CRAWLING TOR'S HIDDEN SERVICES AND DEPICTING THEIR INTERCONNECTIVITY

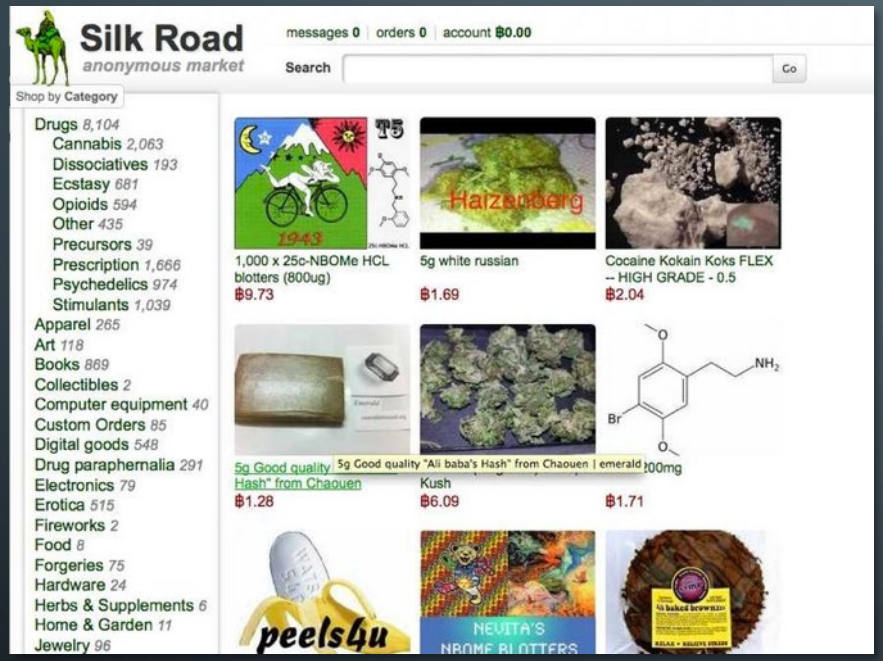
JOHN-LUKE NAVARRO

MENTOR: DR. RODNEY L. SUMMERSCALES

DEPARTMENT OF ENGINEERING AND COMPUTER SCIENCE

Andrews University
J.N. Andrews Honors Program

THE FALL OF (THE) SILK ROAD

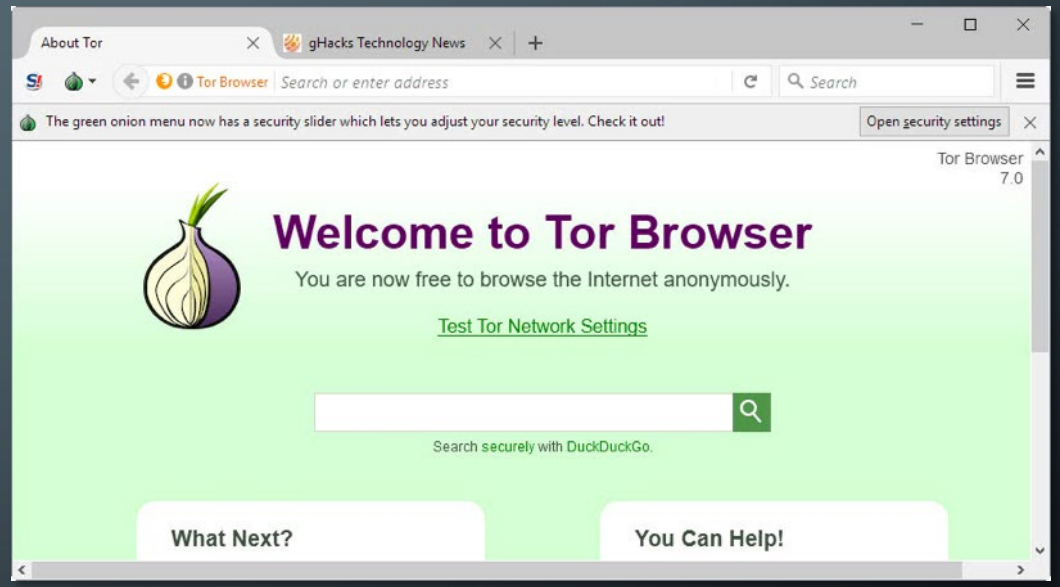


THE TOR NETWORK

- A privacy-centered network
- Anonymizes users and their websites
- Protects users from being tracked
- Wraps information in layers of encryption

