CSci 5271
Introduction to Computer Security
Day 25: Tor and LLM safety

Stephen McCamant

University of Minnesota, Computer Science & Engineering

Outline

Anonymous communications techniques, cont'd

Tor basics

Announcements intermission

Tor experiences and challenges

AI/LLM safety and security

DNSSEC

Traffic analysis

- What can you learn from encrypted data? A lot
- Content size, timing
- Who's talking to who
  - → countermeasure: anonymity

Nymity slider (Goldberg)

- Verinymity
  - Social security number
- Persistent pseudonymity
  - Pen name ("George Eliot"), "moot"
- Linkable anonymity
  - Frequent-shopper card
- Unlinkable anonymity
  - (Idealized) cash payments

Nymity ratchet?

- It's easy to add names on top of an anonymous protocol
- The opposite direction is harder
- But, we're stuck with the Internet as is
- So, add anonymity to conceal underlying identities

Steganography

- One approach: hide real content within bland-looking cover traffic
- Classic: hide data in least-significant bits of images
- Easy to fool casual inspection, hard if adversary knows the scheme

Dining cryptographers

Photo credits: Nick Hopper - UMN CS; Bryan Ford - personal home page; Ian Goldberg - Kat Hanna CC-BY-SA-3.0; Nadia Heninger - personal home page

Dining cryptographers

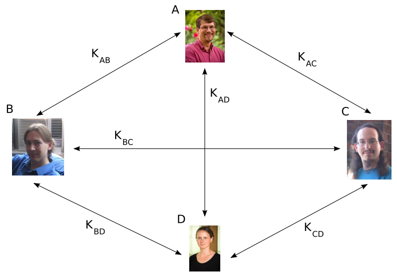
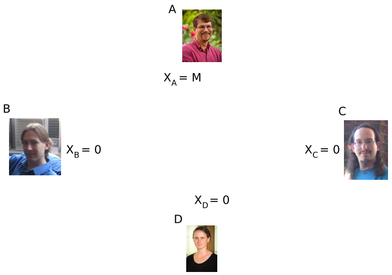


Photo credits: Nick Hopper - UMN CS; Bryan Ford - personal home page; Ian Goldberg - Kat Hanna CC-BY-SA-3.0; Nadia Heninger - personal home page

Dining cryptographers



A

$X_A = M$

B

$X_B = 0$

C

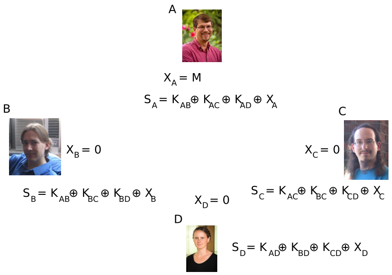
$X_C = 0$

$X_D = 0$

D

Photo credits: Nick Hopper - UMN CS; Bryan Ford - personal home page; Ian Goldberg - Kat Hanna CC-BY-SA 3.0; Nadia Heninger - personal home page

Dining cryptographers



A

$X_A = M$

$S_A = K_{AB} \oplus K_{AC} \oplus K_{AD} \oplus X_A$

B

$X_B = 0$

C

$X_C = 0$

$S_B = K_{AB} \oplus K_{BC} \oplus K_{BD} \oplus X_B$

$X_D = 0$

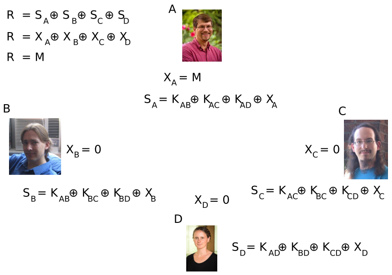
$S_C = K_{AC} \oplus K_{BC} \oplus K_{CD} \oplus X_C$

D

$S_D = K_{AD} \oplus K_{BD} \oplus K_{CD} \oplus X_D$

Photo credits: Nick Hopper - UMN CS; Bryan Ford - personal home page; Ian Goldberg - Kat Hanna CC-BY-SA 3.0; Nadia Heninger - personal home page

