Collecting Data about Internet Censorship and Surveillance on a Global Scale

Jed Crandall
crandall@cs.unm.edu

What I hope to convince you of...

- Internet censorship and surveillance is a global phenomenon that can't be fully understood with current data sets
- TCP/IP side channels are a promising way to collect relevant data that is global and longitudinal in scale

What are people trying to do?

- W.r.t. human rights
  - Read the news
  - Political discussion
  - Advocacy
  - Organize protests
  - Support groups (*e.g.*, LGBT, abuse)
  - Document and call attention to atrocities
- W.r.t. U.S. interests
  - Access, *e.g.*, Radio Free Asia (rfa.org)
  - Free trade

What are people trying to do?

- W.r.t. human rights
  - ~~Read the news~~ Many countries block nytimes.com, *etc.*
  - ~~Political discussion~~ Many social media sites filter results
  - ~~Advocacy~~ Surveillance of advocacy groups in the U.S. and abroad
  - ~~Organize protests~~ Freedom of assembly often a target
  - ~~Support groups (*e.g.*, LGBT, abuse)~~ Overblocking (U.S., middle east)
  - ~~Document and call attention to atrocities~~ YouTube blocked
- W.r.t. U.S. interests
  - ~~Access, *e.g.*, Radio Free Asia (rfa.org)~~ Blocked in many countries
  - ~~Free trade~~ U.S. companies are excluded “due to local laws”

Current data sets are very limited

- Tests executed by volunteers
  - Open Net Initiative's profiles of 62 countries
    - 196 countries in the world (62 is 32%)
    - Sometimes one or two tests per country
  - Tor's Open Observatory of Network Interference
    - 68 countries
    - specific to Tor
- Data that is opportunistic
  - Crandall *et al.*, CCS 2007
  - Kattack *et al.*, SIGCOMM 2014
  - Knockel *et al.*, FOCI 2015
  - Many, many more examples, but data sets are “MPU” sized

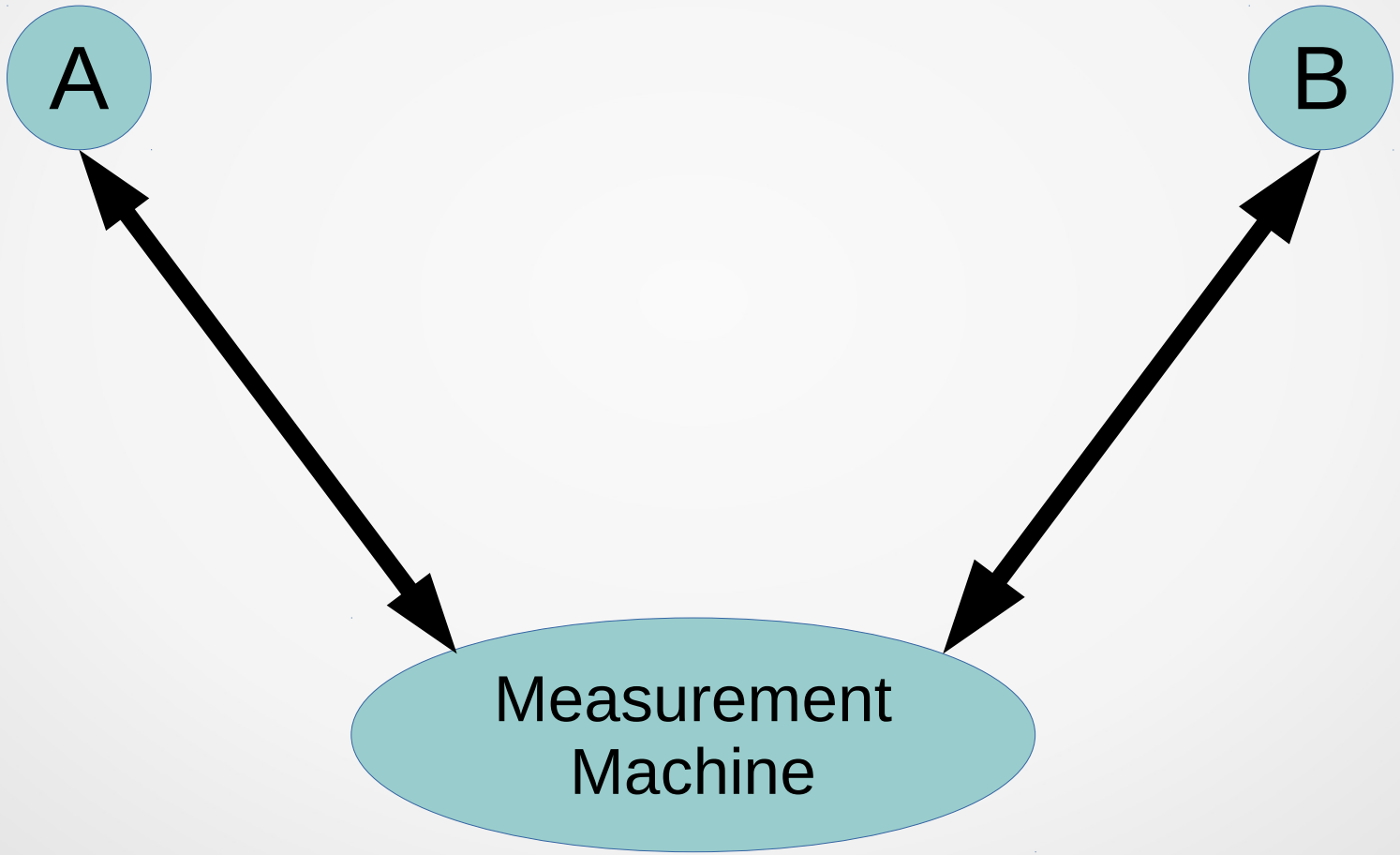
What if we could collect data everywhere, all the time?

