Crowdtrace: Tracerouting and measuring the QoE from the crowd

Ricky K. P. Mok, Amogh Dhamdhere, kc claffy
CAIDA

AIMS 2017
Network measurement platforms

• Outage
• Congestion
• Internet route
• ...
Crowdsourcing

• Thousands of (human) workers work for
  • $$$ money
  • Fun
• Types of tasks
  • Survey
  • Marketing
  • Image/video tagging
  • Network measurement
  • ...
  • Assignment
A missing link
Concern

• The time of outage/congestion event vs. the time when workers conduct the measurement

• Network hosting measurement VP vs access network of worker

• CDN Cache
  • The worker can be served by another GGC, rather than the one found anomalies by the monitors.
Limitations

• No software/code download/installation

CrowdFlower\(^1\): You may not without a separate written agreement with CrowdFlower include tasks that violate our policies, including, but not limited to, ...
(f) tasks that require Contributors to download software or files.

Amazon Mechanical Turk\(^2\): What are some specific examples of HITs that violate Amazon Mechanical Turk policies? ... require Workers to download software that contains any malware, spyware, viruses, or other harmful code

• No binary
• No browser plugins
• Same-origin policy

\(^1\)https://www.crowdflower.com/legal/customer-terms-conditions/
\(^2\)https://www.mturk.com/mturk/help?helpPage=policies
A preliminary study

• Where do the workers come from? Can we easily seek a more specific set of workers?
  • Geographical location
  • Network location

• What is their “office hours”?
  • Can they possibly measure the peak hour congestion?

• which Google Global Cache would they be directed to for youtube videos?

• Can we do traceroute from their computer?
Micro task flow

- Crowdsourcing Dashboard
- Introduction and instructions
- Save and upload a YouTube page
  - Execute traceroute
  - Metadata (+ QoE assessment)
  - Running traceroute (at bg)
  - Report traceroute

- Survey code

- IP, User agent
- GGC serving the worker
- Congested links
- Location, ISP, and QoE
- traceroute
Location and ISP

• Crowdsourcing platforms only support country level filtering.

• We specify the ISP and the city we interested in the title of our task.
  • Location of monitors

5-Minute Internet Research Study For Comcast Users In San Francisco, Ca

Overview
You are conducting a research study about Internet topology and performance in the United States. The length of study is about 5 minutes.

Steps
This study consists of two parts:
• You will be asked to download a webpage and submit that page to our platform.
• You will be instructed to execute traceroute and submit its output to our platform, and will be asked about the service quality of six popular video streaming
GGC

• Save a YouTube video page for us
  • right click ➔ Save page as

• Any automatic way?
  • Same-origin policy

• The first GGC responds to video streaming request.
Running traceroute

- Six destinations are probed
  - Inter-domain congestion project from Ark
  - GGC
- QoE

**Part 2**

Session 1: Please execute the following command in your terminal. You can revisit the instruction here.
```
traceret -d -w 500 -h 10 128.125.0.1 upload_output.txt & traceret -d -w 500 -h 13 131.215.0.1 upload_output.txt & traceret -d -w 500 -h 12 80.231.11.1 upload_output.txt & traceret -d -w 500 -h 30 r1 sn 0 jq n5 googlevideo.com upload_output.txt & echo Complete 1 & 2
```

If you would like to opt-in to our upcoming tasks, please enter your MTurk/Contributor ID: [ ]

Your current physical location: [City] [Please select] [ISP]

Please rate the past experience you perceived for following video streaming services in the last week:

- Excellent: ★★★★★ - Poor: ★★★★ - N/A: Did not use that service last week
- *Only consider your experience in this computer with the same Internet connection.*

- YouTube [ ] [ ] [N/A]
- Vimeo [ ] [ ] [N/A]
- Netflix [ ] [ ] [N/A]
- DirecTV [ ] [ ] [N/A]
- Amazon Video [ ] [ ] [N/A]
- Hulu [ ] [ ] [N/A]

Session 2: After the command is done running, upload the output file (upload_output.txt) below and submit the form.

Choose File: No file chosen

Submit
Deployment

• Crowdsourcing platform
  • Amazon Mechanical Turk
  • CrowdFlower

• Advertised ISPs and locations
  • Cox, Comcast, Time Warner cable
  • San Diego, San Francisco, Boston, Georgia, ...

• Credit
  • 7-13 cents
Low completion rate

• Feb 27 to March 2
• We recorded >110 workers accessed our platform
  • Mostly from CrowdFlower
  • MTurk seems losing workers
• Only 7 workers complete the task with traceroute submitted.
• A few workers submitted non-traceroute output.
WHY???!!!
What did they submit?

• Submit non-traceroute output

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
</table>

• Partial traceroute
• Non-English traceroute
  • Traceroute output is “translated” to the system locale.
ISP

• IP Geolocation and ISP: NetAcuity

• Advertised ISP
  • Comcast: 17% (17 out of 100 workers)
  • Time warner cable: 20% (4 out of 15 workers)

• Significant portion of workers’ IP resolved to web hosting companies
  • Logicweb
  • Egihosting
  • Leaseweb
Geolocation

• Advertised location
  • California: 22% (10 out of 45 workers)
  • Georgia: 9.6% (3 out of 31 workers)
  • Washington: 2.6% (1 out of 38 workers)
GGC

• 38 workers uploaded the YouTube page.

• Geographical location
  • Rochester institute of technology
  • Digital Ocean
  • Level 3

• Network
  • Two Comcast users from Georgia and Illinois are assigned to the same GGC.

• Time
  • Same IP address, different cache assignment
Lesson learnt

• The crowd is not reliable
  • More workers are rejected than accepted
• The workers do not follow instructions
  • They will try to save every single click.
• Break one big task into several ones
  • Get partial data
Future works

• Increase the participation
  • Deploy to more crowdsourcing platform
  • Study their behaviour
  • Adjust the wage
  • Add some fun to the task

• Include subjective QoE assessments
  • Measure the impact of network events on the client’s performance and their QoE