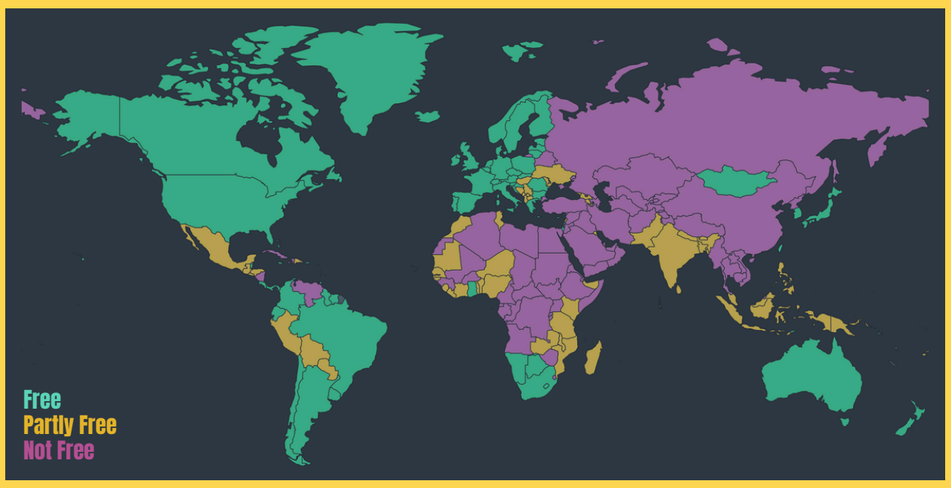
I Know What You Cloned Last Summer

Internet Censorship is Widespread



- A **worldwide** issue
- Many methods to circumvent censorship have been created
- Blockers go to unbelievable lengths to **block** censorship-evasion tools

GitHub & Censorship Circumvention

- Many censorship circumvention tools are hosted on GitHub, and downloads are protected via **TLS**
- GitHub is a **prime target** for censors

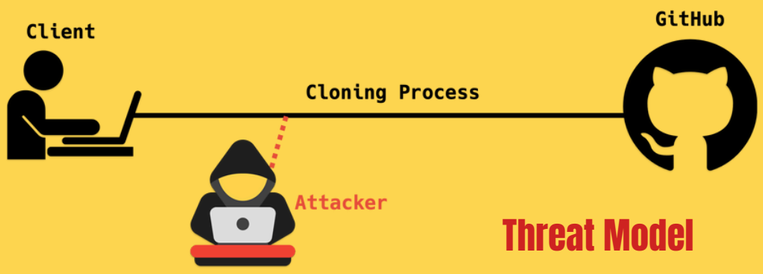
The Case of China v. GitHub

- GitHub was blocked in China for **2 days** back in 2013
- There was a **huge backlash** from Chinese developers
- It was unblocked after 2 days!
- It has **remained unblocked** ever since

Even though most censors are not blocking GitHub, they might have sophisticated ways of detecting which repositories are cloned

Do you think your cloned repositories are a well-kept secret?

GitHub Repository Fingerprinting



Threat Model

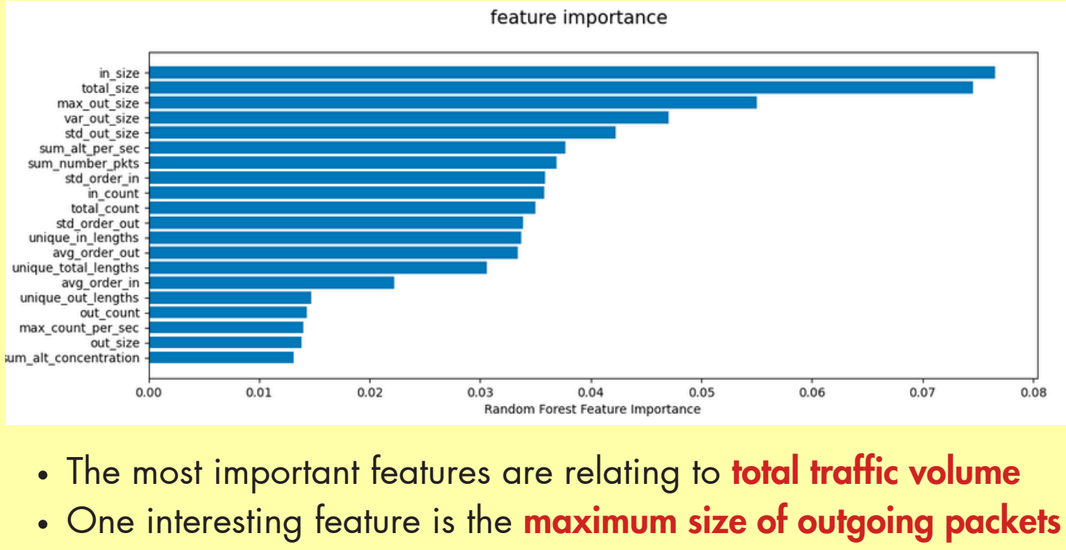
Our Idea:

- Gather a **large** and **representative** dataset of GitHub traffic
- Measure the feasibility of successfully **fingerprinting** the cloning of GitHub repositories

Preliminary Results

Dataset Name	Random Forest Cross Validation	XGBoost Cross Validation	Random Forest Top-2 Accuracy	XGBoost Top-2 Accuracy
25 of each	**96%**	95.00%	**99.5%**	99.00%
Top 100	**97%**	96.00%	**98.97%**	**98.97%**
PET tools	86.00%	**90%**	**95.95%**	94.94%

- Closed world results indicate a **successful** attack
- Top-2 accuracies indicate that the model can **accurately** classify repositories down to **two possible options**



- The most important features are relating to **total traffic volume**
- One interesting feature is the **maximum size of outgoing packets**
- Features correlating to **packet ordering** and **counts** are relevant as well

Methodology

Experimental Testbed

- All data collection was done using a **laptop** to mimic **real world scenarios**
- The laptop was connected to **wifi**
- The cloning was done in **cycles** 24/7 to avoid bias

Datasets

top 100

The Top 100 most starred repositories

25 of each

25 top starred
25 top trending
25 from top collections
25 from top topics

Pet Tools

A handcrafted dataset of prominent PET tools available on GitHub

Pre-processing Approach and Feature Extraction

- .pcap files are converted to traces, which are later converted to sets of **features**
- Features are **manually** created and mostly composed of **summary statistics** of the traces, such as: *total traffic volume, unique lengths*

Classification Setup

- Using **Random Forest** and **XGBoost**
- The tests are done in a **closed world** scenario
- Evaluation is done through **Cross Validation** and **Top-2 Accuracy**

Future Work

- Build a **representative** open world dataset
- Gather the **open world** test results
- Adapt **deep learning** approaches to this setting

UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Sina Kamali, Diogo Barradas
<sinakamali, diogo.barradas> @uwaterloo.ca