Crawling Tor Hidden Service with Dockers

Youngseok Lee
Chungnam National University
University of California, Davis
Tor and Hidden Service

- Tor (The onion router) [https://www.torproject.org/](https://www.torproject.org/)
  - Special network that supports **anonymous** communication using onion routing
  - One kind of Deep web ← → Surface web
  - Called a Dark net
Hidden Service

- Hidden Service
  - Web site or server that only receives inbound connections through Tor
  - Use a special address called "onion address" consisting of 16 alphanumeric characters instead of IP address or common domain
- Many markets where various illegal goods are traded, and communities that deal with dangerous contents

http://3g2upl4pq6kufc4m.onion/ - DuckDuckGo Search Engine
Crawling Tor Hidden Service

- **Problem**
  - Crawl Tor hidden services quickly

- **Constraints**
  - Due to encryption and peer-to-peer connection of Tor, the access to the Tor hidden service is slow
  - Hidden services are often opened and closed frequently

- **Challenges**
  - Slow to collect and analyze dynamically changing Tor hidden services
  - How much computing resources can we put?
Related Work

- Discover hidden service at protocol level [1]
- Observation of Tor service through port scan [2]
- Observed onion requests on global public DNS A and J root nodes and investigate the dynamic nature of the request [3]

Proposal: Virtualized Crawler on the Cloud

- On the cloud
  - Amazon AWS or MS Azure
  - Distributed data centers on the globe
  - Flexible to manage network resources

- Virtualization
  - Docker
  - We can maximize the utilization of computing resources
  - We can probe more Tor hidden services with the same resource
Docker

- Docker is a widely used virtualization software
- Easily deploy virtualized computing resources to the cloud using Docker Image
- Dockers provide containers for abstraction and automation of operating system images
- This allows similar instances of collecting Tor using a large number of instances even for fewer instances.
Docker-based Hidden Service Crawler and Analyzer

- Run one or more docker containers in one instance, such as Microsoft Azure or Amazon EC2.
  - We use a docker to collect and analyze hidden services using fewer instances.
  - Each Docker container, a virtualized Linux server, performs as a Tor crawler.
- Create an image of a dedicated Docker to facilitate configuration and deployment for the crawler.
Crawling Time with Docker

- Experiment on Microsoft Azure’s 2 core 7GB memory instance
- Measures the crawling time according to the number of containers for 100 onion addresses
  - For one container, it takes 3,062 seconds.
  - With five Docker containers, it needs an average of 640 seconds (4.78 times faster than one container)
Hidden Service Page Status

- How Tor hidden services of 9,176 addresses change their states over 9 days.
- During the observation period, the hidden service averaged 1,166 alive during the whole period and 8,010 services were dead.
Content categorization of Hidden Services

- **Marketplace (20%)**
  - illegal transaction services for prohibited goods such as hacking requests, hitman service, cloned credit cards, and drugs take 16%

- **Community (14%)**
  - there are communities about the illegal themes such as drug dealers and pedophilia

- **Bitcoin laundry services (6%)**
  - where users may send Bitcoin with the promotion to receive the returns of 10x to 100x coins.

<table>
<thead>
<tr>
<th>Contents Category</th>
<th>Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marketplace</td>
<td>20</td>
</tr>
<tr>
<td>Community</td>
<td>14</td>
</tr>
<tr>
<td>Hidden Service directory</td>
<td>12</td>
</tr>
<tr>
<td>Personal Blog</td>
<td>9</td>
</tr>
<tr>
<td>Onion address sales page</td>
<td>9</td>
</tr>
<tr>
<td>Unknown page</td>
<td>9</td>
</tr>
<tr>
<td>Bitcoin laundering</td>
<td>6</td>
</tr>
<tr>
<td>Wiki</td>
<td>5</td>
</tr>
<tr>
<td>Apache default setting page</td>
<td>5</td>
</tr>
<tr>
<td>Journalist, Movement organization</td>
<td>4</td>
</tr>
<tr>
<td>Software distribution, sales</td>
<td>4</td>
</tr>
<tr>
<td>Illegal file sharing</td>
<td>3</td>
</tr>
</tbody>
</table>
Summary

● Crawling of Tor hidden services
  ○ Improving performance with virtualization and cloud

● Ongoing work
  ○ Automation of virtualizing and controlling crawling instances on the cloud
  ○ Detail analysis of Tor hidden service content