- What is blocked, when, and where?
  - E.g., nytimes.com, rfa.org, twitter.com, facebook.com, LGBT sites, YouTube, Tor relays
- See the evolution over time of facebook.com and twitter.com, or Content Distribution Networks, becoming available in more countries
  - Then we'd know what kinds of agreements to look for
- More granularity
  - Libraries and schools in the U.S.
  - Do different ISPs within a country block circumvention technologies using different methods?

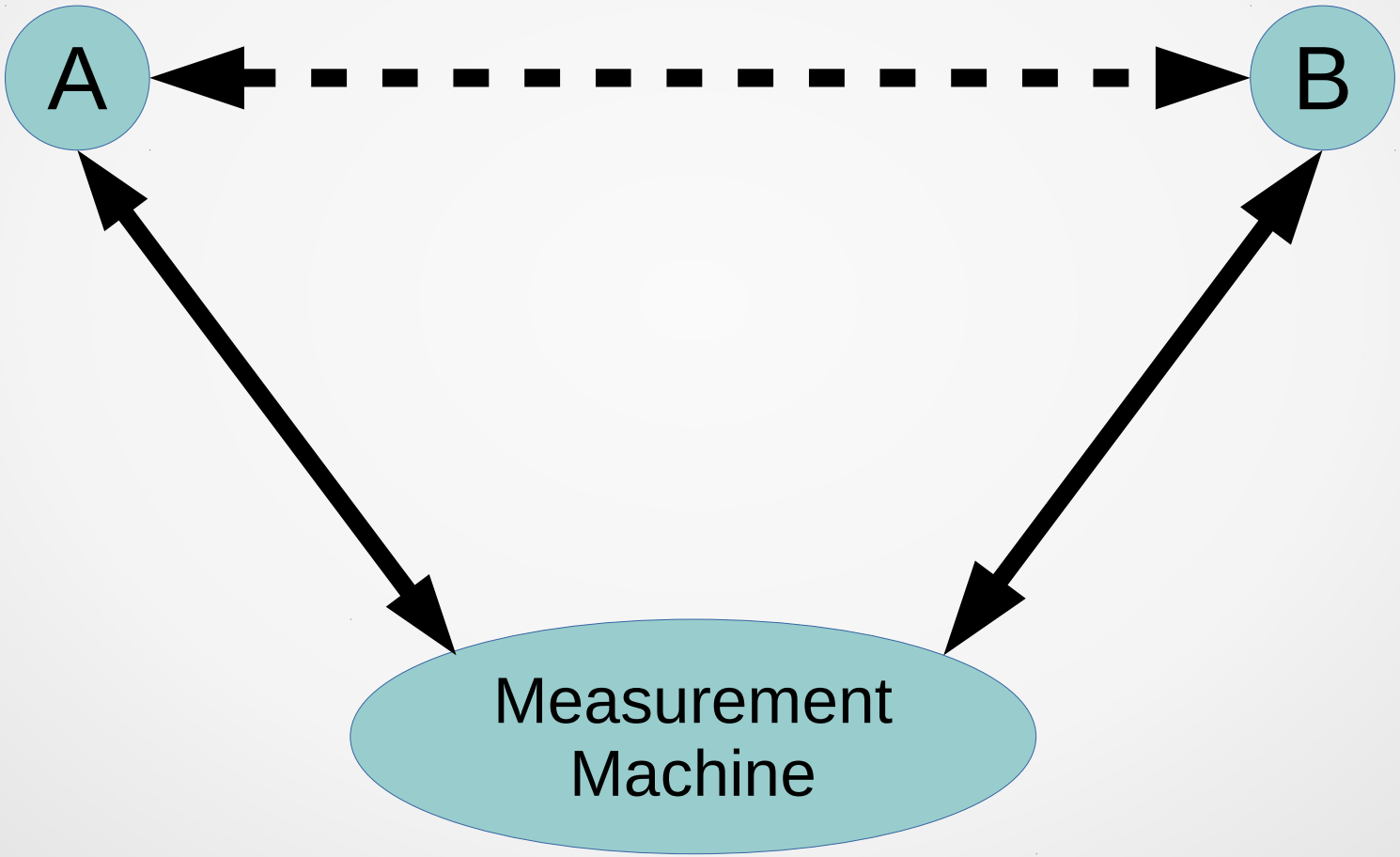
Good news and bad news

- Bad news: we can only do this for DNS (the Domain Name System) and IP (the Internet Protocol) so far
- Good news: we can do this for two important aspects of Internet censorship!
  - DNS: other researchers are working on this
  - IP: what this talk is about
- Note: Encore from SIGCOMM 2015 can do off-path measurements in the application layer, but they essentially run code on a user's machine