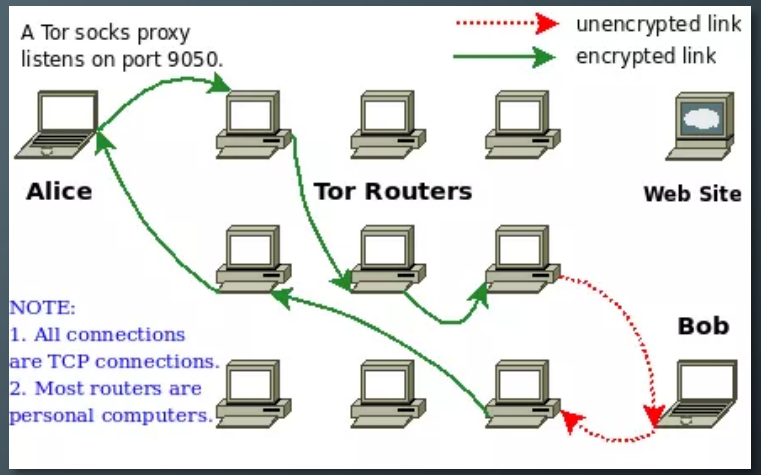
WHAT ARE *HIDDEN SERVICES?*

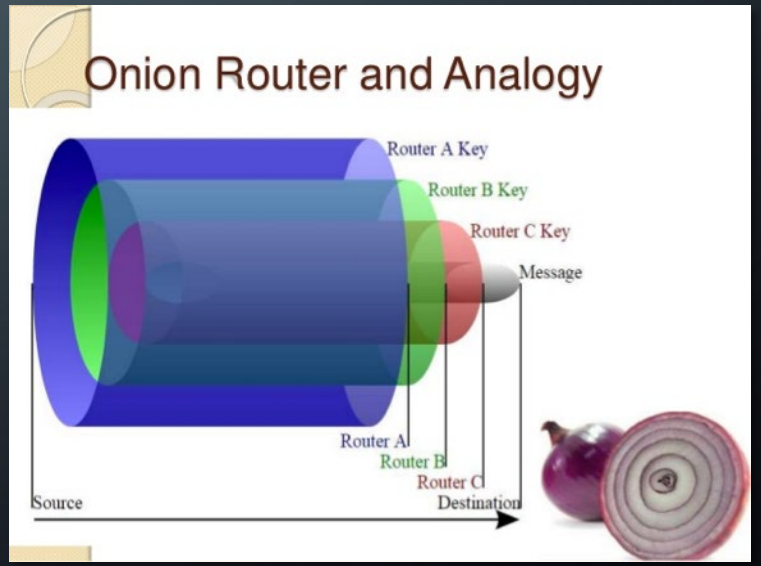
- Websites hosted on Tor
- Same protections as Tor's users
- Identifiable by a *.onion* address
- Example: *silkroad6ownowfk.onion*

