CS 798: Digital Forensics and Incident Response

Lecture 16 - Deep Web and Anonymity

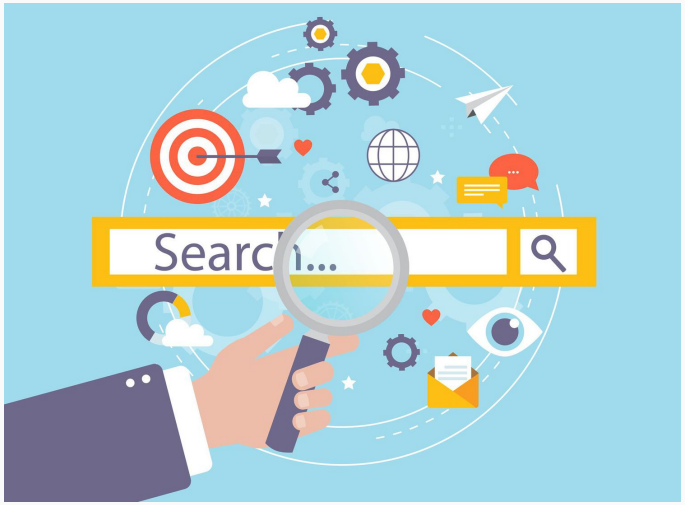
Diogo Barradas

Winter 2025

University of Waterloo

The web is a powerful tool for cybercriminals

- Provides huge source of information, used in:
  - e.g., extortion, privacy violations, identity theft
- The web allows for accessing services for criminal activity
  - e.g., drug selling, weapon selling
- To find services and info, there are powerful search engines
  - e.g., Google, Bing, DuckDuckGo, Shodan



It is also a powerful tool for investigators

- Powerful investigation tool about suspects
  - Find evidence in blogs, social networks, browsing activity
- The place where crimes themselves are carried out
  - Illegal transactions, cyber stalking, blackmail, fraud, etc.

Outline

The Web(s)

The iceberg analogy

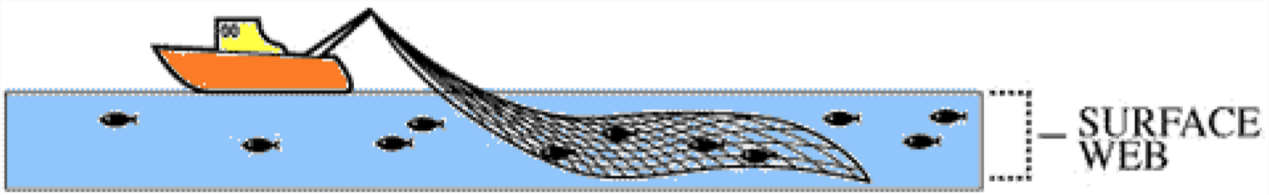
- What is "visible" through typical search engines is minimal
- The Deep Web is not inherently bad: it is just that its content is not directly indexed
- Part of the deep web where criminal activity is carried out is named the Dark Web

The Web(s)

The surface web

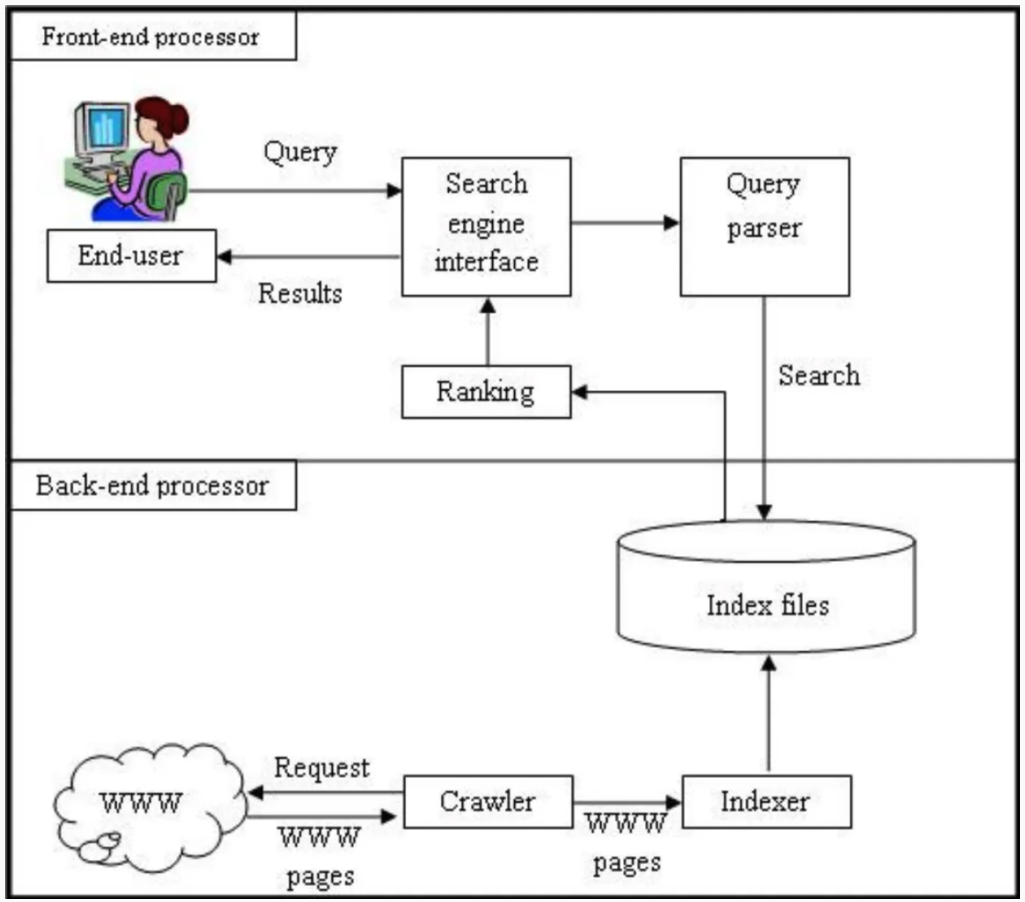
Surfing the Web

- The **Surface Web** is that portion of the World Wide Web that is readily available to the general public and searchable with standard web search engines
- As of June 14, 2015, Google's index of the surface web contains about 14.5 billion pages