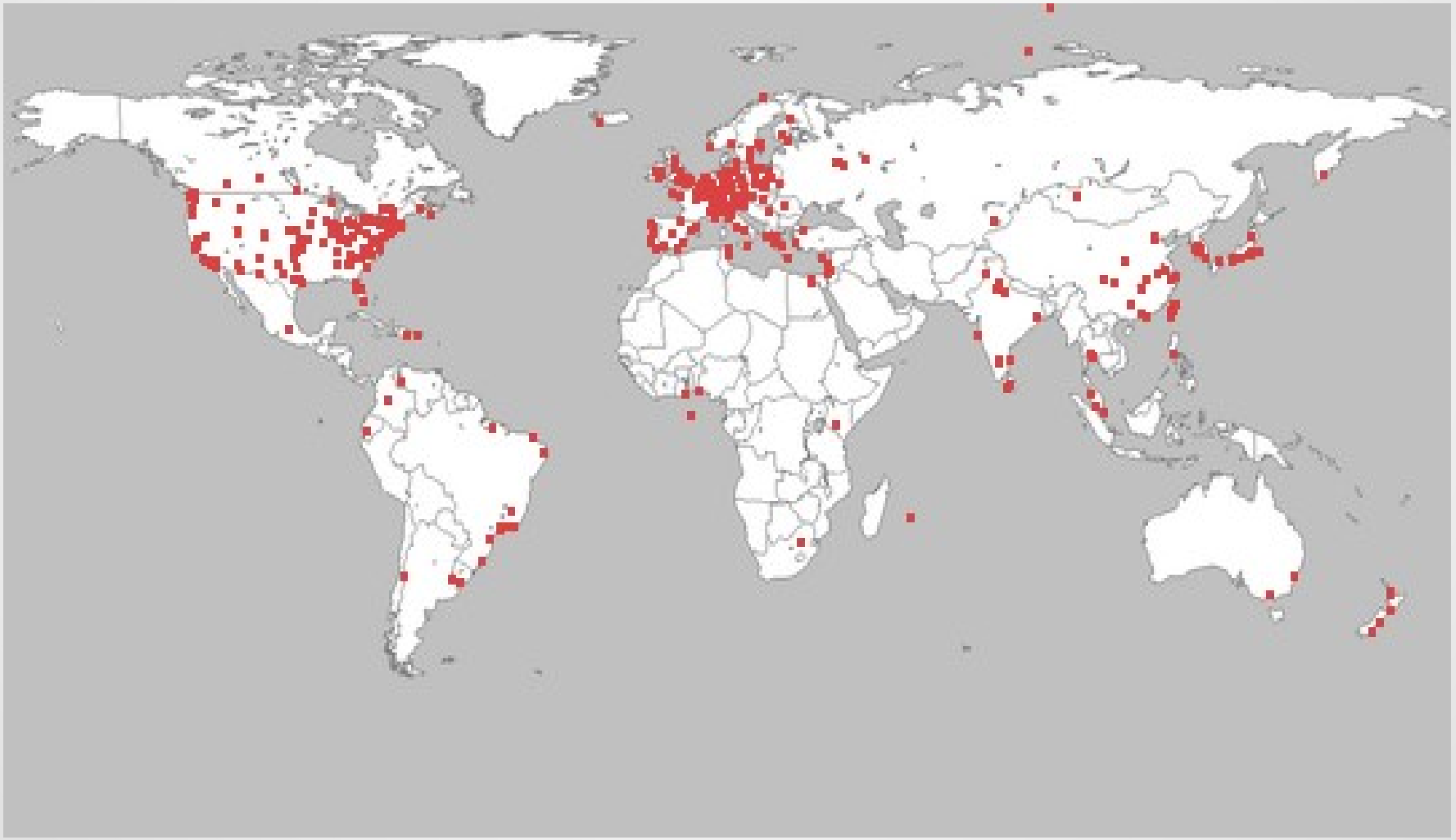
TCP/IP side channels for off-path measurements



TCP/IP side channels for off-path measurements



PlanetLab status map



Analyzing the Great Firewall of China Over Space and Time

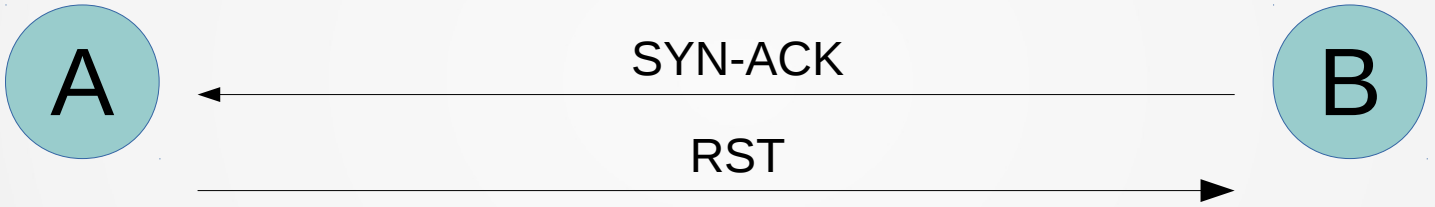
- Roya Ensafi, Philipp Winter, Abdullah Mueen, Jedidiah R. Crandall. *Proceedings on Privacy Enhancing Technologies (PoPETs). 1 (1), 61. DOI: 10.1515/popets-2015-0005. Presented at PETS 2015.*
- Problem statement: Winter and Lindskog (FOCI 2012) observed that roughly 0.5% of Tor relays were accessible from Beijing. Is this true throughout China or just something specific to the one place they measured from?

