess results from non-public data?
- How can we combat rich-get-richer problem?

How to encourage more sharing?

- Carrots vs. sticks
  - Best dataset prize
  - “Seal of approval” and conference publication
  - Limit acceptances, awards?
  - Should we develop an official policy?
- Releasing after delay
  - Maybe some confidentiality issues fade?
  - But is data still useful?
- Without “sacrificing [valuable research] on the altar of openness”

What are best practices for sharing?

- Given scale issues, given privacy restrictions, etc.
- Other sharing models we didn't discuss?
  - Bidirectional sharing for comparison?
  - Pooling several datasets together?
- Examples that surprisingly didn't work

Handling evaluation issues

- Quality of work using non-public data
- When your non-public data is better than a paper reporting with public data
- Comparing results: Changes dues to new approach? New data source? Combination?