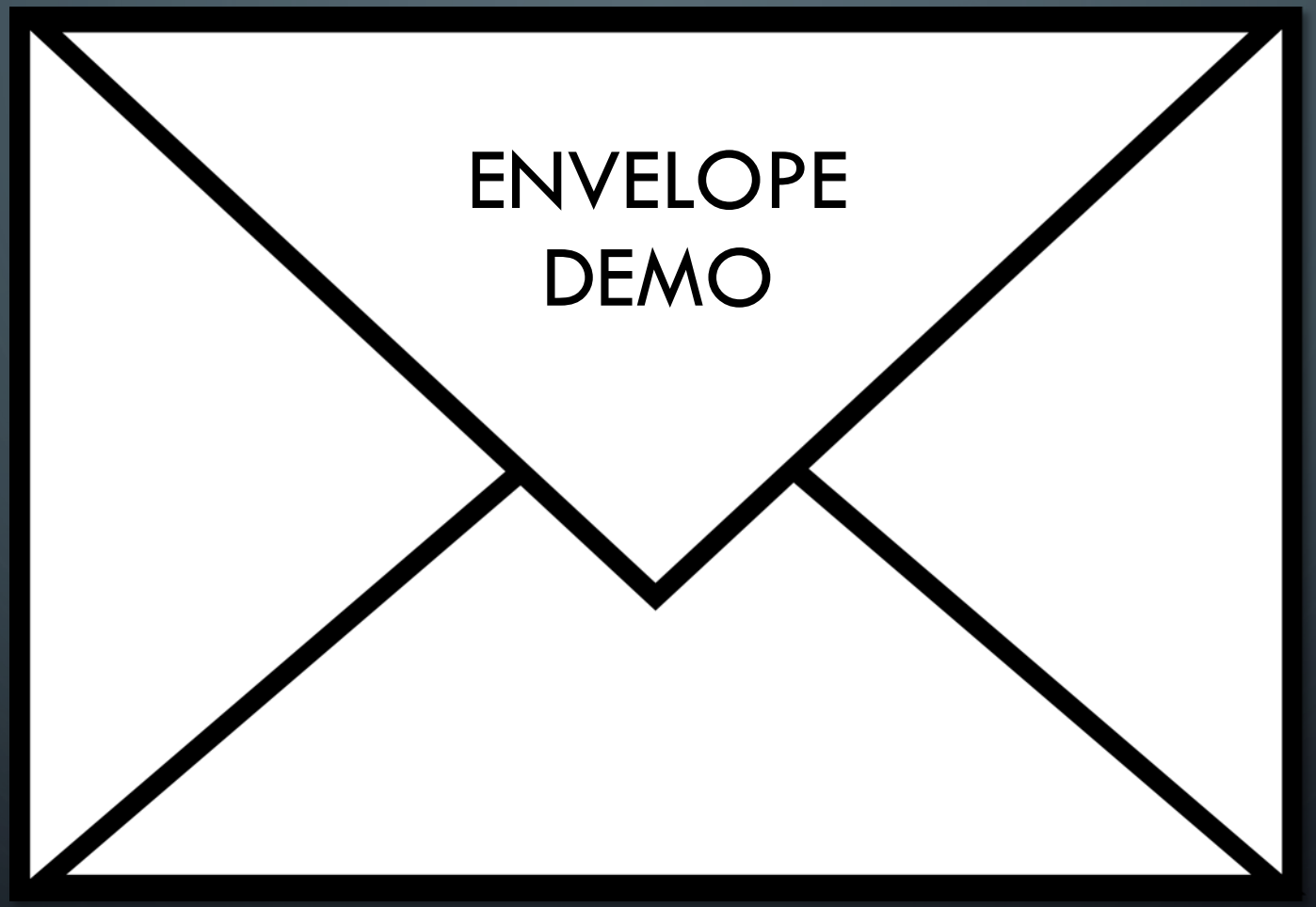


HOW DOES THE TOR NETWORK FUNCTION?

- A *circuit* of servers rotates every 10 minutes.
- Each server in the circuit helps wrap data in layers of encryption
- Each server in the circuit unwraps it's individual layer.



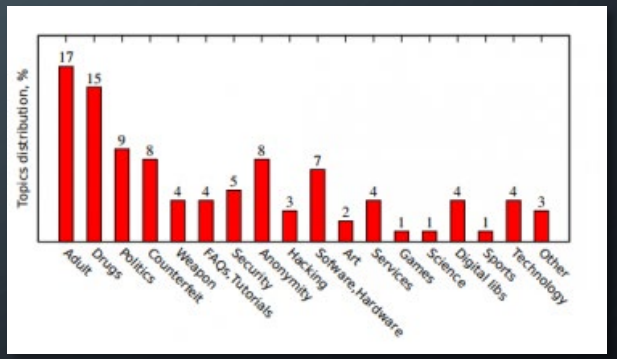



ENVELOPE
DEMO

THE DARKNET WEATHERMAP PROJECT

- SULI Appointment
- Cyber Operations, Analysis, and Research
- Project: Produce daily metrics on Tor content availability, distribution, and “trendiness”
- How can bulk Tor network content be downloaded with relative ease?
- How can the connections between hidden services be depicted?

Argonne NATIONAL LABORATORY

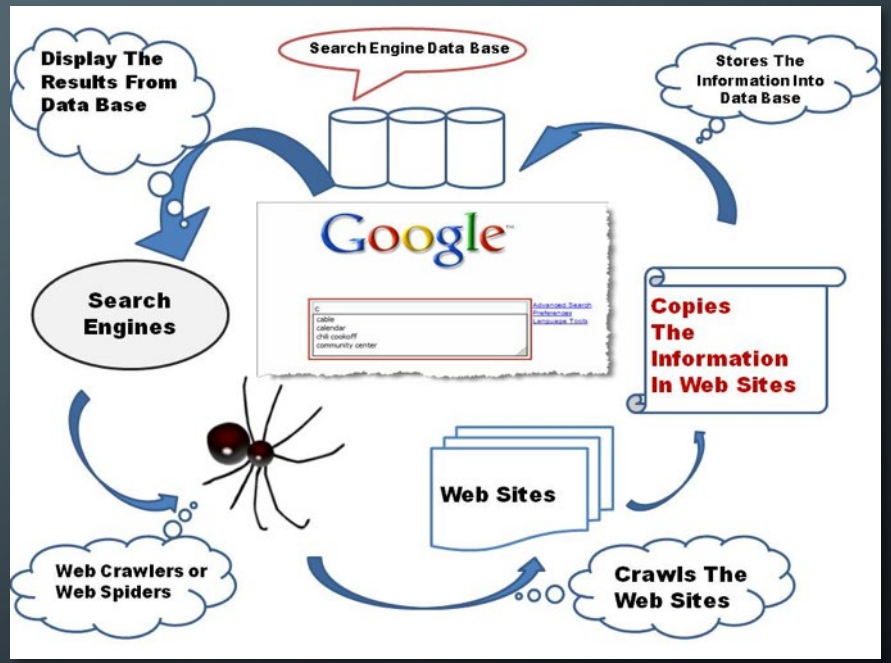


Example from technologyreview.com

WHAT IS A *WEB CRAWLER*?