Hybrid Idle Scan

- A TCP/IP side channel that was first introduced in Ensafi *et al.* (PAM 2013).
- Can tell us if two IP addresses on the Internet can send/receive IP packets to/from each other
  - Assumes one has an open port and the other has a globally incrementing IP identifier (IPID)
  - Also allows us to test layer 4 port information

A TCP/IP handshake

A has no record of ever sending a SYN



A's IPID is incremented by 1 because it sent a RST

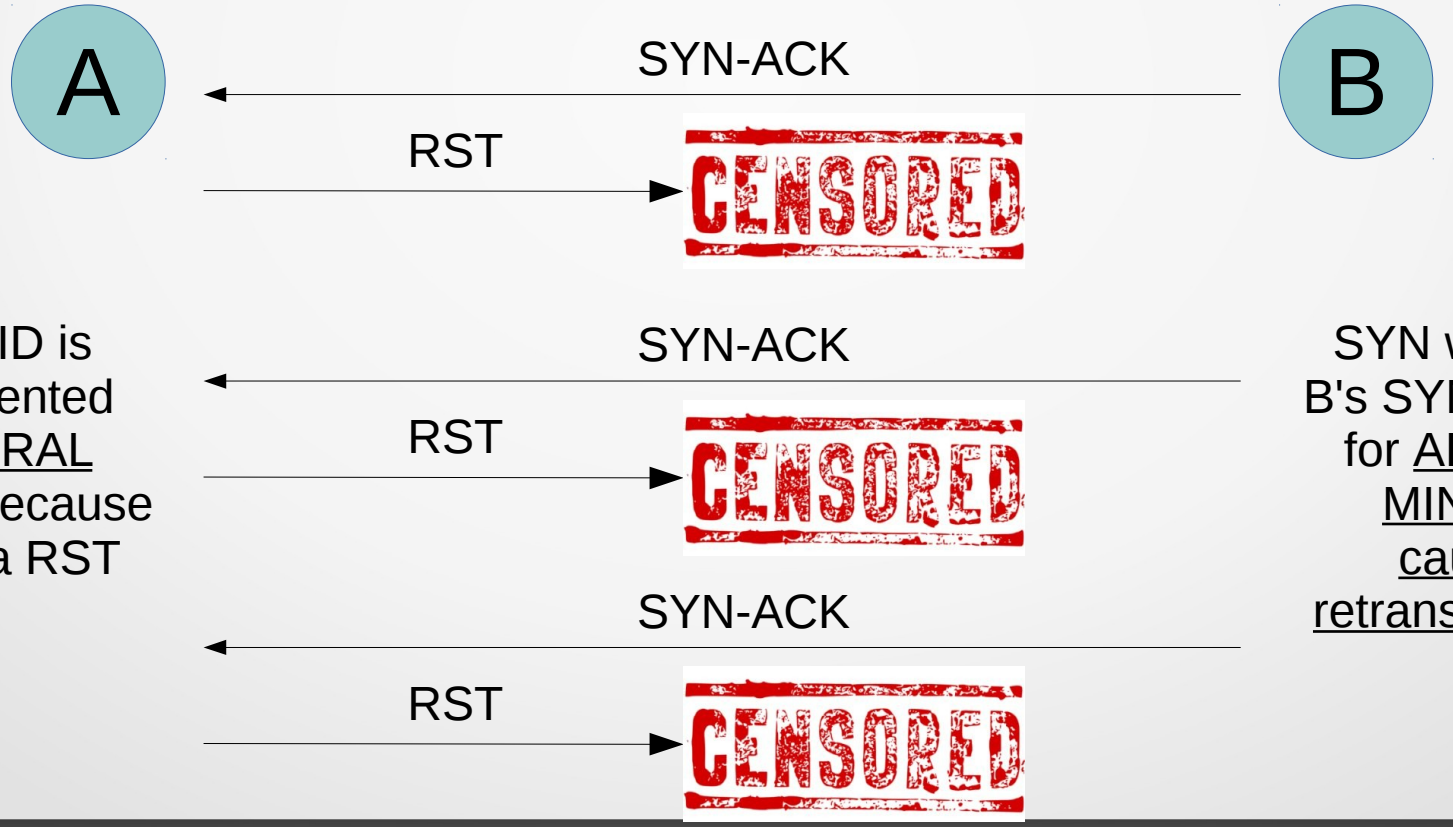
SYN will be in B's SYN backlog for one round-trip time (RTT)

