Packets Lost in the Wild: An Analysis of Empirical Approaches to Measure Internet Censorship

Mohammad Taha Khan

WCP

March 28, 2017

Committee:

Stephen Checkoway (Chair)

Christopher Kanich

G. Elisabeta Marai

COMPUTER SCIENCE
COLLEGE OF ENGINEERING
UIC

What is censorship?

- The **suppression** of **ideas**, words and images that are **offensive** *(American Civil Liberties Union)*

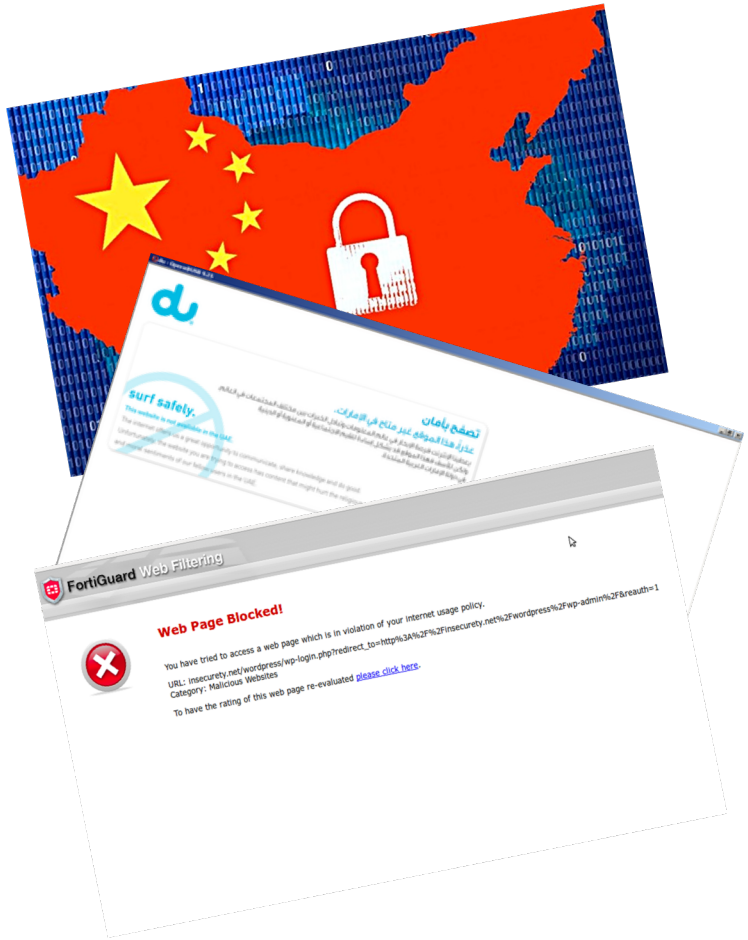
- Carried out authorities, institutions and media outlets

