# Measuring Server-Side Blocking of Tor Users

Mobin Javed
UC Berkeley

In Collaboration with:
Sheharbano Khattak, David Fifield, Sadia Afroz, Srikanth Sundaresan, Vern Paxson,
Steven J. Murdoch, and Damon McCoy

# The Perils of Selective Server-side Blocking

Innocent users suffer due to fate-sharing

- Blocking is generally abuse-based and/or utilizes third-party blacklists.

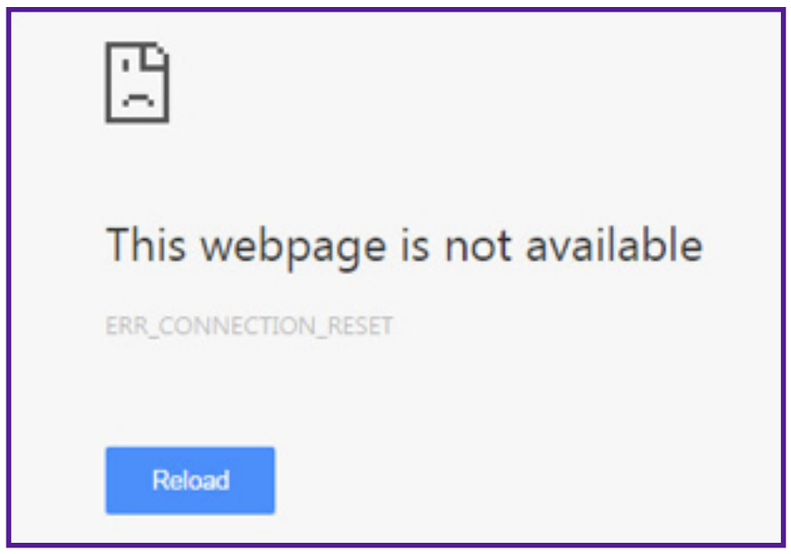
- An important scenario: anonymity networks

    - Sometimes the only rescue tool for users in heavily-censored countries!

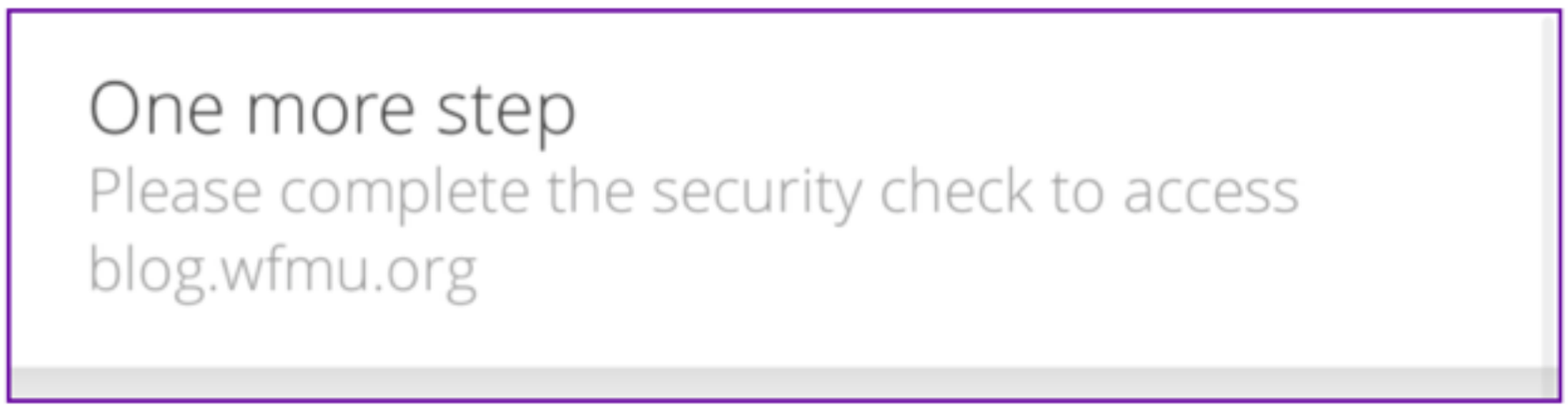
# Example

## Frustrated me using Tor from China …

Network layer blocking




Application layer blocking



# This Talk

- Goal: Quantify server-side blocking of **Tor** at network layer
- Experiment design & tool validation
- Discussion: Modeling web-server “churn”

# Experiment Design

- **High-Level:** scan the entire Internet on port-80 from control nodes and Tor exit nodes.
- Compare results

Tools? 'Tis the era of ZMap!

The Promise: Scan the entire Internet in under 45 minutes!

# Validating ZMap
## Mitigating Measurement Loss

- Does ZMap correctly send/report the packets?
- Measure using experimental set-up
    - 6.7% packet drop at 1Gbps, throttle to 100Mbps
    - Multi-thread configuration buggy, use single-thread

Full Internet Scan takes **7 hours**

# Mitigating Network Loss

- Introduce probe-redundancy
- Temporal churn for back-to-back scans: ~13%
- Need redundancy at shorter-time scales
  - Use a delay of ~7 sec
  - Response rate improves by 1%

Full Internet Scan takes **14 hours**

# Dataset

- Run modified Zmap scans for seven days
    - 4 Tor exit nodes (USA, Netherlands, Romania)
    - 3 Controls (Berkeley, Michigan, Cambridge)
- Scans at different locations synchronized in time
- Success: {SYN-ACK} else Failure

Average Hit Rate: 1.91% (~70 million web-servers)

# Defining Web Footprint

- Web-servers that respond varies across **space** and **time** even for control nodes!

- **Temporal Churn:** up to 17%

  Differing responses over time for the same scan location

- **Spatial Churn:** up to 3.7%

  Differing responses at the same time from different locations

# Defining Web Footprint

- **RAW:** Respond at least once from any location
- **LAX:** Respond at least once from all control nodes
- **STRICT:** Always respond from all control nodes

| RAW: 103 MILLION IP ADDRESSES (aggregated across one week) |
| --- |
| LAX: 96% of RAW |
| STRICT: 50% of RAW |

# Discussion

- Is there an underlying model for web churn?
- Can we characterize various contributing factors?
- What control-plane measurements can we use?

# Questions