COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

Anonymous Communication and Traffic Analysis II

Information Privacy with Applications

David Sidi (dsidi@email.arizona.edu)

Warm-up

- none today!

Small mention of interesting things

- MIT student work on freehaven led to Tor

Continuing last time

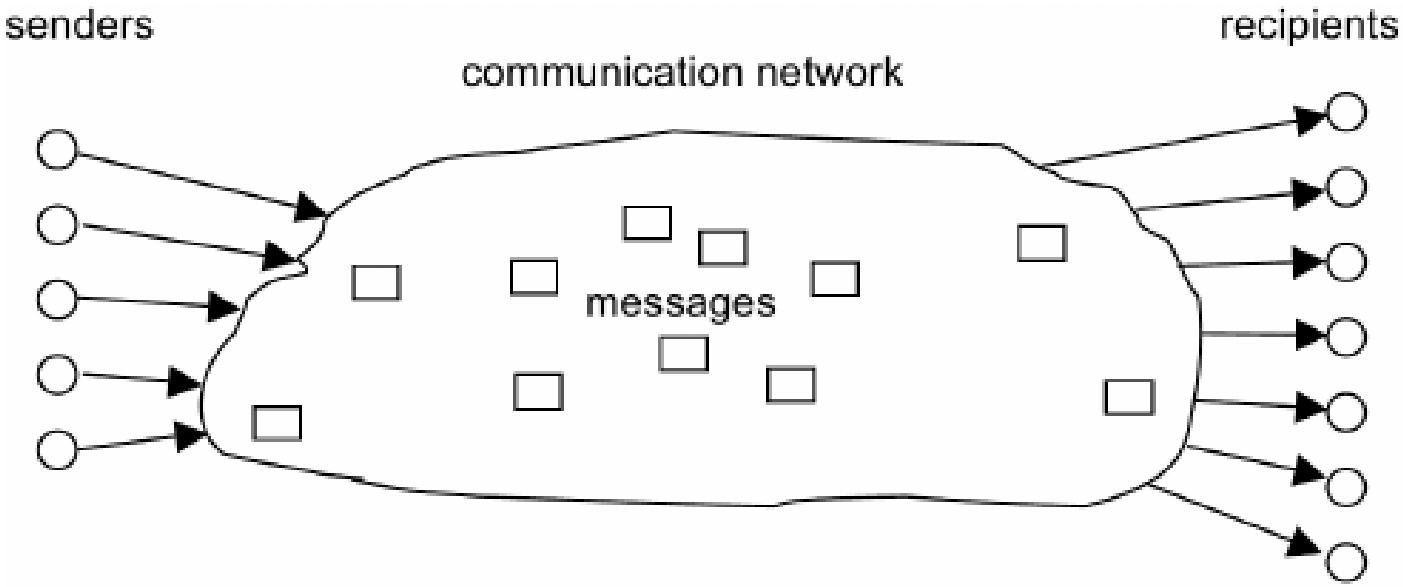
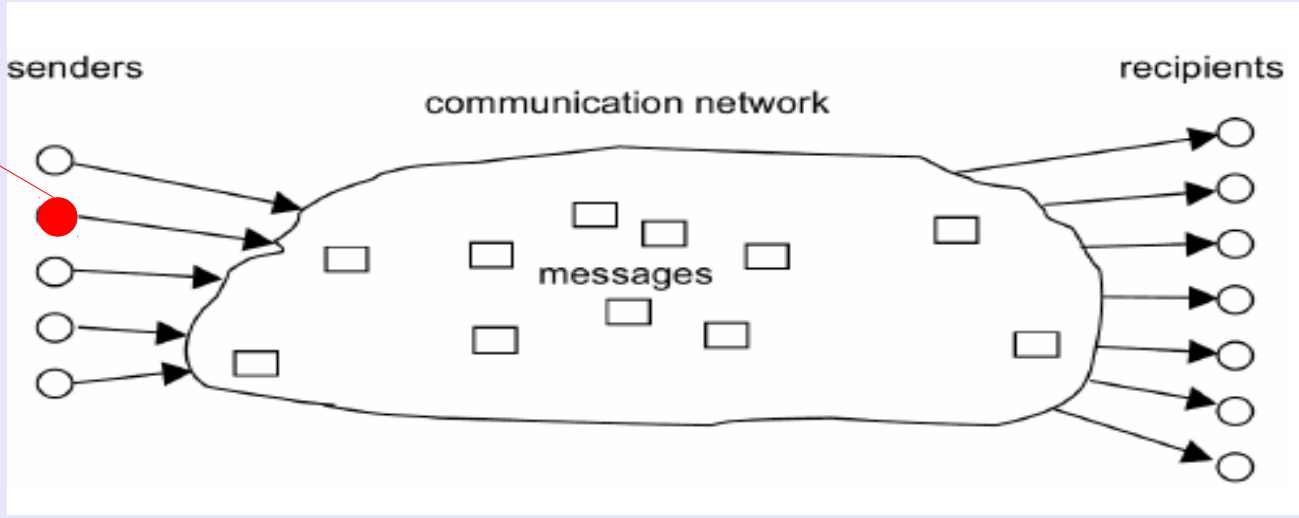