- The **motivations** can be religious, political, moral and even corporate

Censorship: History vs Today

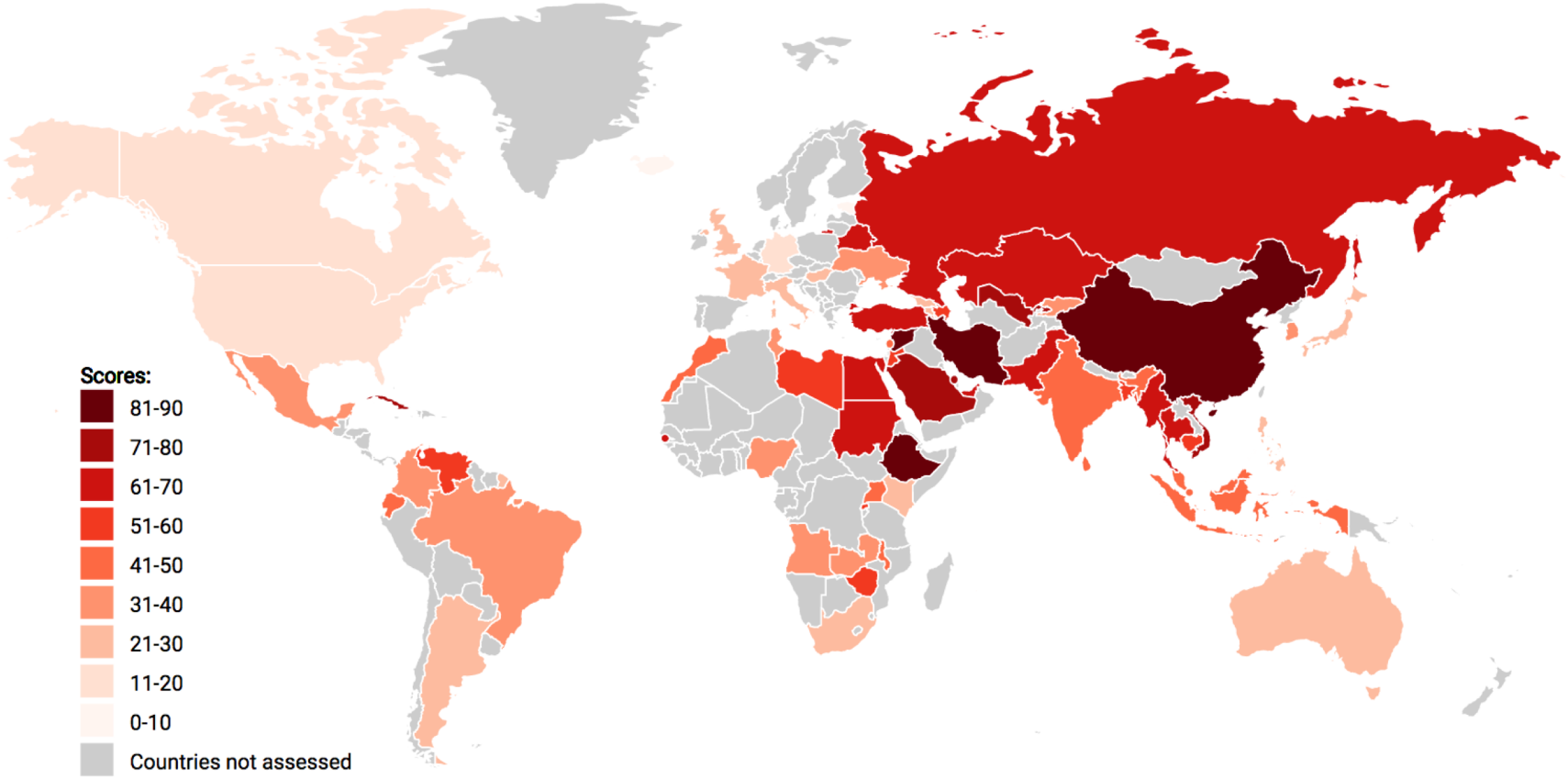
Pre-digital era

The age of the Internet



Global Internet censorship

- More than **66 countries** experience some form of Internet censorship



The Worst Part…

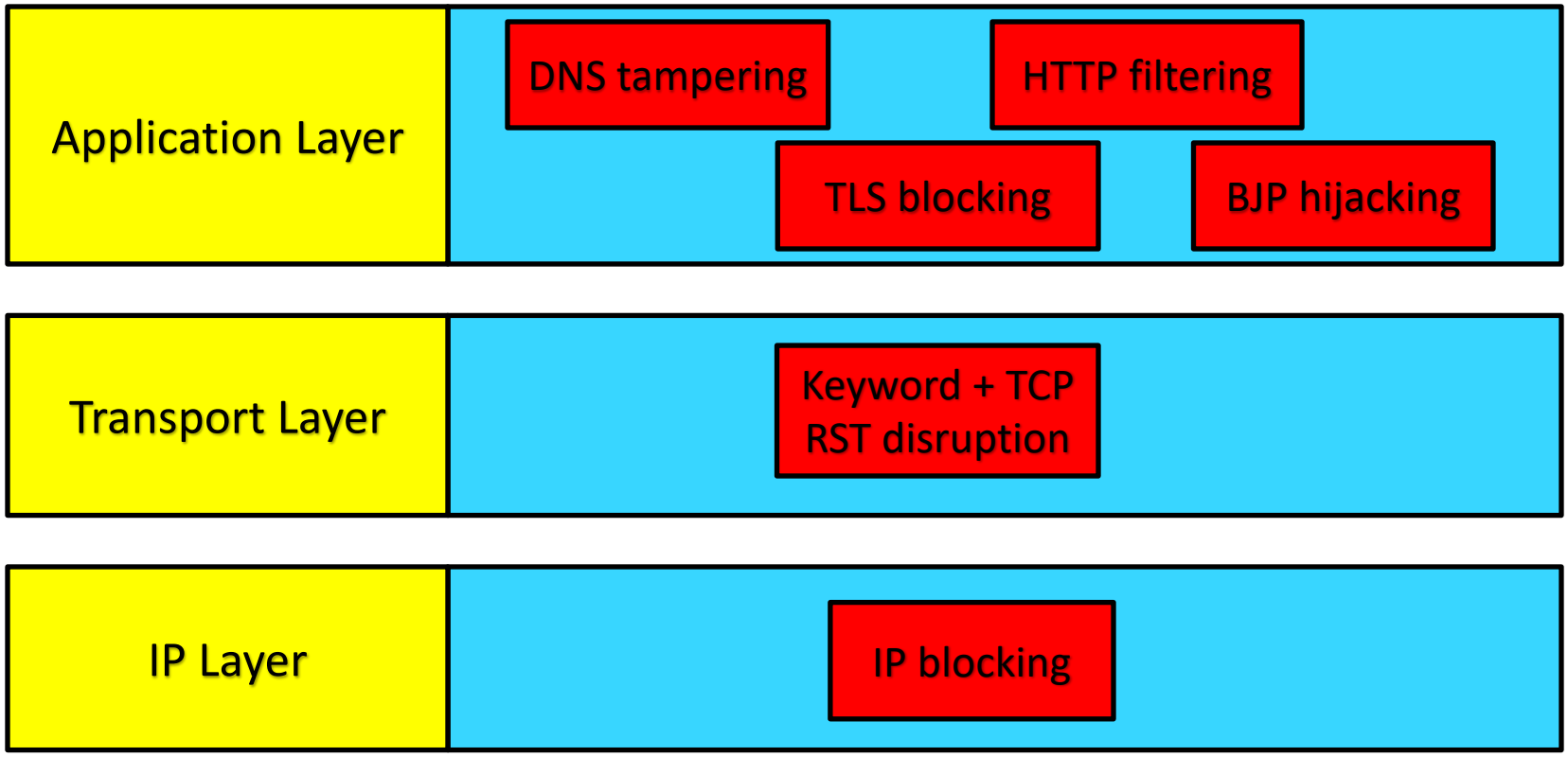
Measuring Internet censorship

- **Who?**
  - Individuals & organizations supporting the idea of the **open Internet**

- **To understand how censorship…**
  - Reduces availability of information
  - Hampers the growth of online communities
  - Impacts activists and civic groups
  - Disrupts economic growth

- To develop **circumvention** mechanisms

Implementation mechanisms



Measurement methodologies

- **Concept Doppler:** Keyword filtering in China
- **URL Filtering Products:** Detection and confirmation of URL filters for censorship
- **Censmon:** Distributed censorship measurements