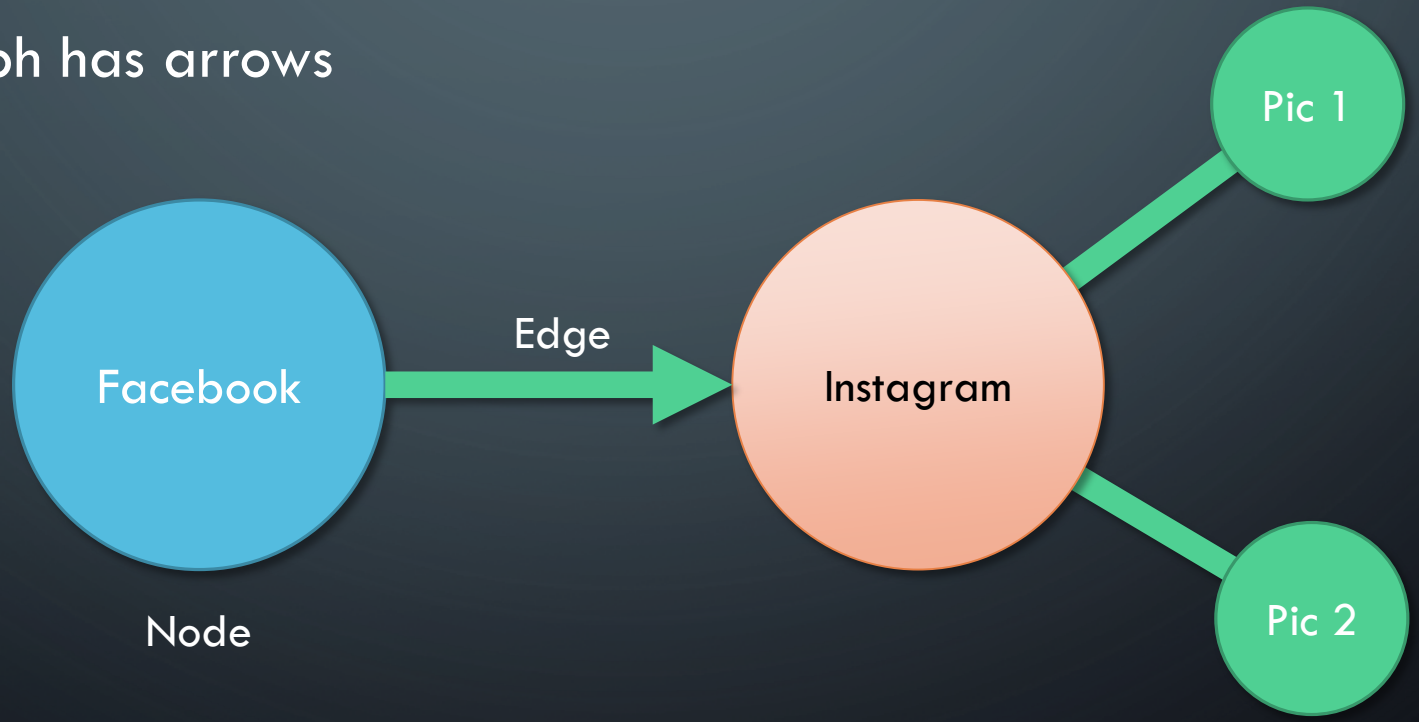
- Automated software that browses the internet and downloads information from websites.

- Used by Google and other Search Engines



WHAT ARE *CONNECTED GRAPHS*?