Dining cryptographers



$R = S_A \oplus S_B \oplus S_C \oplus S_D$

$R = X_A \oplus X_B \oplus X_C \oplus X_D$

$R = M$

A

$X_A = M$

$S_A = K_{AB} \oplus K_{AC} \oplus K_{AD} \oplus X_A$

B

$X_B = 0$

C

$X_C = 0$

$S_B = K_{AB} \oplus K_{BC} \oplus K_{BD} \oplus X_B$

$X_D = 0$

$S_C = K_{AC} \oplus K_{BC} \oplus K_{CD} \oplus X_C$

D

$S_D = K_{AD} \oplus K_{BD} \oplus K_{CD} \oplus X_D$

Photo credits: Nick Hopper - UMN CS; Bryan Ford - personal home page; Ian Goldberg - Kat Hanna CC-BY-SA 3.0; Nadia Heninger - personal home page

DC-net challenges

- Quadratic key setups and message exchanges per round
- Scheduling who talks when
- One traitor can anonymously sabotage
- Improvements subject of ongoing research

Mixing/shuffling

- Computer analogue of shaking a ballot box, etc.
- Reorder encrypted messages by a random permutation
- Building block in larger protocols
- Distributed and verifiable variants possible as well

Anonymous remailers

- Anonymizing intermediaries for email
  - First cuts had single points of failure
- Mix and forward messages after receiving a sufficiently-large batch
- Chain together mixes with multiple layers of encryption
- Fancy systems didn't get critical mass of users

Outline

Anonymous communications techniques, cont'd

Tor basics

Announcements intermission

Tor experiences and challenges

AI/LLM safety and security

DNSSEC

Tor: an overlay network

- Tor (originally from "the onion router")
  - `https://www.torproject.org/`
- An anonymous network built on top of the non-anonymous Internet
- Designed to support a wide variety of anonymity use cases

Low-latency TCP applications

- Tor works by proxying TCP streams
  - (And DNS lookups)
- Focuses on achieving interactive latency
  - WWW, but potentially also chat, SSH, etc.
  - Anonymity tradeoffs compared to remailers

Tor Onion routing

- Stream from sender to D forwarded via A, B, and C
  - One Tor circuit made of four TCP hops
- Encrypt packets (512-byte “cells”) as $E_A(B, E_B(C, E_C(D, P)))$
- TLS-like hybrid encryption with “telescoping” path setup

Client perspective

- Install Tor client running in background
- Configure browser to use Tor as proxy
  - Or complete Tor+Proxy+Browser bundle
- Browse web as normal, but a lot slower
  - Also, sometimes `google.com` is in Swedish

Entry/guard relays

- “Entry node”: first relay on path
- Entry knows the client’s identity, so particularly sensitive
  - Many attacks possible if one adversary controls entry and exit
- Choose a small random set of “guards” as only entries to use
  - Rotate slowly or if necessary
- For repeat users, better than random each time

Exit relays

- Forwards traffic to/from non-Tor destination
- Focal point for anti-abuse policies
  - E.g., no exits will forward for port 25 (email sending)
- Can see plaintext traffic, so danger of sniffing, middleperson, etc.

Centralized directory

- How to find relays in the first place?
- Straightforward current approach: central directory servers
- Relay information includes bandwidth, exit polices, public keys, etc.
- Replicated, but potential bottleneck for scalability and blocking

Outline

Anonymous communications techniques, cont’d

Tor basics

Announcements intermission

Tor experiences and challenges

AI/LLM safety and security

DNSSEC

Note to early readers

- This is the section of the slides most likely to change in the final version
- If class has already happened, make sure you have the latest slides for announcements

Outline

Anonymous communications techniques, cont'd

Tor basics

Announcements intermission

Tor experiences and challenges

AI/LLM safety and security

DNSSEC

Anonymity loves company

- Diverse user pool needed for anonymity to be meaningful
  - Hypothetical Department of Defense Anonymity Network
- Tor aims to be helpful to a broad range of (sympathetic sounding) potential users

Who (arguably) needs Tor?