Image credit (before modification):
Christina Pöpper
Ruhr-University Bochum

Senders and recipients are thought of as rows in a database table

	MAC	Browser_fingerprint	IP	Sites_visited
SNDER_1	00:a0:ef:eb:5v:ff	af7f098c39728f8cb676e3df82ced01a149ee3aa92af2b88c20c4948a5fad5fd		torproject.org, ischool.arizona.edu, maps.google.com...
SNDER_2	00:c0:ff:dd:ff:ef	a5fad5fdd01a149eeaf7f098c39728f8cb676e3df82ce3aa92af2b88c20c4948		nytimes.com, purple.com



Question

- Suppose I include in a record a person's weight and height as 150 lbs, 5'3".
- Now suppose further that I do so for a database of male UA basketball players. Is the player anonymous?
- Where might you find combinations of attributes that are as rare as a short UA basketball player, but in other contexts?
  - Web: DNS, third-party tracking, browser fingerprinting

DNS

- You are required by law to say that DNS is “like a phone book” give it an easy-to-remember name get an IP address for the server hosting the website
- Forwarding DNS server
- Recursive DNS server (or resolver)
- (Root nameserver)
- (Top Level Domain nameserver)
- Authoritative nameserver

COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

DNS

```
$dig arizona.edu

; <<>> DiG 9.9.5-9+deb8u14-Debian <<>> arizona.edu
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 14058
;; flags: qr rd ra; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 4096
;; QUESTION SECTION:
;arizona.edu.                   IN     A

;; ANSWER SECTION:
arizona.edu.            6813IN     A      128.196.128.233

;; Query time: 32 msec
;; SERVER: 208.67.222.222#53(208.67.222.222)
;; WHEN: Mon Oct 23 12:43:03 MST 2017
;; MSG SIZE  rcvd: 56
```

DNS

- Suppose I use a VPN to tunnel my traffic to a server I control. What can you learn about me from my DNS requests?

DNS

- Suppose I use a VPN to tunnel my traffic to a server I control. What can you learn about me from my DNS requests?
- Many sites to do with local things in Tucson, AZ
- Sites to do with the University of Arizona
- Sites for groups with small memberships (Xerocraft)
- Many hits for a site with a public record attached to one person (sidiprojects.us)

COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

Browser fingerprinting

- UserAgent
- Language
- Color Depth
- Screen Resolution
- Timezone
- Has session storage or not
- Has local storage or not
- Has indexed DB
- Has IE specific 'AddBehavior'
- Has open DB
- CPU class
- Platform
- DoNotTrack or not
- Full list of installed fonts (maintaining their order, which increases the entropy), implemented with Flash.
- A list of installed fonts, detected with JS/CSS (side-channel technique) - can detect up to 500 installed fonts without flash
- Canvas fingerprinting
- WebGL fingerprintingPlugins (IE included)
- Is AdBlock installed or not
- Has the user tampered with its languages 1
- Has the user tampered with its screen resolution 1
- Has the user tampered with its OS 1
- Has the user tampered with its browser 1
- Touch screen detection and capabilities
- Pixel Ratio
- System's total number of logical processors available to the user agent.