- Used to show relationships between objects
- Directed graph has arrows

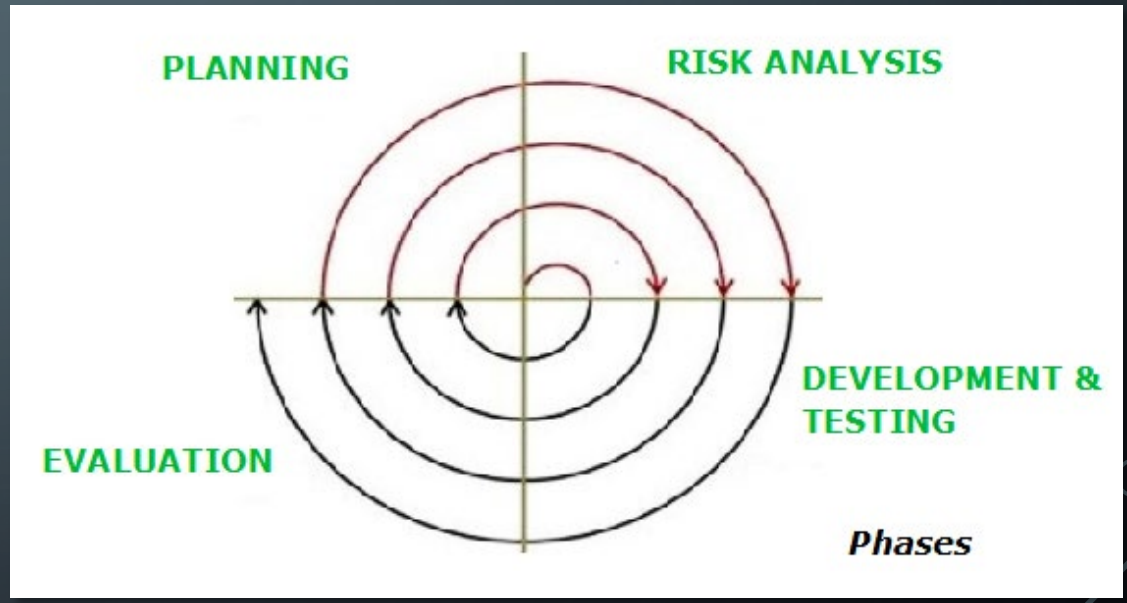


PREVIOUS WORK

- Cryptopolitik and the Darknet – Moore & Rid
  - Developed tools to download and analyze Tor content
- The Tor Dark Net – Gareth & Savage
  - Operated multiple Tor data-mining projects, emphasized manual textual analysis
- Towards a Comprehensive Insight on the Thematic Organization of the Tor Hidden Services – Spitters, et al.
  - Developed another Tor data mining tool, implemented alternative searching strategies.
  - Emphasized extensive textual analysis on resulting data.

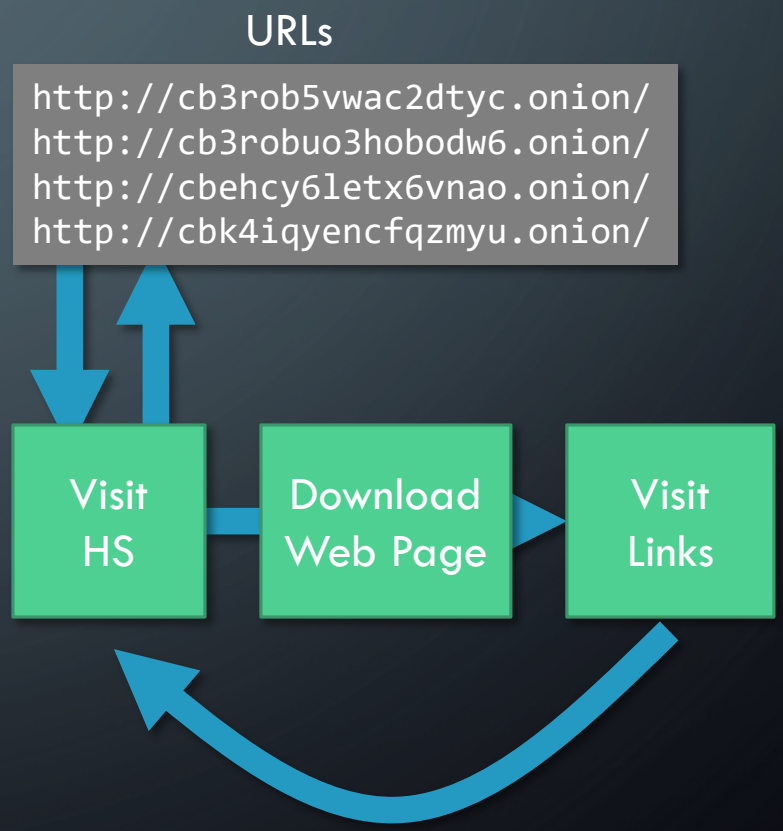
SOFTWARE DEVELOPMENT STRATEGIES

- Adhered to Spiral Model of Software Development
- Frequent code reviews
- Lots of testing, debugging, and documentation
- Frequent meetings with mentors



THE TOR WEB CRAWLER – CRAWLER BASICS

- Written entirely in Python
- Crawler needs an initial list of Hidden Service URLs
- Visit hidden service, download content, click links, repeat



THE TOR WEB CRAWLER – DEALING WITH DATA

- Dataset contains:
  - Downloaded HTML
  - Links to other hidden services
- Avoids downloading too much data
  - Page-count limits
  - Depth limits