How a typical search engine works

- Architecture of a typical search engine

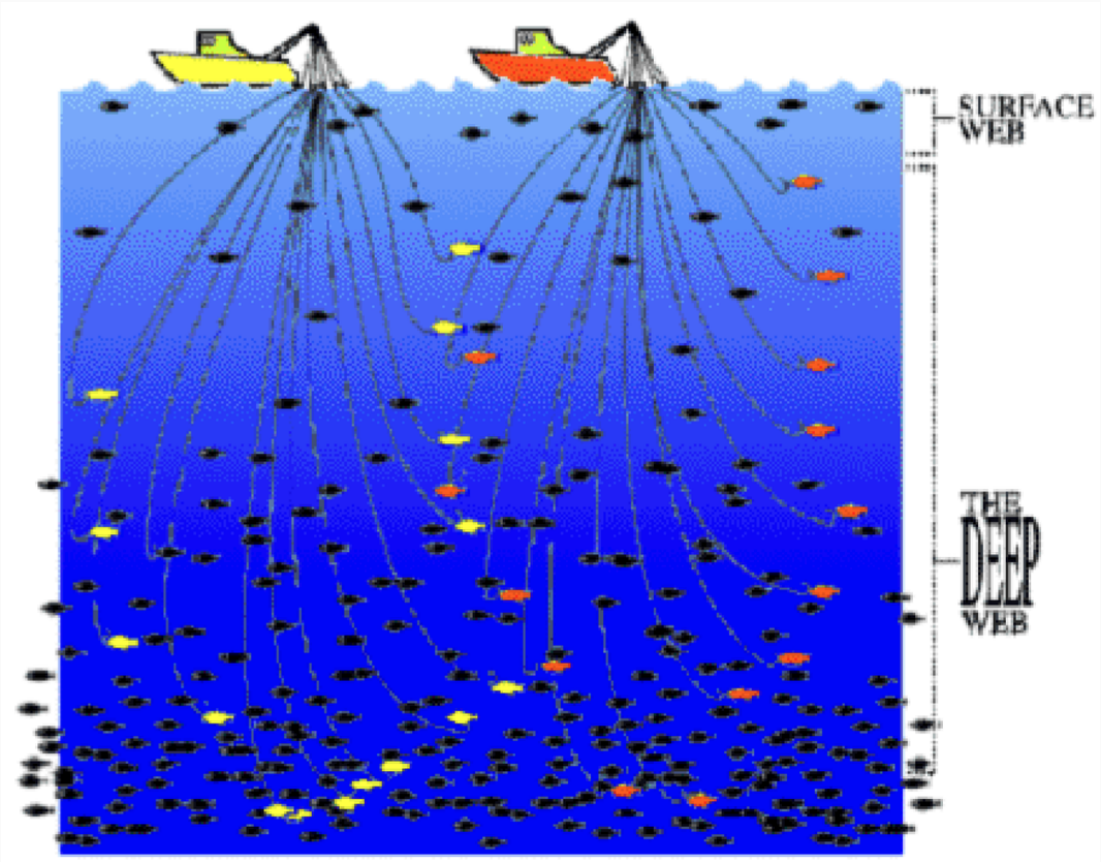


The Web(s)

The deep web

The deep web

- The **Deep Web** is the part of the Web which is not indexed by conventional search engines and therefore does not appear in search results
- Why is it not indexed by typical search engines?



Some content cannot be found through URL traversal

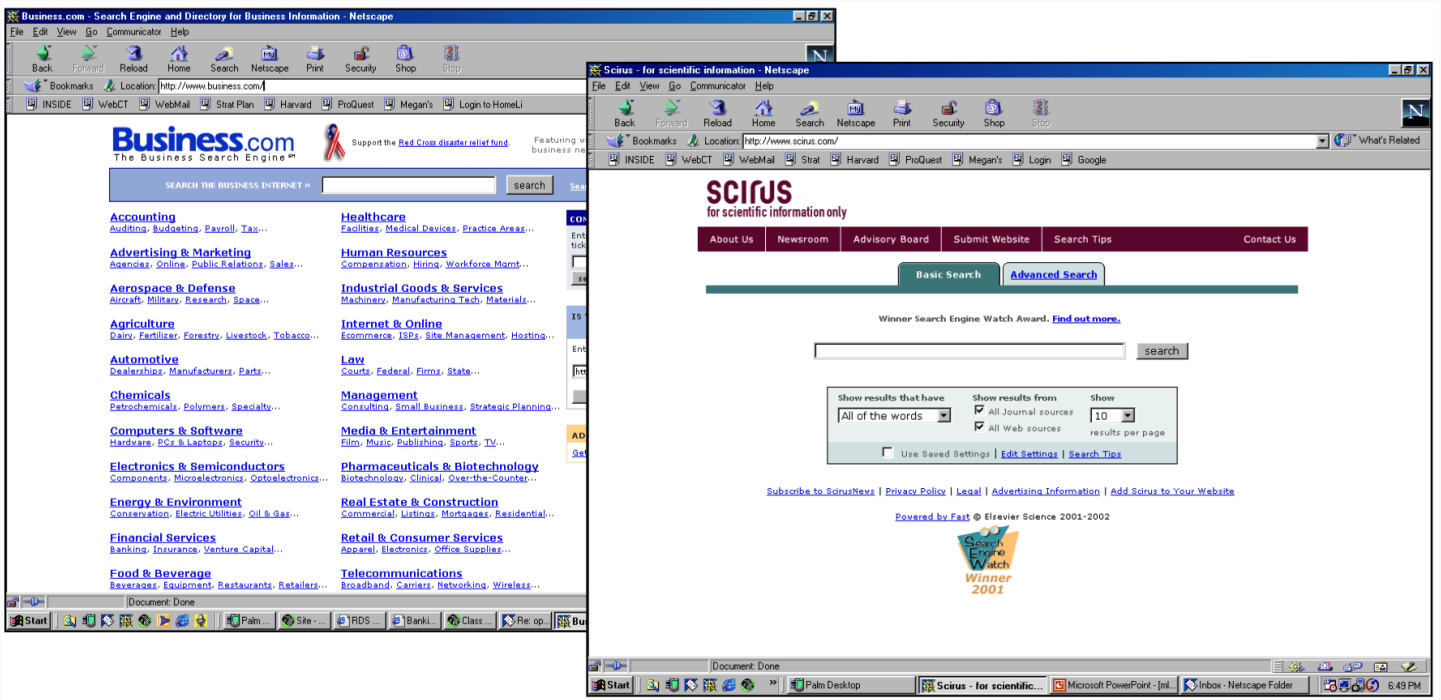
- **Dynamic content**:
  - Response to a query or accessed only through a form
- **Unlinked contents**:
  - Pages without any links to them (orphan)
- **Private web**:
  - Sites requiring registration and login
- **Limited access web**:
  - e.g., sites with captchas

Other times, content access is restricted or difficult

- Crawling restrictions by site owner
  - Use a `robots.txt` file to keep files off limits from spiders
- Crawling restrictions by the search engine
  - A page may be found this way: `http://www.website.com/cgi-bin/getpage.cgi?name=sitemap`
  - Most search engines will not read past "?" in the URL
- Limitations of the crawling engine
  - e.g., real-time data – websites change rapidly

Specialized search engines abound

- There's dozens of specialized search engines



A particularly interesting search engine

- Shodan lets the user find specific types of computers connected to the internet using a variety of filters
  - Routers, servers, traffic lights, security cameras, heating systems
  - Control systems for water parks, gas stations, water plants...
- Why is it interesting?
  - Many devices use “admin” as username and “1234” as password, and you can connect via a web browser

SHODAN

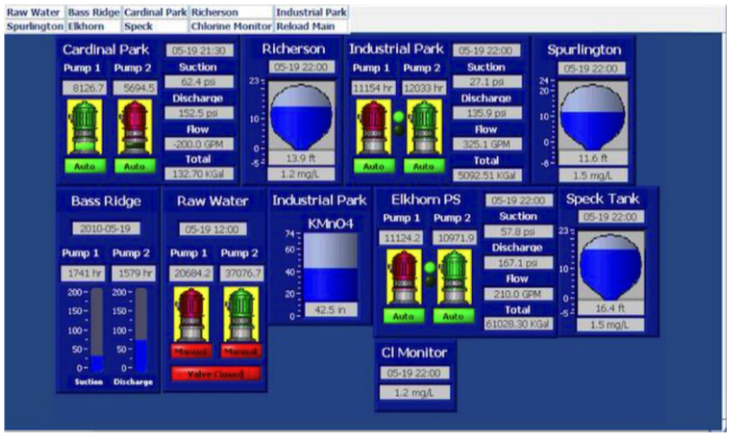
How does Shodan work?

"Google crawls URLs - I don't do that at all. The only thing I do is randomly pick an IP out of all the IPs that exist, whether it's online or not being used, and I try to connect to it on different ports. It's probably not a part of the visible web in the sense that you can't just use a browser. It's not something that most people can easily discover, just because it's not visual in the same way a website is."