SYN/ACK is dropped due to, *e.g.*, censorship

A's IPID is NOT incremented

SYN will be in B's SYN backlog for ABOUT A MINUTE, causing retransmissions

RSTs are dropped due to, *e.g.*, censorship



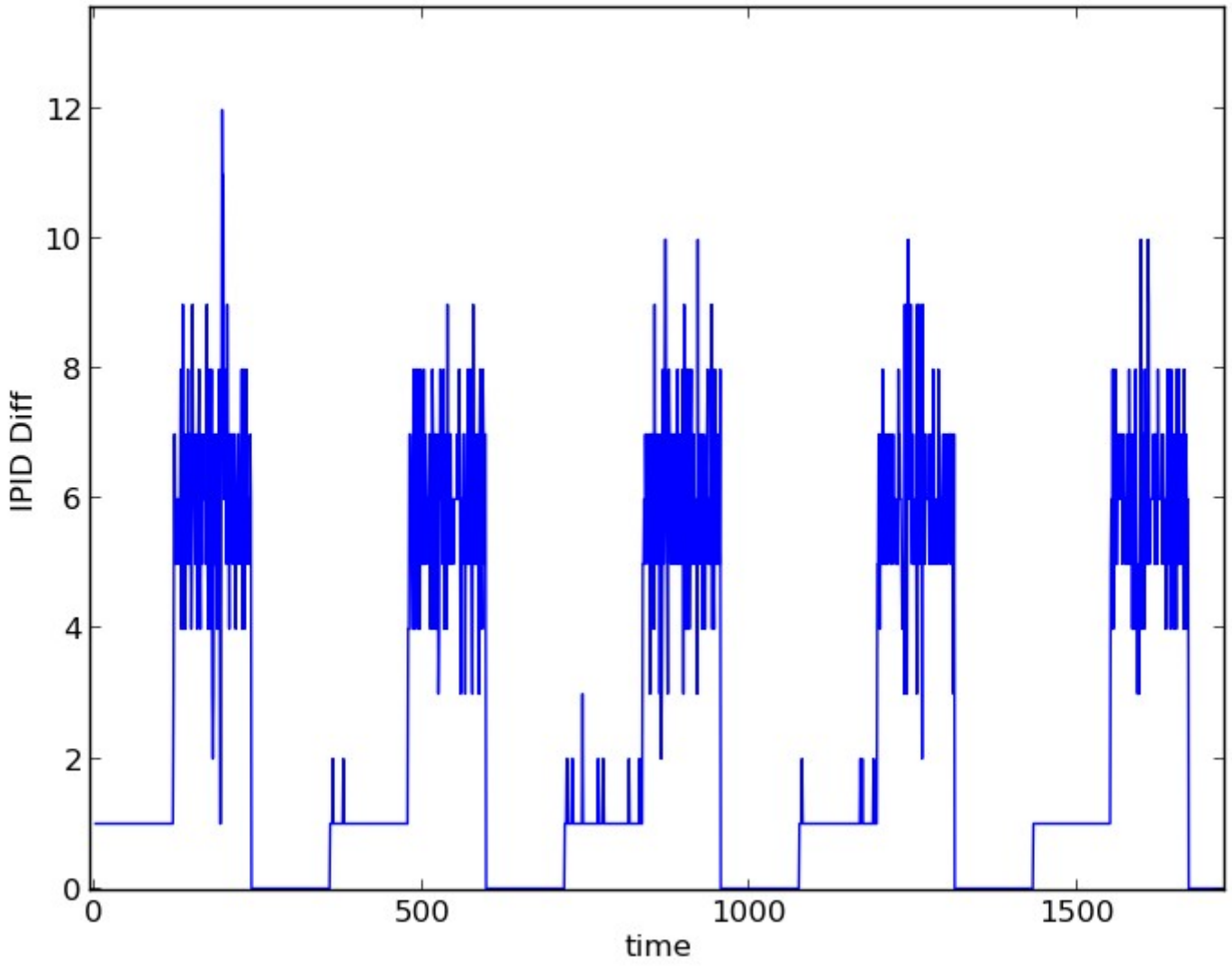
A's IPID is incremented SEVERAL TIMES because it sent a RST