Dataset Sample

```
MyHiddenWebPage:
    - homepage.html
    - pictures.html
    - contact.html
    - social.html
    - links.txt
```

URL:	https://www.andrews.edu/services/honors/research/		
Depth:	0	1	2

THE TOR WEB CRAWLER - SECURITY

- Tor integration also protects crawler
- User-agent rotations
  - Pretend to be a human!
  - Avoid software that blocks bots
- Keyword blacklists
  - Prevent downloading unwanted content
  - Avoid undesirable websites in the future

AVOID:
- Classify
- Classified
- Military Secret
- Redacted

THE CONNECTIVITY GRAPH PROGRAM

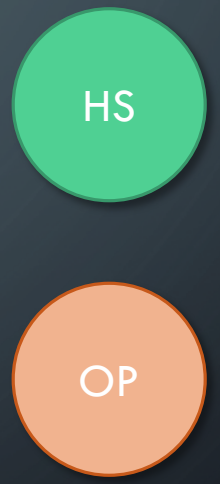
- Written in Python with the Networkx library
- Takes crawler datasets as input
- For each crawled hidden service:
  - Insert the URL as a node
  - Check the links file, insert edges between two matching nodes

Folders

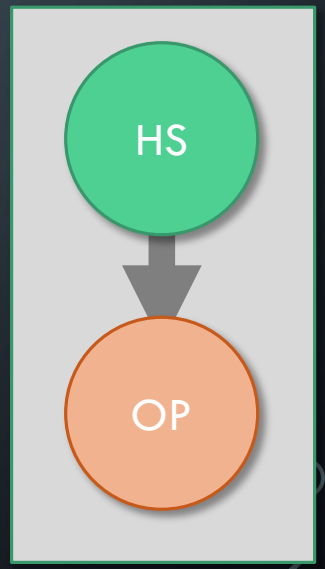
hiddensite
- Links
  - helloworld.onion
  - otherplace.onion