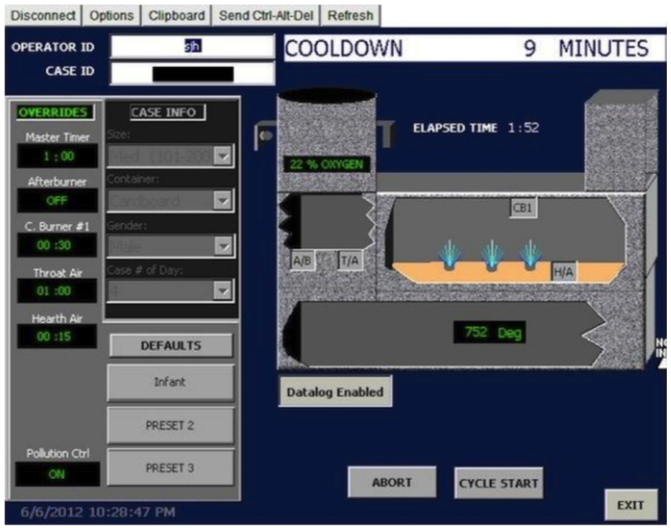
John Matherly, Shodan's creator

Interesting findings on Shodan

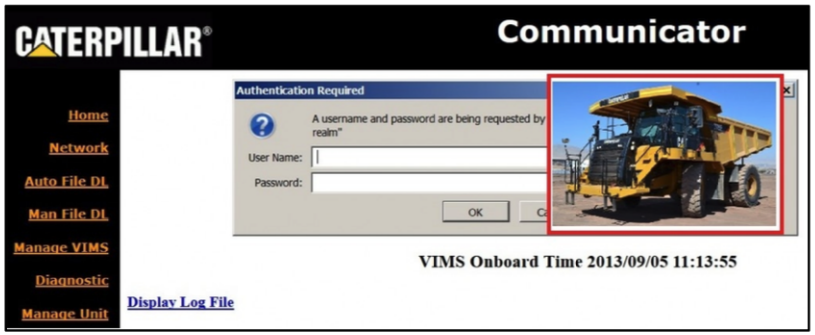
Webcam



Controls for water treatment facility



Controls for a crematorium



Controls of Caterpillar trucks

The Web(s)

The dark web

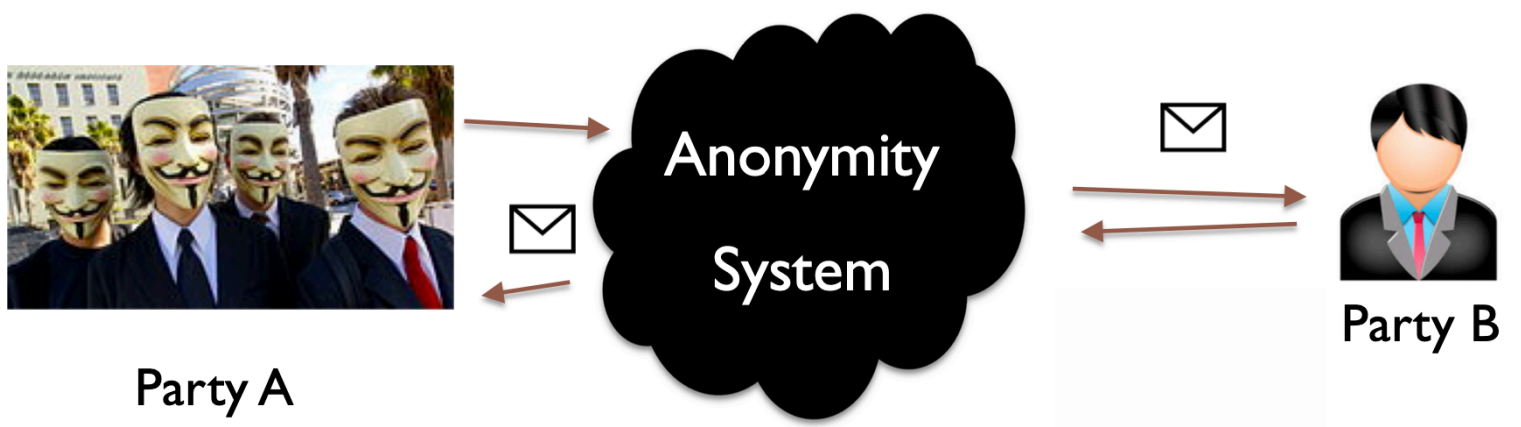
The dark web

- Mostly relies on anonymous communication systems
- We will learn more about this next

Anonymous Communication

Anonymous communication systems

- Aim to conceal the identity of communicating parties
- Can make a forensic investigator's life harder



Some simple notions of anonymity

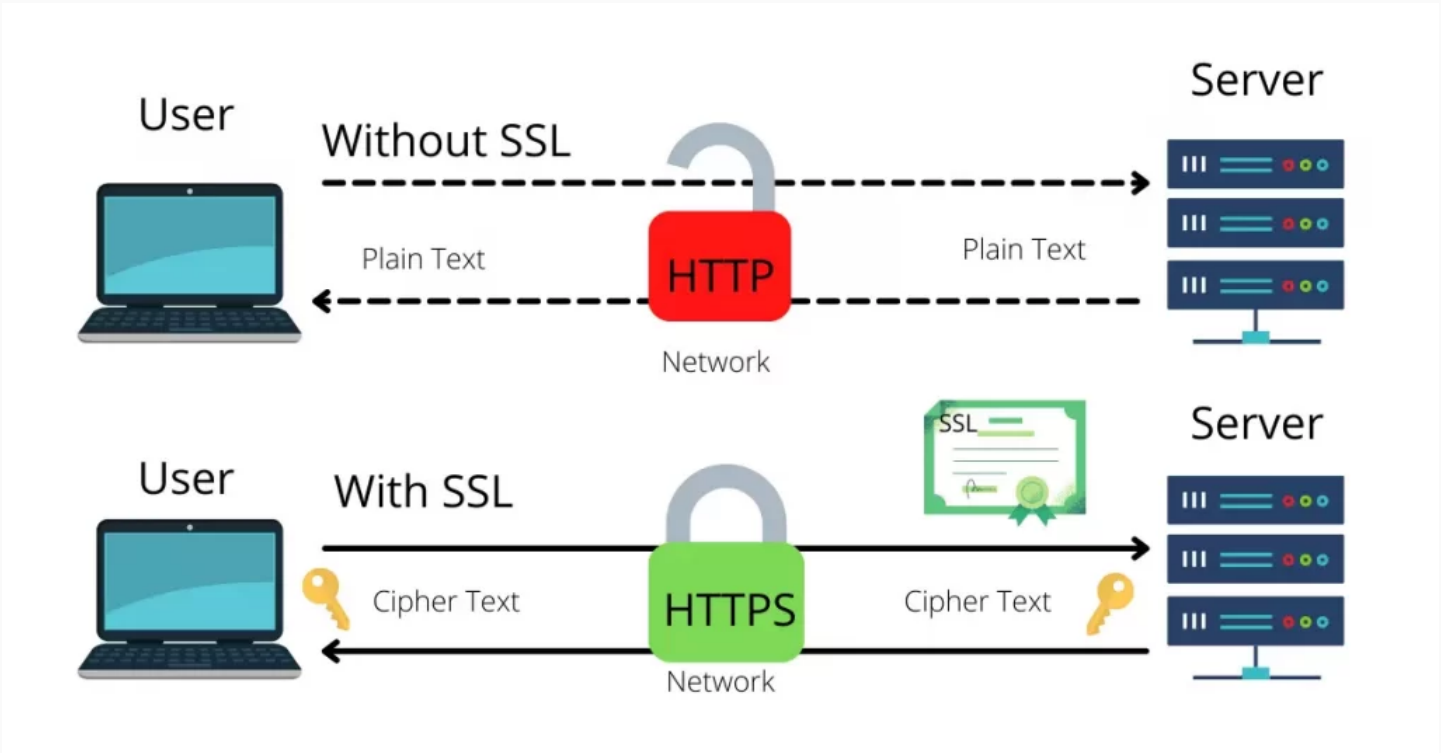
- **Sender anonymity**
  - Can't tell who sent this message
- **Receiver anonymity**
  - Can't tell who received this message
- **Sender-receiver anonymity**
  - Can't tell if A and B are respectively the sender and receiver of this message