- **Encore:** Browser based cross origin censorship measurements

1. Concept Doppler

- **Understand** and **quantify** the state of keyword filtering by the Great Firewall of China

- Keyword filtering is **granular**.

- Main **contributions**:
  - The **detection** and **mapping** of **filtering routers**
  - Development of an efficient keyword **extraction** and **probing** technique

Filtering device discovery

- Generate a list of **target servers** in China
  - Google query for domains ending in **.cn**

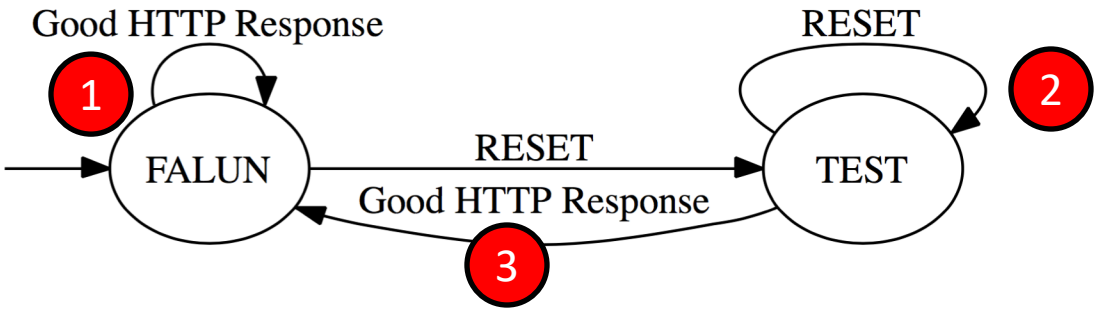
- TTL based ***firewall router*** discovery **algorithm**
  1. Establish a TCP connection
  2. Send a packet containing filtered keyword with increasing TTL values e.g. **TTL =0,1…**
  3. On receiving RST packet, identify location of the firewall routers using the last probe.
  4. Close connection to avoid idling

Discovering blacklist keywords

- Use **Latent semantic analysis (LSA)** to develop a comprehensive blacklist of keywords.