otherplace
- Links
  - buystuff.onion

Nodes

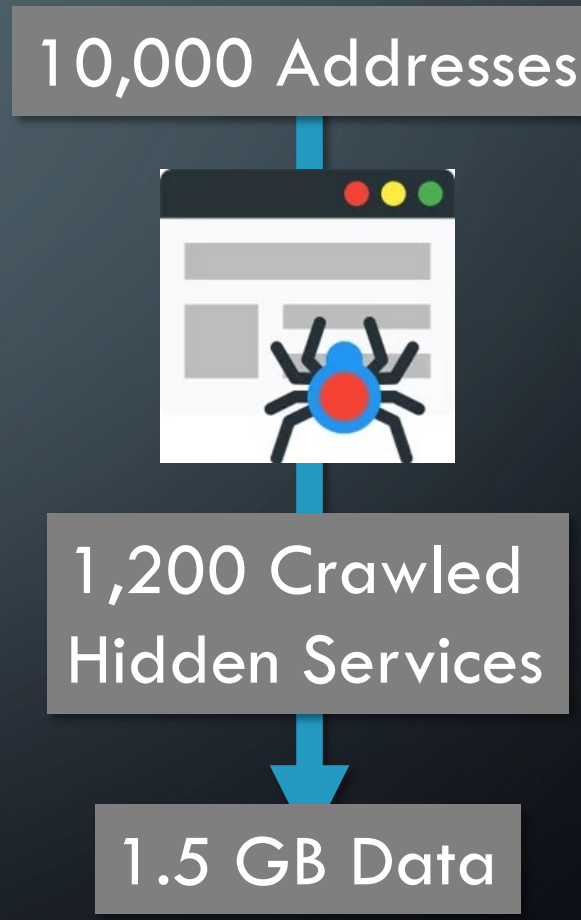


Edges



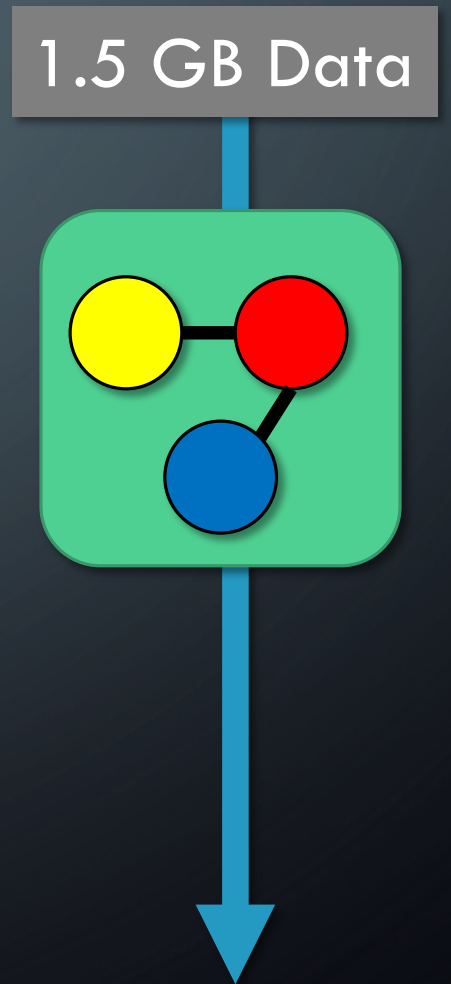
TOR WEB CRAWLER PERFORMANCE

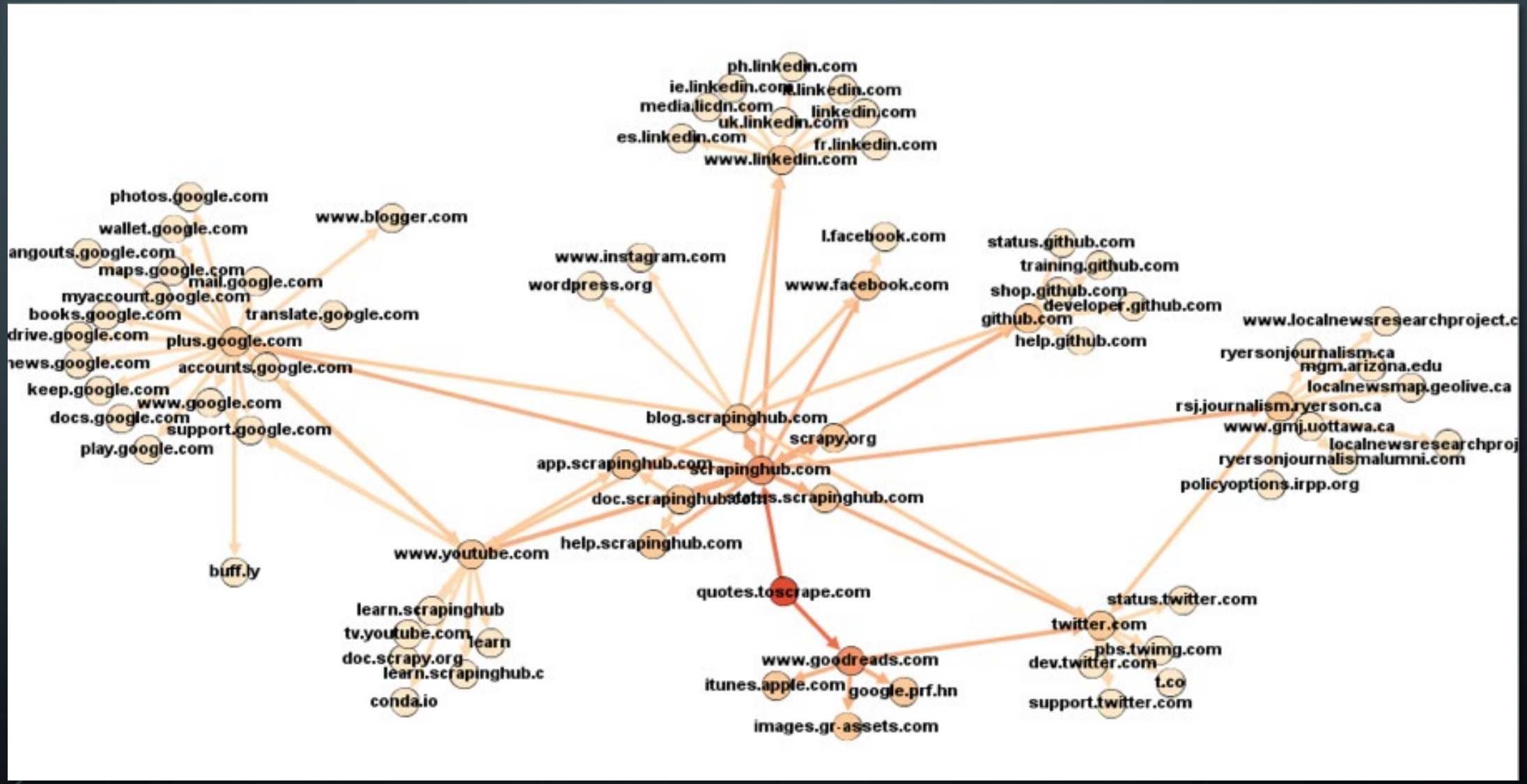
- Most extensive session based on starting list of ~10,000 URLs
- Over the course of four hours:
  - ~1,200 hidden services successfully crawled
  - ~1.5 GB total HTML downloaded
- Acceptable results, as roughly 85% of hidden services are short-lived (Owen & Savage)