Desired properties

- **Unlinkability**
  - The inability of link two or more items of interests to break anonymity, like packets, events, people, actions, etc.
- **Unobservability**
  - Items of interest are indistinguishable from all other items

Example of a connection with no anonymity

- A typical connection towards a server is trivially linkable
- Regardless of content being intelligible or not

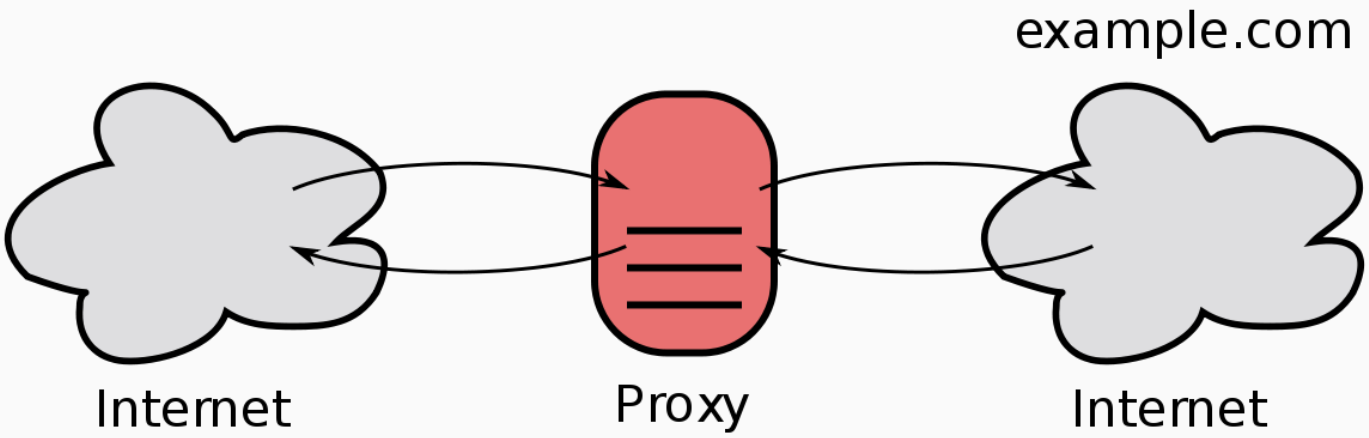


Anonymous Communication

Anonymizing proxies

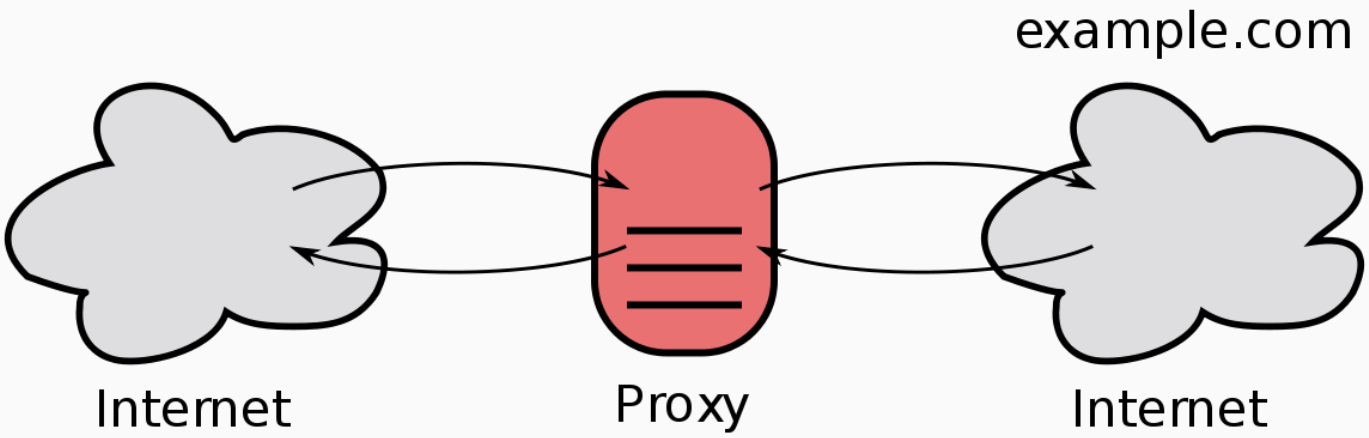
A basic approach for achieving anonymity

- Web proxies are set up to allow forwarding of HTTPS traffic
  - e.g., `https://www.anonymizer.com`, `http://www.bind2.com/`



A basic approach for achieving anonymity

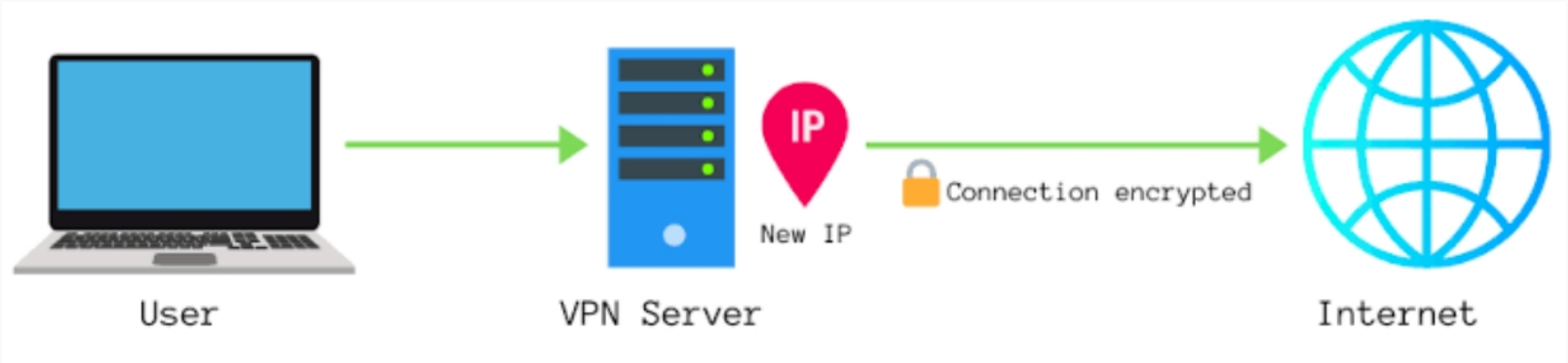
- Web proxies are set up to allow forwarding of HTTPS traffic
  - e.g., `https://www.anonymizer.com`, `http://www.bind2.com/`



What can an investigator located in different vantage points learn about this connection?