- Use 12 seed keywords to extract correlated keywords from Chinese language Wikipedia

- Probing ***search.yahoo.cn*** to test for plausible keywords:



Evaluation – Concept Doppler

- The firewall discovery algorithms assumes **identical packet routes**
- Current framework leverages **RST packets**
- The testing methodology is **Asymmetric**
- The LSA based technique uses **Gaussian distribution** of textual noise

2. Censorship via URL filters

- Third party URL products have a **dual use** for **censorship**

- ONI has documented several instances over the past 10 years

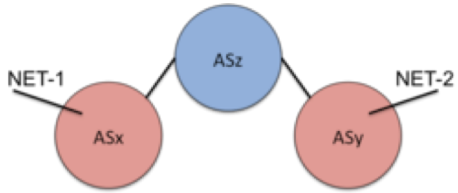
- Main **contributions**:
  - Identifying installations of URL filtering products
  - Confirming their use for Internet censorship

Identifying installations

- Searching the complete web space
  - Use **Shodan** to collect IP information and HTTP header metadata
  - Identify from keywords and country TLDs e.g. *proxysg, macafee, blockpage.cgi*

SHODAN

- Validating the installations:
  - WhatWeb proxy
  - IP to AS Mappings



Confirming use for censorship

- In-network testing
  - Measurement clients are setup in **suspected ASes**
  - **Control experiments** confirm the state of **blocking**
- Domain submission testing
  - Create domains containing **potentially objectionable content**
  - Submitted to **enterprises**

Evaluation – URL Filtering Detection

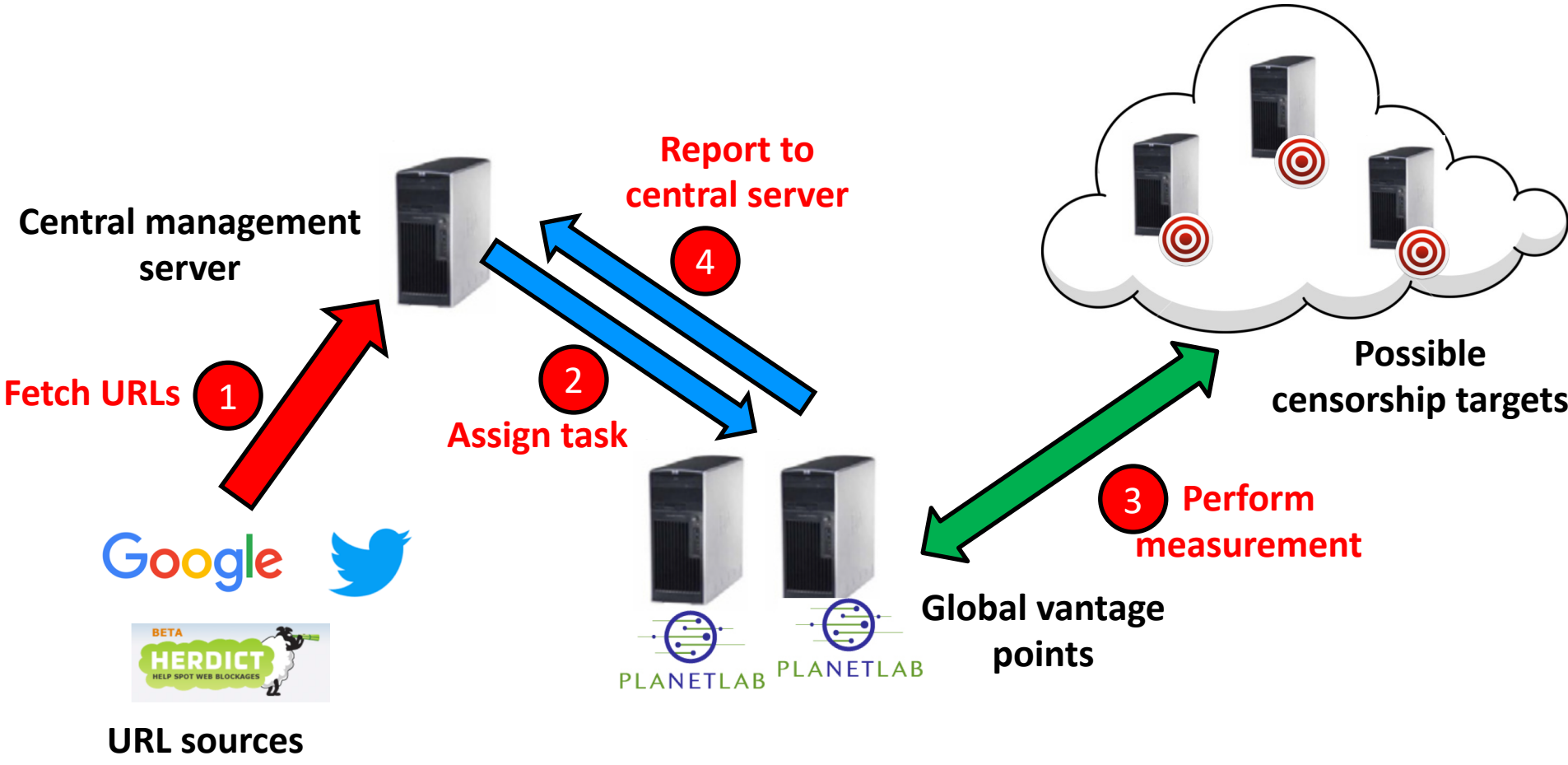
- Discovers installations that are **only globally visible**
- Scanning becomes **harder** with **newer technologies** like IPv6
- **Scalability** is an issue for **large scale measurements**
- Device vendors economically benefit and can collude make **installations undetectable**