- Consumers concerned about web tracking
- Businesses doing research on the competition
- Citizens of countries with Internet censorship
- Reporters protecting their sources
- Law enforcement investigating targets

Tor and the US government

- Onion routing research started with the US Navy
- Academic research still supported by NSF
- Anti-censorship work supported by the State Department
  - Same branch as Voice of America
- But also targeted by the NSA
  - Per Snowden, so far only limited success

Volunteer relays

- Tor relays are run basically by volunteers
  - Most are idealistic
  - A few have been less-ethical researchers, or GCHQ
- Never enough, or enough bandwidth
- P2P-style mandatory participation?
  - Unworkable/undesirable
- Various other kinds of incentives explored

Performance

- Increased latency from long paths
- Bandwidth limited by relays
- Recently 1-2 sec for 50KB, 3-7 sec for 1MB
- Historically worse for many periods
  - Flooding (guessed botnet) fall 2013

Anti-censorship

- As a web proxy, Tor is useful for getting around blocking
- Unless Tor itself is blocked, as it often is
- *Bridges* are special less-public entry points
- Also, protocol obfuscation arms race (uneven)

Hidden services

- Tor can be used by servers as well as clients
- Identified by cryptographic key, use special rendezvous protocol
- Servers often present easier attack surface

Undesirable users

- P2P filesharing
  - Discouraged by Tor developers, to little effect
- Terrorists
  - At least the NSA thinks so
- Illicit e-commerce
  - "Silk Road" and its successors

Intersection attacks

- Suppose you use Tor to update a pseudonymous blog, reveal you live in Minneapolis
- Comcast can tell who in the city was sending to Tor at the moment you post an entry
  - Anonymity set of 1000 → reasonable protection
- But if you keep posting, adversary can keep narrowing down the set

Exit sniffing

- Easy mistake to make: log in to an HTTP web site over Tor
- A malicious exit node could now steal your password
- Another reason to always use HTTPS for logins

Browser bundle JS attack

- Tor's Browser Bundle disables many features try to stop tracking
- But, JavaScript defaults to on
  - Usability for non-expert users
  - Fingerprinting via NoScript settings
- Was incompatible with Firefox auto-updating
- Many Tor users de-anonymized in August 2013 by JS vulnerability patched in June

Traffic confirmation attacks

- If the same entity controls both guard and exit on a circuit, many attacks can link the two connections
  - "Traffic confirmation attack"
  - Can't directly compare payload data, since it is encrypted
- Standard approach: insert and observe delays
- Protocol bug until recently: covert channel in hidden service lookup

Hidden service traffic conf.

- Bug allowed signal to guard when user looked up a hidden service
  - Non-statistical traffic confirmation
- For 5 months in 2014, 115 guard nodes (about 6%) participated in this attack
  - Apparently researchers at CMU's SEI/CERT
- Beyond "research," they also gave/sold info. to the FBI
  - Apparently used in Silk Road 2.0 prosecution, etc.

Outline

Anonymous communications techniques, cont'd

Tor basics

Announcements intermission

Tor experiences and challenges

AI/LLM safety and security

DNSSEC

Kinds of AI safety concerns

- AI failure and misuse: present-day negative consequences of AI not being smart enough, or being used by adversarial people
- AI alignment: long-term risks of AI behavior being inconsistent with human values

Business and social context

- Recent advances in AI are novel software being driven by big tech companies
- Short-term concern is showcasing the technology as useful and low-risk
  - Worthy of future investment but only light regulation
- The reading is a whitepaper from OpenAI around the time GPT-4 was released
  - Incentives to not leave risks out, but make them seem manageable

Normal security concerns

- Companies deploying LLMs have most of the normal security concerns
  - E.g., running a large public web site
- For commercial providers, keeping the models secret is a critical requirement

Relevance of threat modeling

- For AI-specific concerns, the main intersection with security is thinking about adversarial threats
- Main adversaries are:
  - Malicious users (short term)
  - Rogue AIs (longer term)

Unwanted/harmful content

- "Unwanted" for generative AI covers both:
  - Unwanted by the user: not following directions
  - Unwanted by the provider: fulfilling user requests would harm third parties or damage the provider's reputation

Exemplary harms from a chatbot

- Facilitating disinformation and political influence
  - Avoid things social media platforms have been criticized for
- Facilitating development of weapons
  - E.g., help an individual or low-resource group build a biological weapon
  - Support going beyond web search results