Similar properties for other anonymizing proxies

- Consider "anonymizing and secure" VPNs
  - Usually legitimate businesses
  - Accountable and prone to subpoenas



Sources of digital evidence

- **Messages exchanged through the system**
  - Email headers if an anonymous mailer is used
  - Network traces of the communication
- **Log files of the anonymous proxy**
  - Typically include source IP addresses and timestamp records
- **If the anonymous proxy is a paying service we can do more**
  - Request subscriber information for the account that was using the anonymizer to send the web based email

Anonymity networks

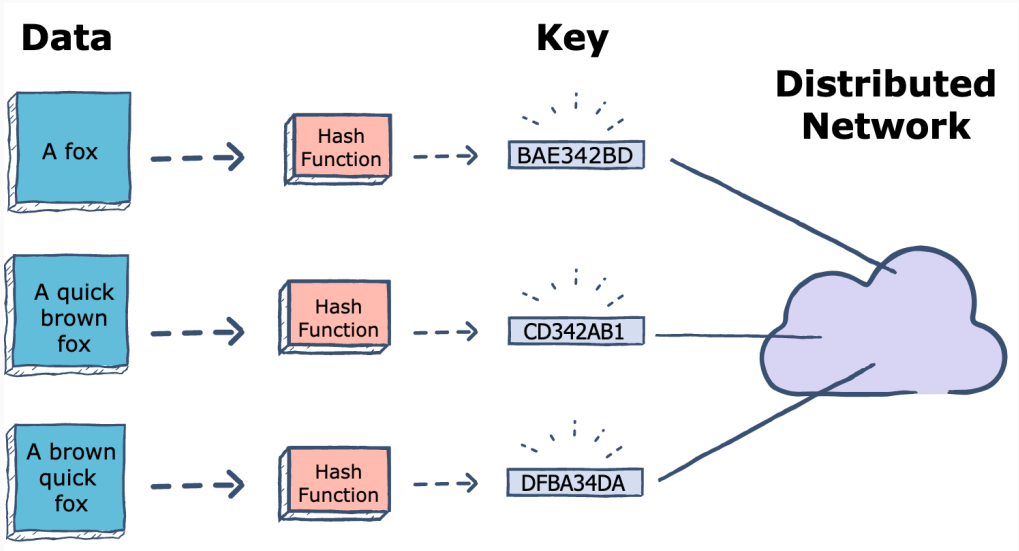
- **Anonymity networks** aim to overcome the limitations of proxy servers and VPNs
  - In particular, remove "single points of failure"
  - Also decrease the chances of successful traffic analysis
- Anonymity networks typically forward traffic through a set of nodes to make communication patterns hard to determine
- They can have multiple purposes, such as:
  - Anonymous Web access
  - Anonymous and decentralized file hosting/retrieval
  - Anonymous access to websites that are only accessible within the anonymity network itself

Anonymous Communication

Freenet

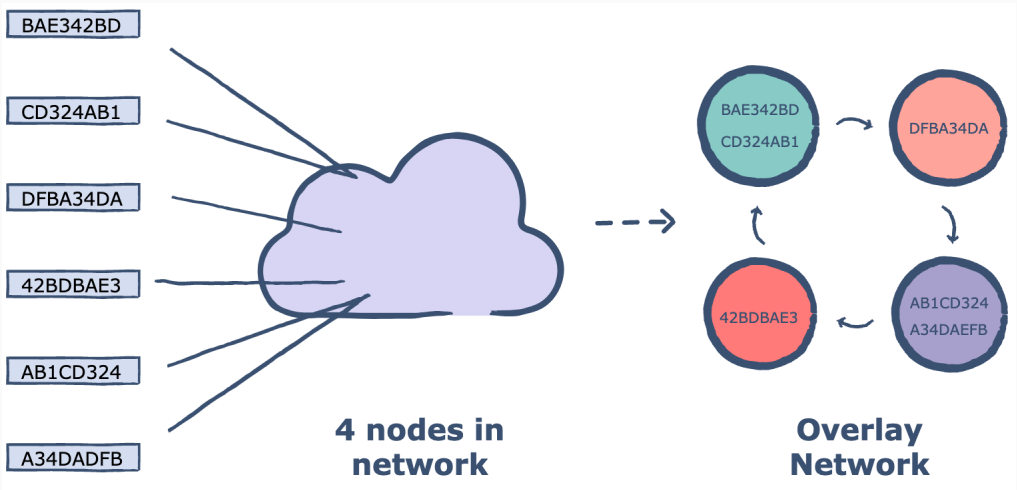
A primer on distributed hash tables

- A **distributed hash table (DHT)** is a decentralized storage system
  - Stores key-value pairs
  - Each node in a DHT is responsible for keys along with the mapped values
  - Provides lookup and storage schemes similar to a hash table
  - Any node can efficiently retrieve the value associated with a given key



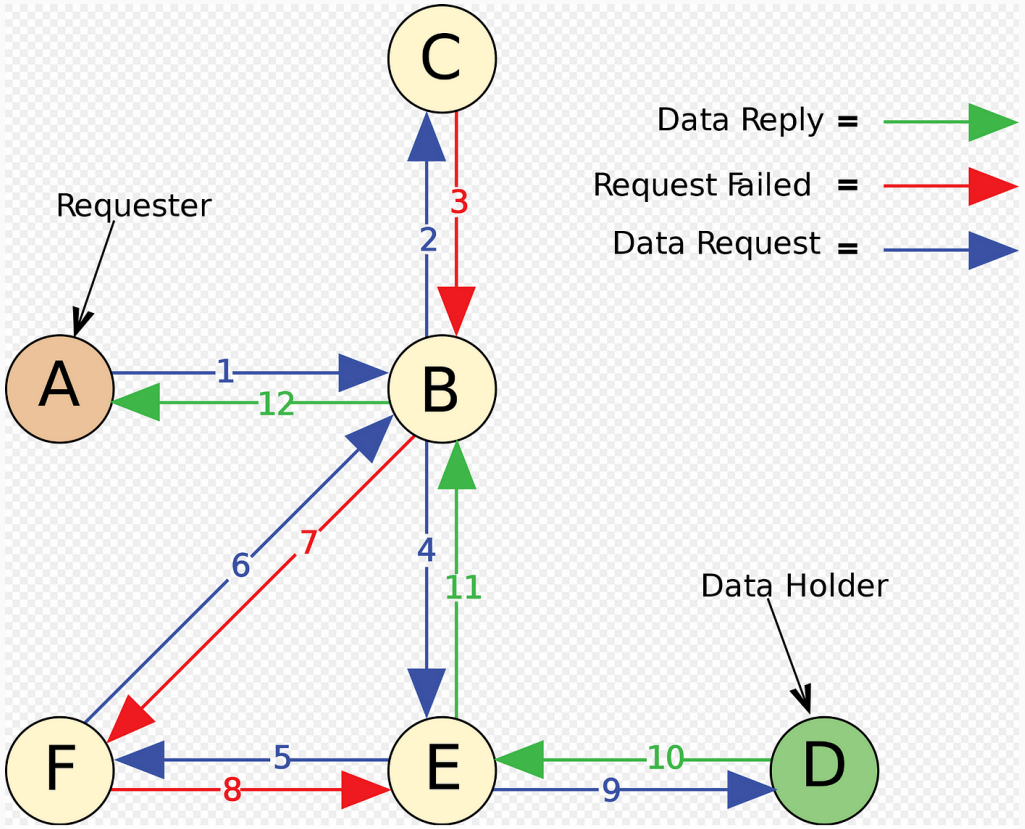
Overlay networks

- An **overlay network** connects nodes in a DHT
  - Each node maintains a set of links to other nodes (its neighbours on a routing table)
  - A node picks its neighbors according to a certain structure
- Overlay networks often make use of **key-based routing**
  - **Essential property**: for any key k, each node either has a node ID that owns k or has a link to a node ID which is closer to k



Key-based routing in action

- The request moves through the network from node to node, backing out of a dead-end (step 3) and a loop (step 7) before locating the desired file



Freenet (now called Hyphanet)

- A decentralized P2P network
  - Files are split into small chunks and stored across many nodes
  - Provides anonymity to producers and consumers
- Freenet is **self-contained**
  - i.e., you can only access contents that live on Freenet itself