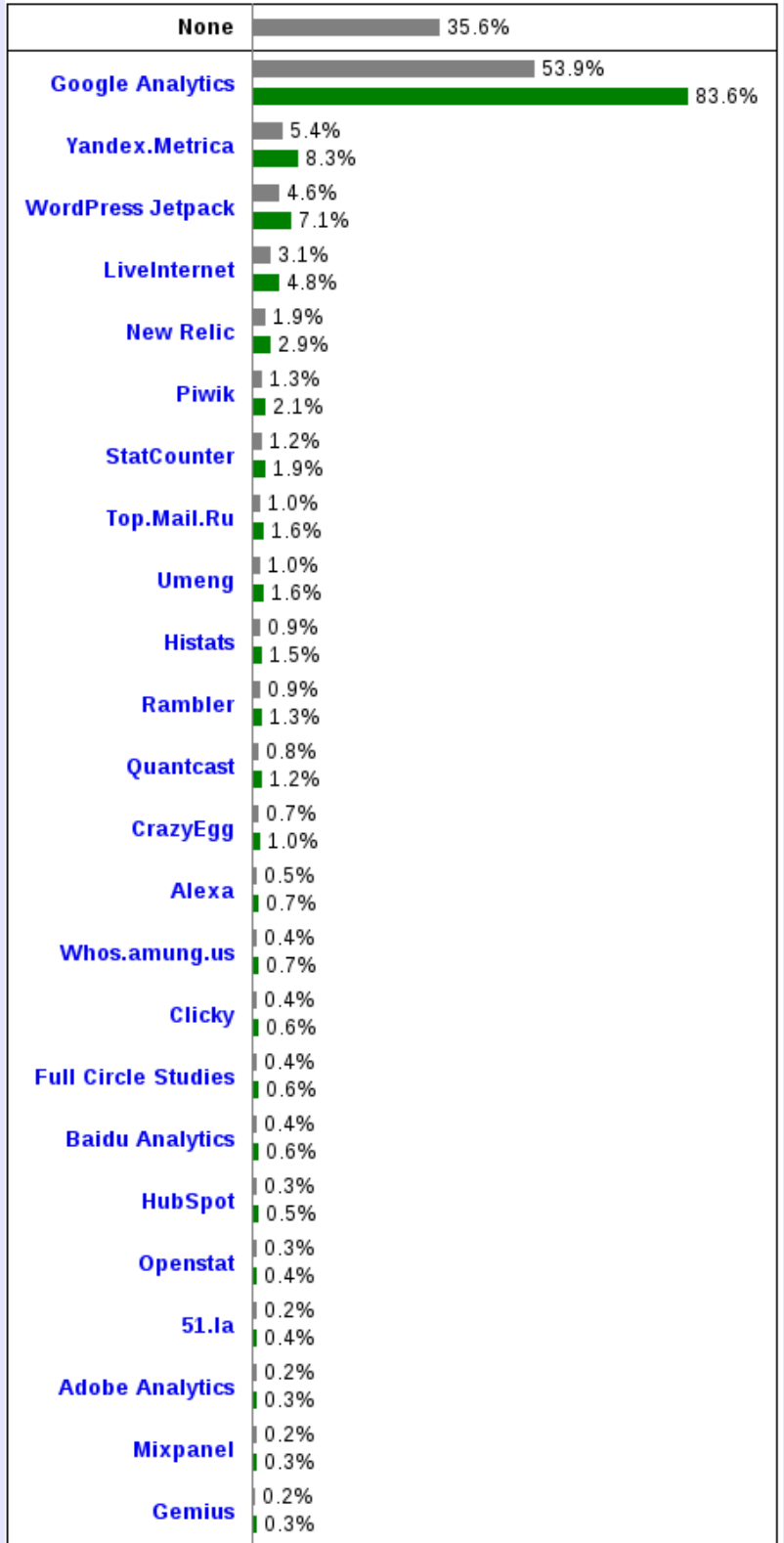
Browser fingerprinting

- Multi-monitor detection,
- Internal HashTable implementation detection
- WebRTC fingerprinting
- Math constants
- Accessibility fingerprinting
- Camera information
- DRM support
- Accelerometer support
- Virtual keyboards
- List of supported gestures (for touch-enabled devices)
- Pixel density
- Video and audio codecs availability
- Audio stack fingerprinting