SYN will be in B's SYN backlog for ABOUT A MINUTE, causing retransmissions

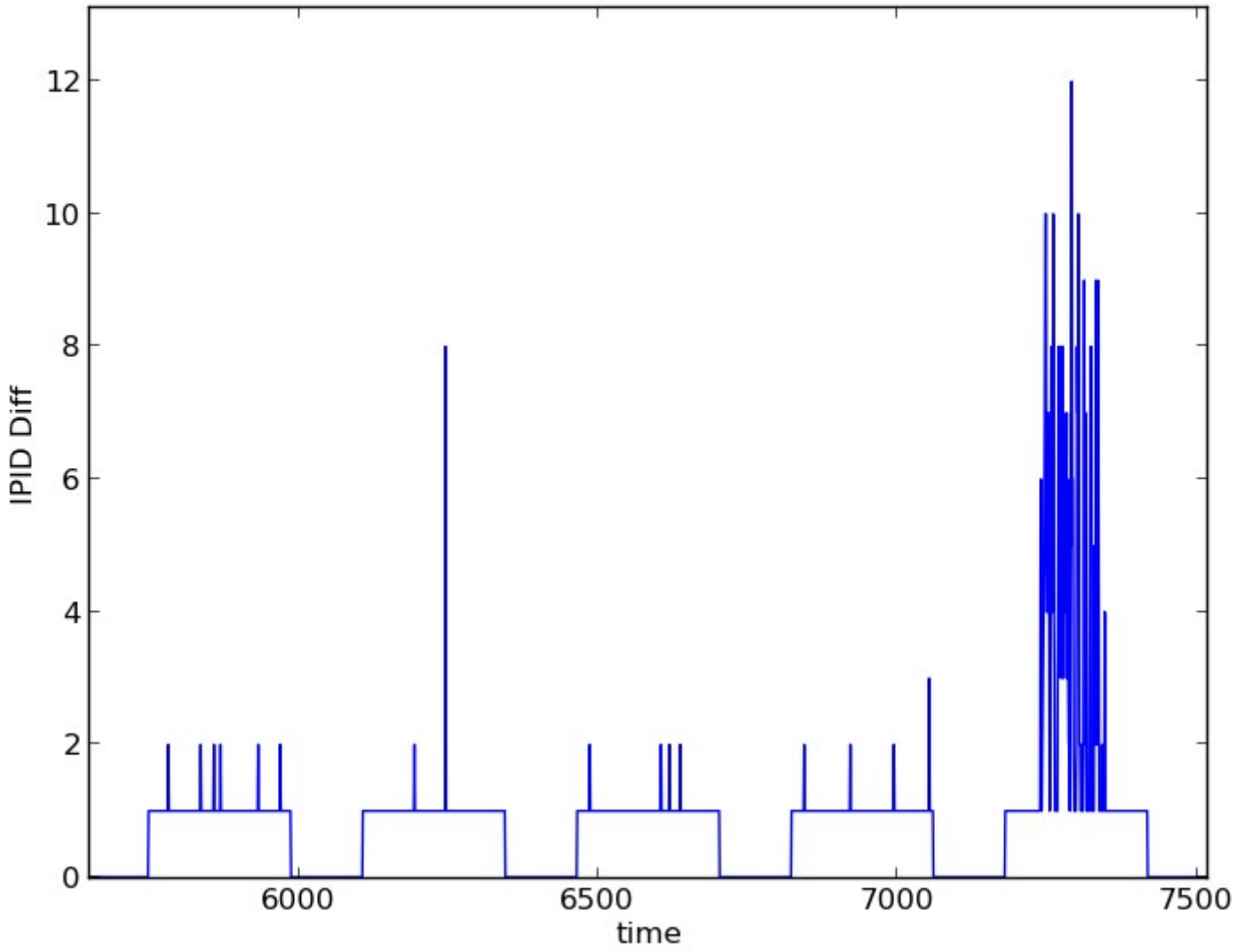
SYN backlog and IPID are shared, limited resources

- Can measure A (*a.k.a.*, the client)'s IPID by sending our own SYN/ACKs and noticing differences
- Could measure B (*a.k.a.*, the server)'s SYN backlog by sending out own SYNs, but the state of the backlog is implied in A's IPID