Freenet's plausible deniability

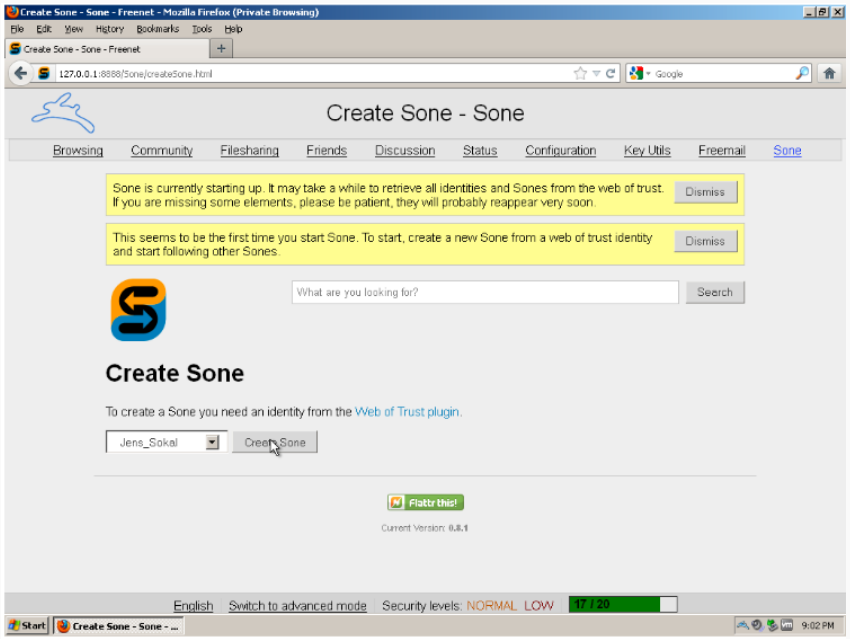
- Data requests are forwarded across several nodes
  - None of which knows whether they received an original or forwarded data request
- Difficult to distinguish nodes that share and access data from those who merely relay it

"I have nothing to do with that!"

What can I host on Freenet?

- Freenet supports different kinds of applications beyond mere file storage
- These applications live on “freesites” (websites accessible only through Freenet)

Newsgroup-like messaging, File sharing, Encrypted messages



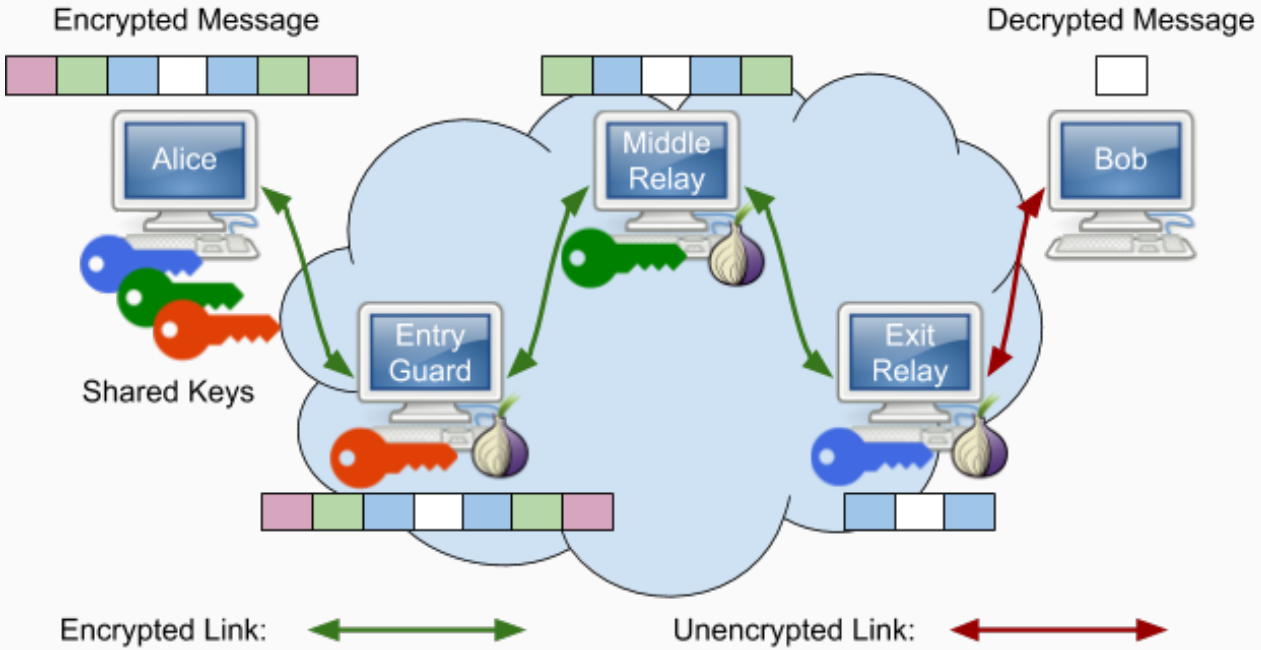
Social networks

Anonymous Communication

Tor

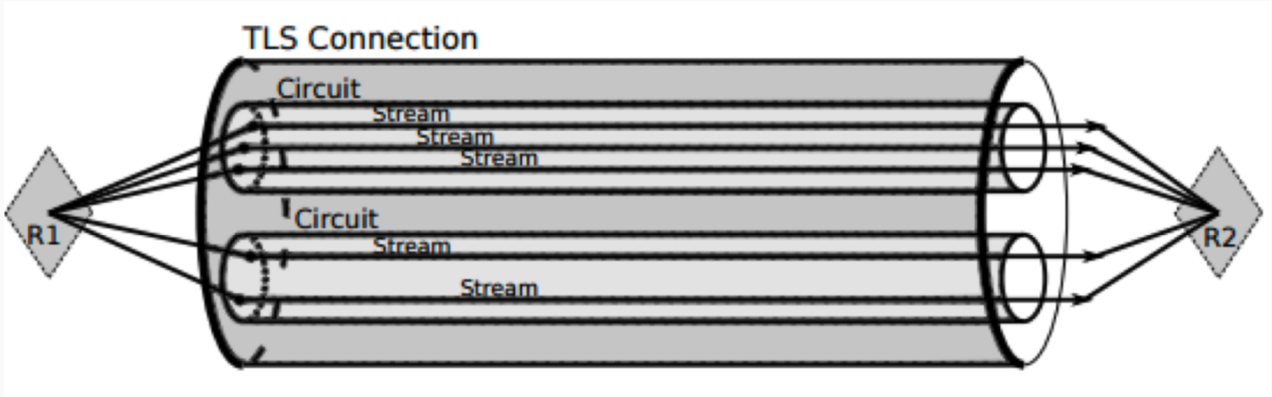
Onion routing

- **Idea**: Use public-key cryptography to establish a circuit with pairwise symmetric keys between relays
  - Sender chooses sequence of relays: some may be honest some controlled by an investigator; sender controls length of path
  - Routing info for each link is encrypted with the node's pubkey
  - Each relay learns only the identity of the previous/next relay



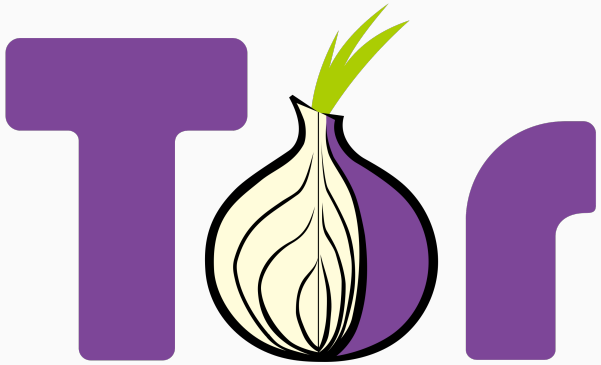
Circuits and streams multiplexing

- After the circuit has been built, a client can start transmitting and receiving data over this circuit
- All TCP connections of the user's application are translated into Tor streams which are multiplexed over the circuit