LLMs in computer security

- Lowest-hanging fruit is augmenting social engineering
- What about finding security bugs?
  - Dual use between defenders and attackers
  - Not yet very effective, interesting cases are harder than other code-support tasks
  - But could be a cause of a high-profile harmful incident

Emergent risks

- Scaling LLMs have often shown novel capabilities
  - Which ones are most concerning in amplifying AI risk?
- Planning, pursuing goals (positive applications too)
- Self-replication (e.g., compare computer worm)
- Real world influence and deception
  - Example: TaskRabbit to solve a CAPTCHA

Medium-term concerns

- Economic disruption
  - E.g., widespread job losses and unemployment
- Acceleration: positive feedback increasing the rate of AI development
  - Reckless competition towards AI goals
  - AI facilitating science and technological development

Some reasons alignment is hard

- Humans already can't agree among themselves on universal values
- Human desires have a lot of implicit side conditions and unstated restrictions
- We don't understand many details of how LLMs work internally
- If AIs become smarter than people, why would they want to obey us?

Hypothetical endpoints

- Paperclip maximizer
    - Seemingly simple goal + great capability = deeply undesirable result
- Will super-human AIs treat humans the way humans have treated non-human animals?
    - Extreme loss of agency is possible without destruction
    - Many different example animals and possible perspectives
    - Too close of an analogy may be unrealistic, since AI may be much less like us than animals are

Precaution and p(doom)

- A trending conversation topic is comparing estimates on the probability of a catastrophic outcome from AI
- Surprisingly many people working in AI have a significant p(doom)
    - Progress is inevitable, or it would be worse without me
- Choosing not to pursue technology because of downside risks is rare
    - Compare: nuclear weapons and energy

Outline

Anonymous communications techniques, cont'd

Tor basics

Announcements intermission

Tor experiences and challenges

AI/LLM safety and security

DNSSEC

DNS: trusted but vulnerable

- Almost every higher-level service interacts with DNS
- UDP protocol with no authentication or crypto
    - Lots of attacks possible
- Problems known for a long time, but challenge to fix compatibly

DNSSEC goals and non-goals

+ Authenticity of positive replies
+ Authenticity of negative replies
+ Integrity

— Confidentiality

— Availability

First cut: signatures and certificates

- Each resource record gets an `RRSIG` signature
    - E.g., `A` record for one name→address mapping
    - Observe: signature often larger than data
- Signature validation keys in `DNSKEY` RRs
- Recursive chain up to the root (or other "anchor")

Add more indirection

- DNS needs to scale to very large flat domains like `.com`
- Facilitated by having single `DS` RR in parent indicating delegation
- Chain to root now includes `DS`es as well

Negative answers

- Also don't want attackers to spoof non-existence
  - Gratuitous denial of service, force fallback, etc.
- But don't want to sign "`x` does not exist" for all `x`
- Solution 1, `NSEC`: "there is no name between `acacia` and `baobab`"

Preventing zone enumeration

- Many domains would not like people enumerating all their entries
- DNS is public, but "not that public"
- Unfortunately `NSEC` makes this trivial
- Compromise: `NSEC3` uses password-like salt and repeated hash, allows opt-out

DANE: linking TLS to DNSSEC

- "DNS-based Authentication of Named Entities"
- DNS contains hash of TLS cert, don't need CAs
- How is DNSSEC's tree of certs better than TLS's?

Signing the root

- Political problem: many already distrust US-centered nature of DNS infrastructure
- Practical problem: must be very secure with no single point of failure
- Finally accomplished in 2010
  - Solution involves 'key ceremonies', international committees, smart cards, safe deposit boxes, etc.

Deployment

- Standard deployment problem: all cost and no benefit to being first mover
- Servers working on it, mostly top-down
- Clients: still less than 20%
- Will probably be common for a while: insecure connection to secure resolver

What about privacy?

- Users increasingly want privacy for their DNS queries as well
- Older DNSCurve and DNSCrypt protocols were not standardized
- More recent "DNS over TLS" and "DNS over HTTPS" are RFCs
- DNS over HTTPS in major browsers might have serious centralization effects

Next time

- How usability affects security