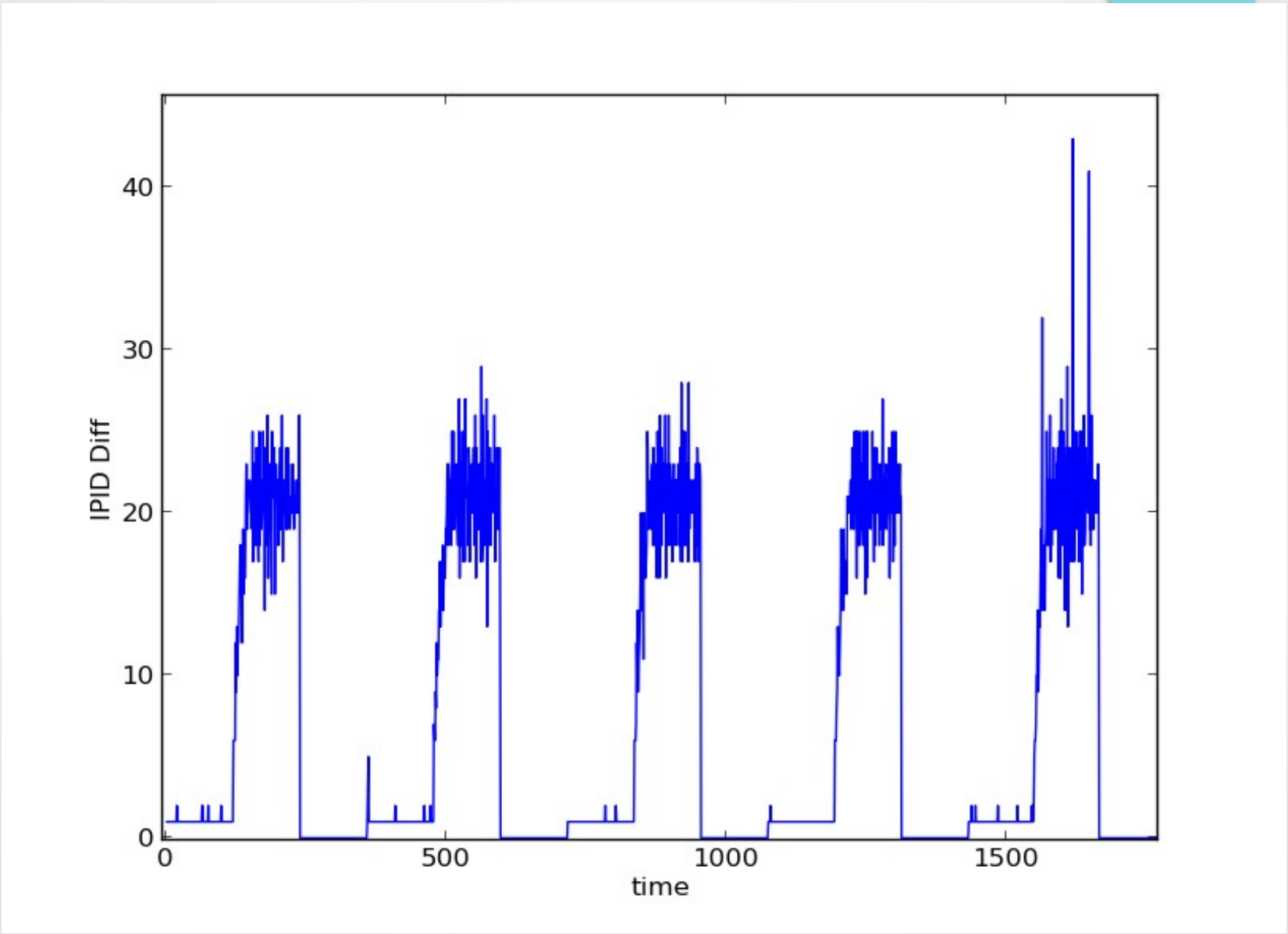
No dropping of packets is occurring



Packets from server to client are dropped



Packets from server to client okay, but from client to server are dropped



Methodology

- Linear combination of regressors fit to an ARMA (autoregressive moving average) model
  - A fourth case (Error) is added
- Details about the challenges our experimental methodology was designed to overcome are in the paper
  - Churn in the Tor network
  - Diurnal patterns
  - Machines that go down
  - Assumptions about ports, machine behaviors
  - Geographically stratified sampling
  - Clients outside China as controls
- Other experiments and results are also in the paper

Data for all client/server pairs

Results

Client	Server	$S \rightarrow C$ (%)	None (%)	$C \rightarrow S$ (%)	Error (%)
CN	Tor–Relay	116,460 (81.52)	555 (0.39)	786 (0.55)	25,061 (17.54)
CN	Tor–Dir	8,922 (64.91)	31 (0.23)	2,696 (19.61)	2,097 (15.25)
CN	Web	306 (1.23)	15,663 (62.95)	2,688 (10.80)	6,226 (25.02)
EU	Tor–Relay	18 (0.20)	8,589 (96.79)	22 (0.25)	245 (2.76)
EU	Tor–Dir	2 (0.25)	776 (96.76)	0 (0.00)	24 (2.99)
EU	Web	19 (1.23)	1,333 (86.28)	95 (6.15)	98 (6.34)
NA	Tor–Relay	45 (0.39)	11,022 (94.48)	33 (0.28)	566 (4.85)
NA	Tor–Dir	4 (0.37)	1,025 (94.73)	3 (0.28)	50 (4.62)
NA	Web	32 (1.52)	1,794 (85.06)	98 (4.65)	185 (8.77)

What we learned

- Winter and Lindskog's (FOCI 2012) observations are true throughout China
  - Routing and active probing likely play a role
  - Heterogeneity of implementation is a possibility

Internet censorship and surveillance – the present

Internet censorship and surveillance – the future

TCP/IP side channels

- IPID (or IPv6 fragment ID)
- SYN backlog
- IP fragment cache
- ARP forwarding table
- More to come...