Third-party analytics

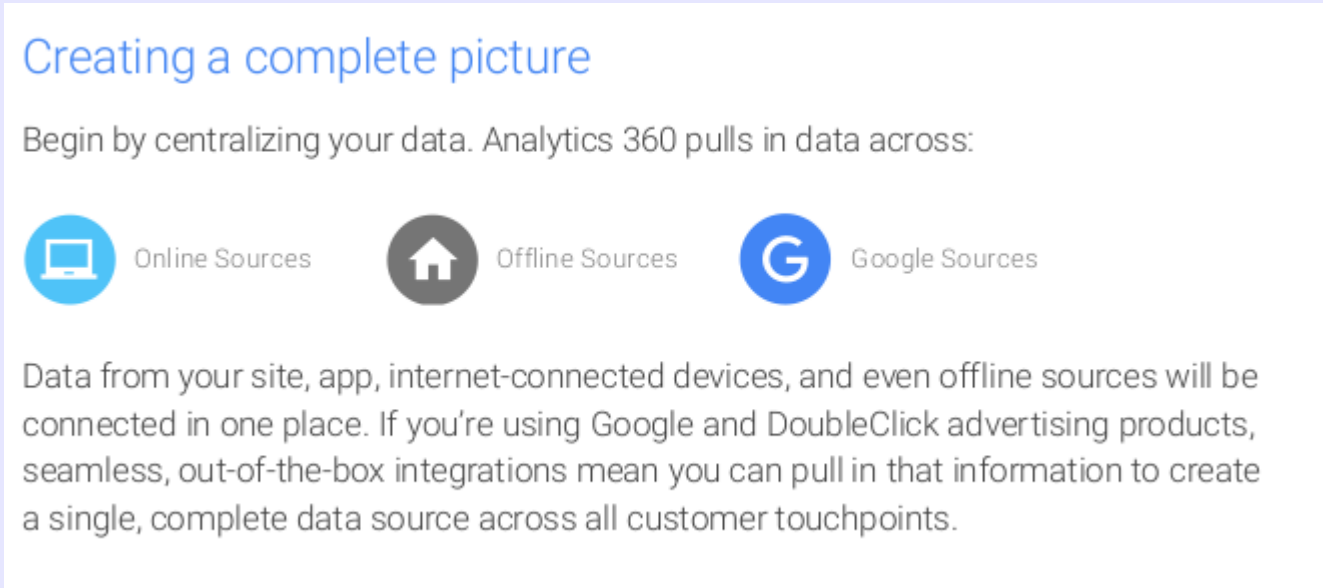
- 53.9% of all sites use Google Analytics

https://w3techs.com/technologies/overview/traffic_analysis/all



Third-party analytics

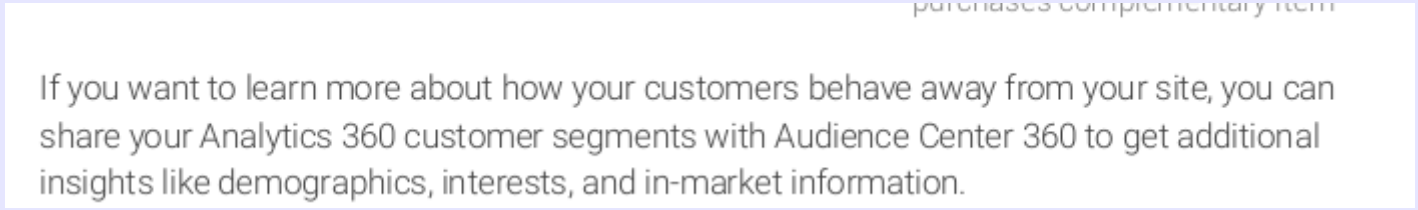
- 53.9% of all sites use Google Analytics



https://www.google.com/analytics/analytics/features/

Third-party analytics

- 53.9% of all sites use Google Analytics



https://www.google.com/analytics/analytics/features/

Discussion: “Identifiability”

- Why might it be too simple to say that for a sender S, every other potentially-different sender is either completely indistinguishable from S or not?
- 2 minutes alone, 2 minutes with a partner, then we’ll talk as a class

Question

- What is problematic about this definition of anonymity? “Anonymity is thus defined as the state of being not identifiable within a set of subjects, the anonymity set.” (Danezis and Diaz 3)

Question

- What is problematic about this definition of anonymity? “Anonymity is thus defined as the state of being not identifiable within a set of subjects, the anonymity set.” (Danezis and Diaz 3)
- We need to say something about an adversary and an attack model

- **Definition 1.2** From an adversary's perspective, anonymity of a subject s means that the adversary cannot achieve a certain *level of identification* for the subject s within the anonymity set. (Torra)

- Torra's terminology is confusing. His own idea of `n-Confusion' is in the background here :-).
- Simplifying, the point is: for each row in the database and some auxiliary information, we have a distribution over the subjects.
- The closer to uniform this distribution is, the "less identifiable" an entity is within the anonymity set, and the better the anonymization

Anonymity metrics

- Reviewing the discussion from a recent breakout session in a workshop on Privacy as Engineering Practice, Deirdre Mulligan said that there is a need for more formal measures of privacy (including anonymity)
- privacy loss in terms of information flow analysis, measures that take into account inference and not only disclosure, etc.