Tor: the second generation onion router

- Tor is a volunteer-based low-latency anonymity network
  - Anonymizes traffic by forwarding it through a circuit of relays
  - Essentially based on onion routing
  - Using Tor makes it more difficult for non-global adversaries to trace Internet activities

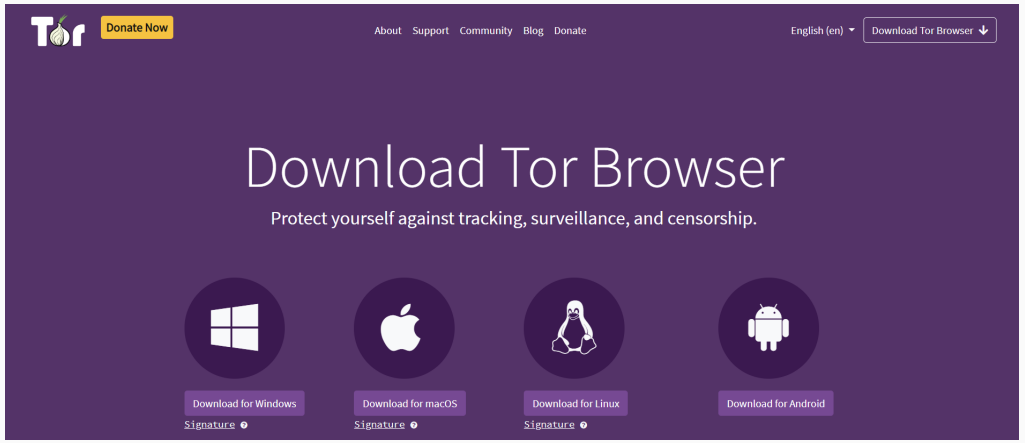


Who uses Tor?

- Pretty much everyone can use Tor, and it is quite useful for individuals facing specific circumstances
  - Journalists
  - Law enforcement
  - Human rights activists
  - Intelligence/military personnel
  - Victims of abuse
  - Regular Internet users
  - Criminals?

How to use Tor?

- Download, install, and execute the Tor client
  - The client acts as a SOCKS proxy
  - The client builds and maintains circuits
- Configure a browser to use the Tor client as a proxy
  - Any app that supports SOCKS proxies will work with Tor
  - You can also use The Tor Browser

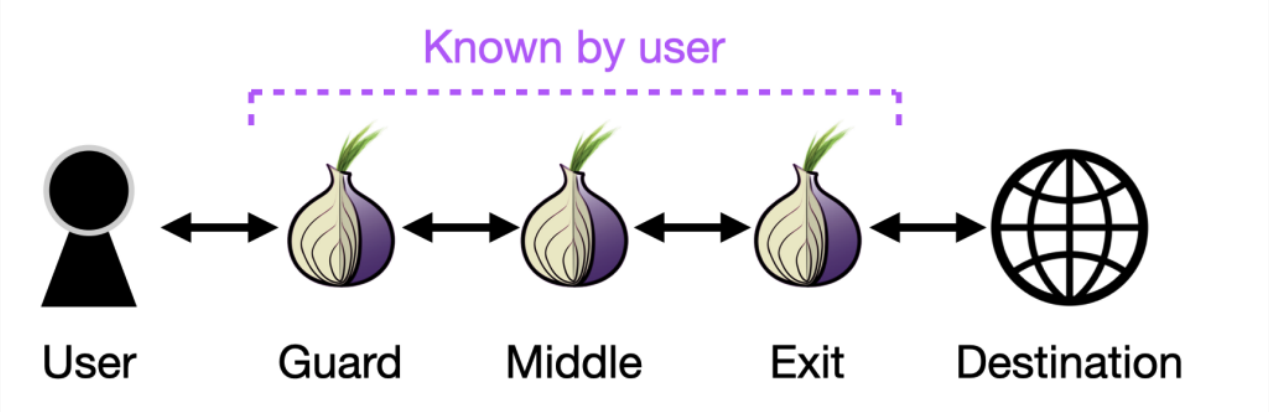


How are relays selected?

- Check the Tor **consensus file**
  - Hosted by trusted directory servers
  - Lists all known relays
- Tor does not select relays randomly
  - Chance of selection is proportional to bandwidth
  - Also several rules:
    - Same relay not chosen twice for the same circuit
    - No more than one relay in a given /16 subnet
    - Entry node must be a guard relay

Hosting and accessing anonymous content in Tor

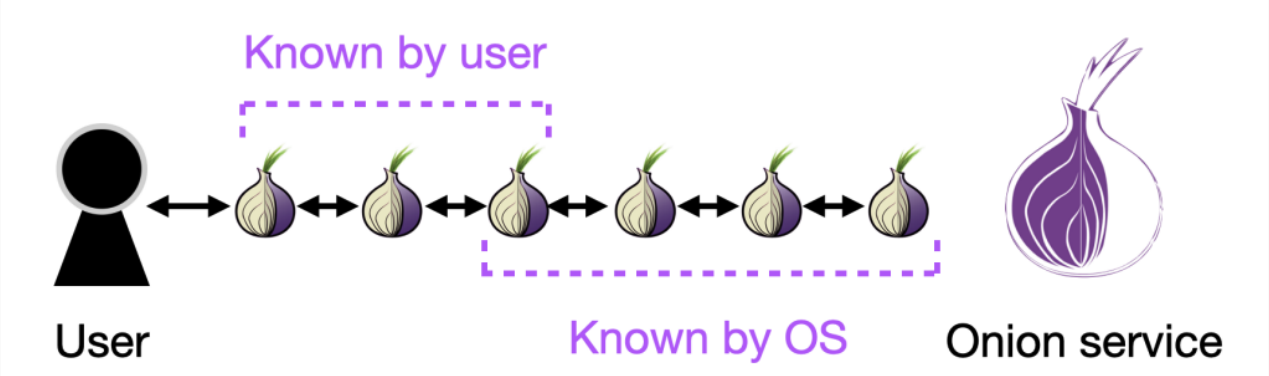
- Traditional Tor circuits are good for hiding the source of traffic
  - But the destination is typically a "clearweb" website
  - i.e., a website with a publicly known IP address



- What if I want to run websites that cannot be located?
  - i.e., shroud the website's IP address

Tor onion services

- Tor onion services are accessible only via Tor
  - They are identified by an identity public key
    - e.g., `vww6ybal4bd7szmgncyruucpgfkqahzddi37ktceo3ah7ngmcopnpyyd.onion`
  - A circuit now consists of 6 relays (3 chosen by the client, and 3 chosen by the onion service)



Investigating the Dark Web

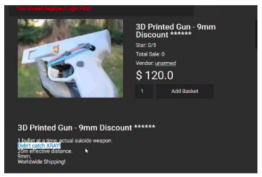
A note on the dark web

- **Darknets** fuel the dark web, and can be a haven for cyber-criminal activities
- Darknets are overlay networks which use the public Internet but require specific software or authorization to access
  - e.g., “freesites” Freenet/Hyphanet
  - e.g., Tor onion services

Sale of narcotics

Fraud and identity theft


“Cybercrime-as-a-Service”



Gun trafficking


Abuse material

The dark side of anonymity (Freenet)

PRESS RELEASE

Lincoln Man Sentenced to 21 1/2 Years in Prison for Production of Child Pornography

Friday, August 5, 2022

Share >

For Immediate Release

U.S. Attorney's Office, District of Nebraska

Acting United States Attorney Steven Russell announced that Matt Tibbels, 59, of Lincoln, Nebraska, was sentenced today in Lincoln by United States District Judge John M. Gerrard for production of child pornography. Tibbels was sentenced to 21 1/2 years in prison and 10 years of supervised release. There is no parole in the federal system. Tibbels was additionally ordered to pay $29,000 which will contribute to funds established for victims of these types of cases.