CONNECTIVITY GRAPH PROGRAM PERFORMANCE

- Successfully reads crawler output files
- Generates graphs in multiple formats
  - These can be viewed by external applications
- Can be made more legible by applying graph-drawing algorithms



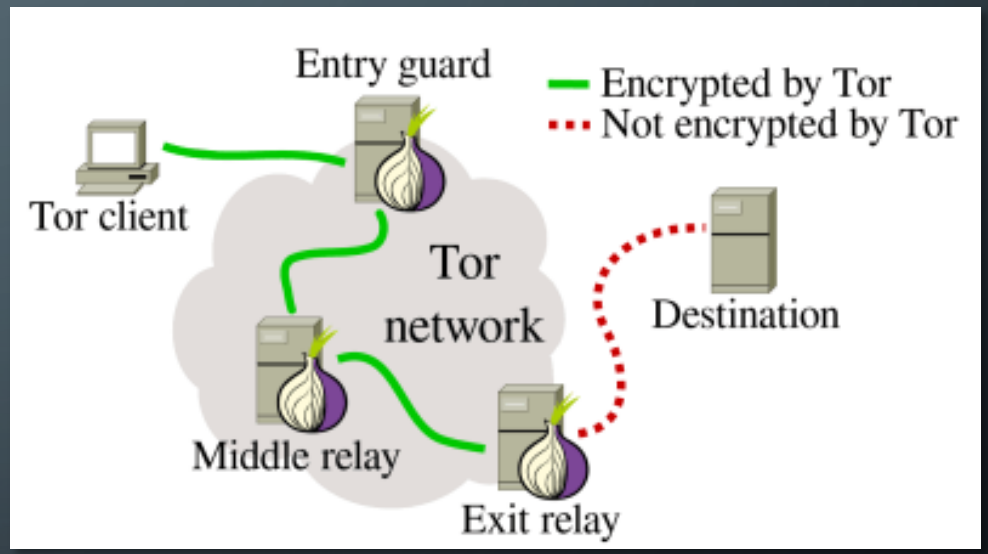

ph.linkedin.com
ie.linkedin.com
media.licdn.com
es.linkedin.com
uk.linkedin.com
linkedin.com
fr.linkedin.com
www.linkedin.com
photos.google.com
wallet.google.com
www.blogger.com
maps.google.com
mail.google.com
myaccount.google.com
books.google.com
translate.google.com
plus.google.com
accounts.google.com
keep.google.com
www.google.com
docs.google.com
support.google.com
play.google.com
buff.ly
www.youtube.com
learn.scrapinghub
tv.youtube.com
learn
doc.scrapy.org
learn.scrapinghub.c
conda.io
www.instagram.com
wordpress.org
l.facebook.com
www.facebook.com
blog.scrapinghub.com
scrapy.org
app.scrapinghub.com
scrapinghub.com
doc.scrapinghub.com
help.scrapinghub.com
quotes.toscrape.com
www.goodreads.com
itunes.apple.com
google.prf.hn
images.gr-assets.com
status.github.com
training.github.com
shop.github.com
developer.github.com
github.com
help.github.com
www.localnewsresearchproject.c
ryersonjournalism.ca
mgm.arizona.edu
localnewsmap.geolive.ca
rsj.journalism.ryerson.ca
www.gmj.uottawa.ca
localnewsresearchproj
ryersonjournalismalumni.com
policyoptions.irpp.org
status.twitter.com
twitter.com
pbs.twimg.com
dev.twitter.com
t.co
support.twitter.com

PROJECT DIFFICULTIES

- Crawler accidentally downloaded stolen personally identifiable information (PII)
- Security concern
- Hard drives confiscated and destroyed!

CONCLUSIONS

- The Tor Web Crawler was able to effectively traverse the Tor network and download hidden service content with no difficulties
- The Connectivity Graph program was successful at depicting links between hidden services



ACKNOWLEDGEMENTS

Josh Lyle – Argonne National Laboratory

Dr. Rodney L. Summerscales – Department of Engineering & Computer Science

SOURCE CODE

https://github.com/Argonne-National-Laboratory/torantula

BIBLIOGRAPHY

- Moore, Daniel, and Thomas Rid. "Cryptopolitik and the Darknet." *Survival*, vol. 58, no. 1, Feb. 2016, pp. 7–38., doi:10.1080/00396338.2016.1142085.
- Owen, Gareth, and Nick Savage. "The Tor Dark Net." *Global Commission on Internet Governance*, ser. 20, 30 Sept. 2015. 20, CIGI publications, www.cigionline.org/publications/tor-dark-net.
- Spitters, Martijn, et al. "Towards a Comprehensive Insight into the Thematic Organization of the Tor Hidden Services." *2014 IEEE Joint Intelligence and Security Informatics Conference*, 2014, doi:10.1109/jisic.2014.40.

QUESTIONS?