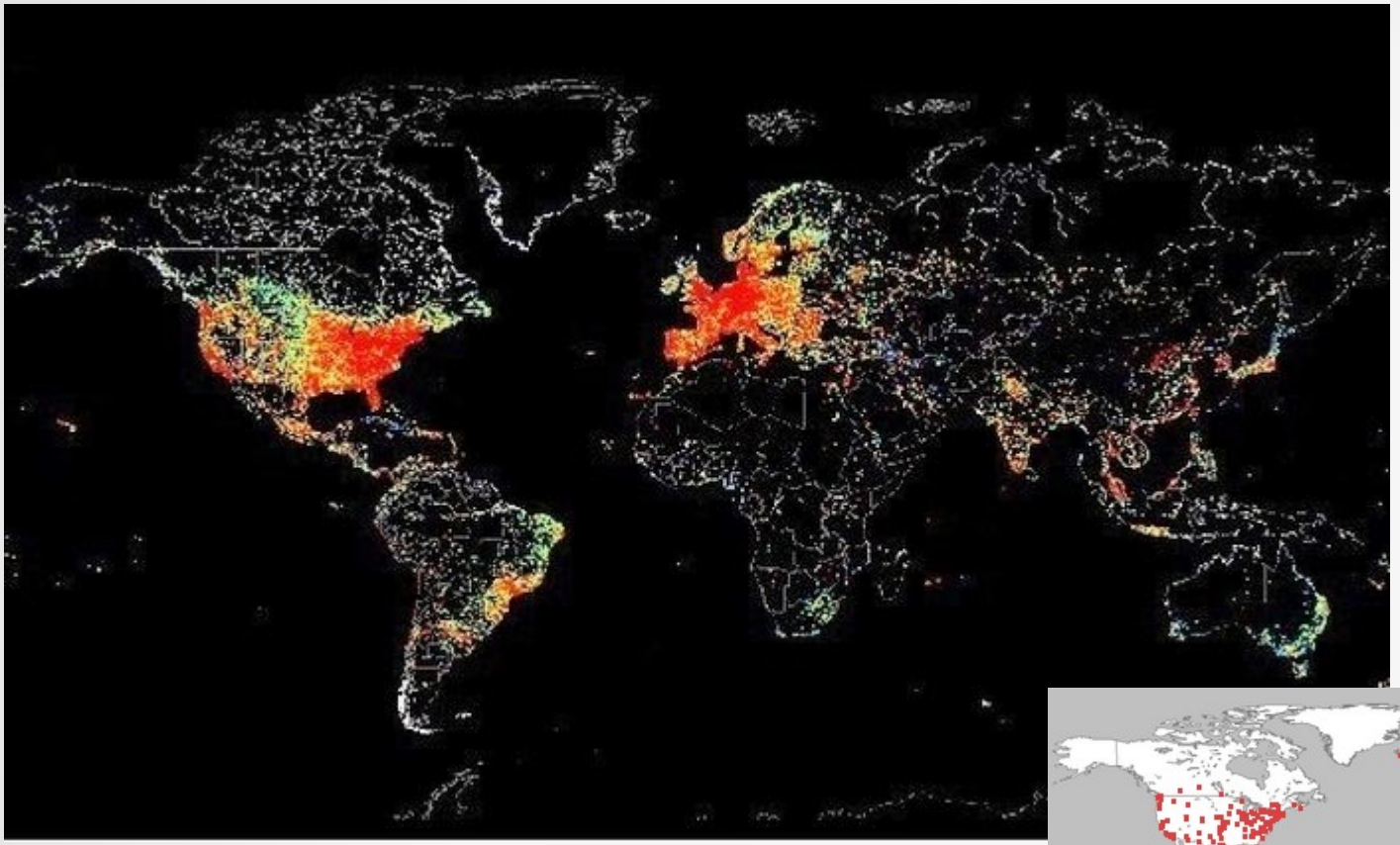
Applications of TCP/IP side channels

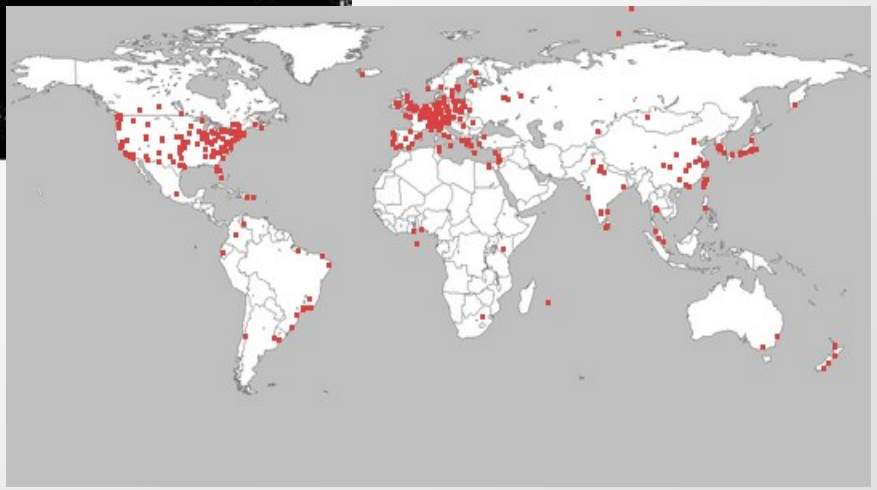
- IP layer (and some layer 4 info like ports) censorship
  - PAM 2013, PETS 2015
- Off-path round-trip time (RTT) measurements
  - INFOCOM 2015, unpublished
- Off-path packet loss measurements
- Off-path Maximum Transmission Unit (MTU) measurements (tunnels, network virtualization)
  - Unpublished
- Counting packets sent between remote hosts
  - FOCI 2014

What about surveillance?

- Better understand market segments, such as chat programs in Asia
  - Surveillance can be built in
  - Crypto is often implemented very poorly
- Routing issues
  - MTU, RTT, packet loss
- Man-in-the-middle attacks
  - RTT

Our goal





How goes the battle?

- Going live in the coming weeks (funded by an NSF NeTS Large)...
  - 10 Gbps *unfiltered* Internet connection
  - 220TB of usable storage (343 TB raw)
- Collaborating with the Citizen Lab at the University of Toronto
  - Ethical issues, including IRB
    - Relying heavily on a vibrant community
  - What to measure?
    - Again, must engage with the community
- All data collected will be freely available

Questions?

- Thank you!
- Thanks to the people who do all the work: Danny Adams, Geoff Alexander, Adnan Bashir, Antonio Espinoza, Jeffrey Knockel, Brandon Lites, Ben Mixon-Baca, Meisam Navaki, Rajkumar Pandi, Jong Park, Erin Sosebee, Xu (Shane) Zhang
- This material is based upon work supported by the National Science Foundation under Grant Nos. 0844880, 0905177, 1017602, 1314297, 1420716, 1518523, 1518878. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.