This case began as a part of a Freenet peer-to-peer investigation by the FBI concerning a Freenet user who was receiving child pornography. The FBI traced IP addresses associated with the investigation to Tibbels's residence. The residence was also operated as an in-home daycare. On October 5, 2021, a search warrant was executed where multiple devices were seized. A forensic examination of the devices revealed a total of 887 child pornography image files and 147 child pornography video files. While reviewing other electronic devices seized, FBI agents observed a video and seven images that depicted a child changing clothes in the main bedroom of the Tibbels' family home. This child was positively identified as a child who had attended that daycare.

The dark side of anonymity (Tor)

PRESS RELEASE

Four Men Sentenced to Prison for Engaging in a Child Exploitation Enterprise on the Tor Network

Monday, August 12, 2019

Share

For Immediate Release

Office of Public Affairs

The creator and lead administrator of a highly sophisticated Tor-network-based website dedicated to the sexual abuse of children was sentenced Friday, along with three others, for their roles in this global child exploitation enterprise.

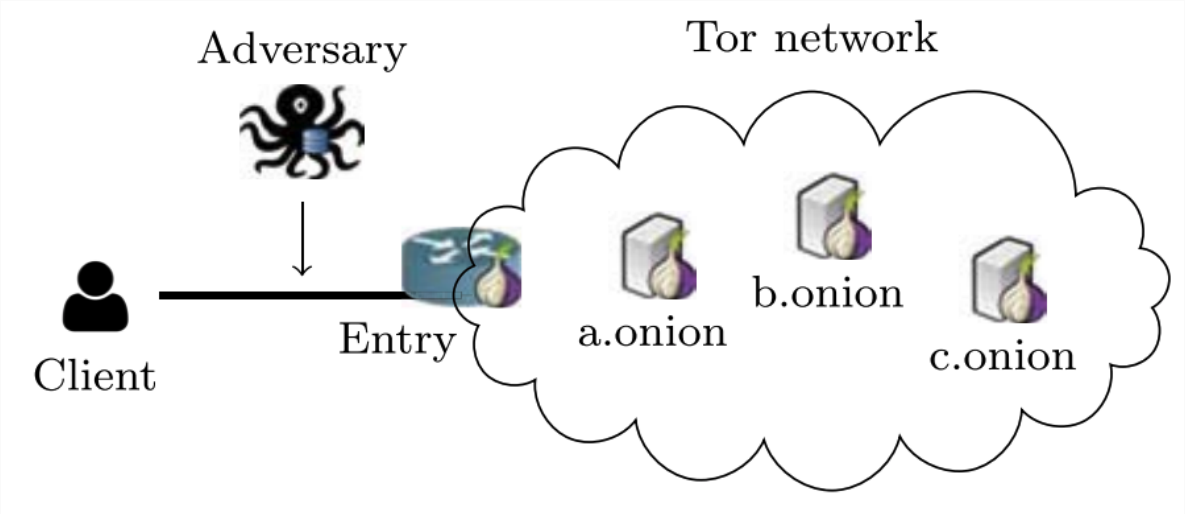
Towards forensically sound methods for deanonymization

- **Forensic soundness**:
  - A method that is based on a testable hypothesis, has a known error rate, follows existing standards, and uses generally accepted methods
- **Identifying uploaders and downloaders on Freenet**:
  - TLDR: Investigators planted nodes on Freenet and waited for them to be used as relays for illegal contents

A Forensically Sound Method of Identifying Downloaders and Uploaders in Freenet, Levine et al., CCS'20

Towards forensically sound methods for deanonymization (2)

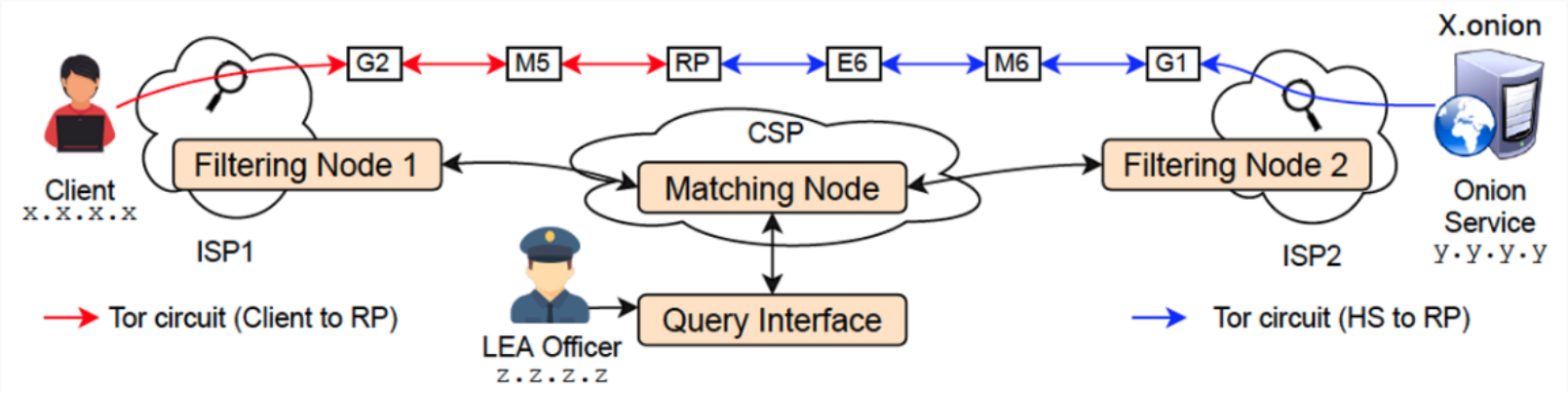
- **Onion service fingerprinting**:
  - TLDR: Website fingerprinting attacks can identify traffic patterns that are unique to a webpage



How Unique is Your .onion? An Analysis of the Fingerprintability of Tor Onion Services, Overdorf et al., CCS'17

Towards forensically sound methods for deanonymization (3)

- **Traffic correlation**:
  - TLDR: Investigators collect Tor flow information in different vantage points and attempt to match clients with servers hosting OSes



Flow Correlation Attacks on Tor Onion Service Sessions with Sliding Subset Sum, Lopes et al., NDSS'24

Takeaways

- The Deep Web includes the largest bulk of the Web, and there are specific search engines for searching through it.
- Anonymity systems enable general users to preserve their privacy online, but can be used for criminal purposes
- Anonymity systems have evolved so much that that it is hard for forensic investigators to thwart them

Pointers

- **Textbook:**
  - Chapter 23.3–4 [Casey]
- **Literature:**
  - Freenet: A Distributed Anonymous Information Storage and Retrieval System
  - Tor: The Second-Generation Onion Router
  - Reports of Child Sexual Abuse Materials (CSAM) on Tor, Freenet, and i2P
  - Shining Light on Internet-based Crimes Against Children (YouTube)
  - A Forensically Sound Method of Identifying Downloaders and Uploaders in Freenet
  - How Unique is Your .onion? An Analysis of the Fingerprintability of Tor Onion Services
- **Acknowledgements:**
  - Adapted and extended from Nuno Santos's Forensics Cyber-Security course at Técnico Lisbon
  - DHT/overlays drawings from educative.io