3. Censmon

- **Censmon…**
  - Is a based on a client server model.
  - Collects automatic measurements

- **Salient features** of the design
  - Planet Lab Nodes
  - Multiple plugin feeds
  - Identify the filtering technique used

Censmon Design



Censmon Approach

Tasks at the node

1. Make a **DNS request** for domain resolution
2. Establish a **TCP connection** on port 80
3. Make request to **dummy server** and **target URL**

Tasks at the server

1. Repeat experiments due to **network failures**
2. **Match Whois** records with the DNS responses
3. Hashes HTML responses for **partial filtering**

Evaluation - Censmon

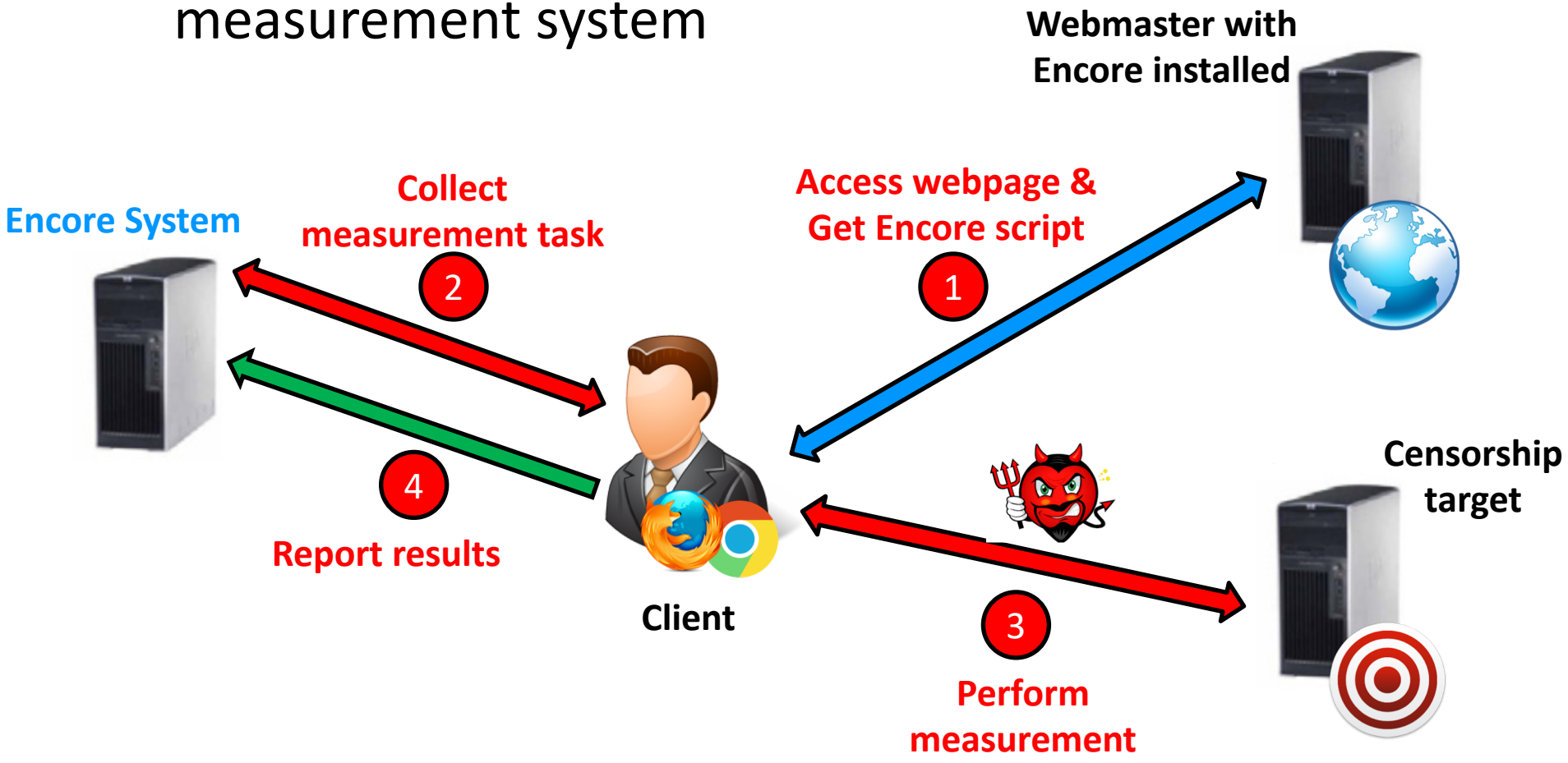
- Tested in 2500 domains, 193 **censored**.
- HTTP filtering (**48.5%**) DNS (**18.2%**) IP (**33.3%**)
- Planet lab nodes are **limited vantage points**
- **No validation** for HTML response filtering detected

4. Encore

- **Vantage points** are essential in measuring censorship
- Current systems achieve this by
  - Measurement clients
  - VPN services
  - Local crowdsourcing.
- These approaches incur an **overhead**
- **Encore** harnesses cross origin requests in browsers
- Current browsers allow cross origin requests for **images, stylesheets, iframes and scripts**

The design of Encore

- Encore requires webmasters to install the measurement system



Efficient measurement characteristics

- **Developing** intelligent measurement tasks
  - URL expansion
  - HTTP archives
  - Task selection
- Detection of domain vs URL filtering
  - **Complete domains:** multiple resources **blocked**
  - **Specific URLs:** iframe (timing information) & scripts
- **Control setup** to validate the measurements

Evaluation - Encore

- **Deployment**
  - 17 Webmasters
  - 141k measurements from 88K IP addresses
- Approach raises **ethical concerns**
- Encore **depends** on **reliable webmasters**
- Integration with more **secure browsers**

No silver bullet!!!

Takeaways…

- **Moving forward…**
  - Minimize user involvement
  - Globally diverse and safe vantage points
  - Collaboration of technologists and social scientists

- Create **a global repository** of measurement results

- Development of a product that provides **circumvention** and performs in a **measurement** decoupled manner

Conclusion

- Researchers have come up with various ways to measure censorship
- Measuring censorship is a **non-trivial** and **dynamic** problem.
- Active area of **research** and **development**

System	Region	Blocking Detection	Methodology
Concept Doppler	China	Keyword Filtering	External probing / LSA
URL Filtering	MENA	URL Filtering	External scans/ In-network testing
Censmon	Global	DNS, IP, URL Filtering	Overlay network (PlanetLab)
Encore	Global	Domain, URL Filtering	Cross origin browser requests

Thank You!

Questions?

System	Region	Blocking Detection	Methodology
Concept Doppler	China	Keyword Filtering	External probing / LSA
URL Filtering	MENA	URL Filtering	External scans/ In-network testing
Censmon	Global	DNS, IP, URL Filtering	Overlay network (PlanetLab)
Encore	Global	Domain, URL Filtering	Cross origin browser requests

Latent Semantic Analysis