Anonymity metrics

- Degree of anonymity is a distribution 1 - p, where p is the probability assigned to the senders in the anonymity set (Reiter and Rubin 1998)
  - I talk about senders just for convenience, this applies more generally
- Think “worst case” -- who’s got the highest probability of being identified, and how high is that probability?
- This doesn’t account for how evenly distributed the probability is over the anonymity set, in the sense that it just depends on the greatest probability

Anonymity metrics

- Suppose a user u has 0.1 degree of anonymity. Consider two scenarios s1 and s2
  - s1: 2 users u and v, with v also 0.1 degree anonymity
  - s2: 1000 users, all users distinct from u with the same degree anonymity (which is less than 0.001)
- With degree of anonymity as measure, both have equal anonymity

Degree of anonymity

- One way to account for evenness of distribution is with Information Theory
- “Effective size”: Roughly, how many bits does the attacker need to identify a member of the anonymity set?

Degree of anonymity

- "Effective size": Roughly, how many bits does the attacker need to identify a member of the anonymity set?
- Less roughly, use the Shannon entropy. For Ψ the set of users, $\mathcal{U}$ the posterior of a user being the sender given a message,

Definition 2. *We define the effective size* $\mathcal{S}$ *of an* r *anonymity probability distribution* $\mathcal{U}$ *to be equal to the entropy of the distribution. In other words*

$$\mathcal{S} = -\sum_{u \in \Psi} p_u \log_2(p_u)$$

where $p_u = \mathcal{U}(u, r)$.

Degree of anonymity

- Less roughly, use the Shannon entropy.
  - This gives an expected value
- Danezis and Diaz also mention defining degree of anonymity as $\log_2(N)$, for N the number of users
- Can also use min entropy for worst case
- They don't mention it, but all these are related

Degree of anonymity

- $H_0 = \log |X|$
- $H_1 = -\sum_{i=1}^{n} p_i \log(p_i)$
- $H_\infty = -\log(\max_i p_i)$

{Sender, Receiver, Relationship} Anonymity

- Sender/receiver anonymity: For a given message m, what is the probability that m came from sender/receiver A?
  - Note this depends on who the adversary is: the recipient? A global passive adversary? An active adversary? ...
- Relationship anonymity: For a given message m, what is the probability that m came from sender A and went to destination B?
- Relationship* anonymity: What is the probability that sender A is communicating with destination B?
  - a persistent relationship, not a single message or request exchange
- Question: how can you have sender anonymity but not relationship* anonymity? Hint: a universal generalization implies all instantiations.

“Suppose the network provides perfect sender anonymity, i.e., any message exiting the network is equally likely to have originated from any active sender. By observing these messages, however, the attacker can easily infer that all of them have the same destination. For every active sender, the attacker can thus determine with 100% certainty that this sender is communicating with the website, completely breaking relationship anonymity.”

Shmatikov and Wang, ‘Measuring Relationship Anonymity in Mix Networks’

COLLEGE OF SOCIAL & BEHAVIORAL SCIENCES
School of Information

Anonymity networks

Anonymity networks address the traffic analysis problem

- Chaum: “Keeping confidential who converses with whom, and when they converse”
- Contrast with secrecy of message content

Anonymity networks can involve trusted or semi-trusted relays

- Trusted parties are not adversaries: they can break anonymity
- Semi-trusted parties don’t all collude

Trusted relays

- Example: Nym servers
  - a server keeps a dictionary between real and pseudonymous emails
  - request comes to the remailer, which forwards it, gets the response, and returns it to the user
  - Example: anon.penet.fi
- Other Examples: Anonymous proxies (startpage.com), VPNs

Trusted relays

- Problem: messages are all linked
  - Stylometric attacks: the frequency of function words in the English language can be used in the long term to identify users (Rao & Rohatgi (2000), “Can Pseudonymity Really Guarantee Privacy?”)
  - Correspondent sets of each nym
- Anonymity is compromised if one node is compromised. (“Single point of failure.”)
  - lots of incentive to coerce
  - or if the node is not honest
- Fails bitwise indistinguishability: sometimes traffic analysis can deanonymize
  - http proxy example
  - timing correlation

Semi-trusted relays

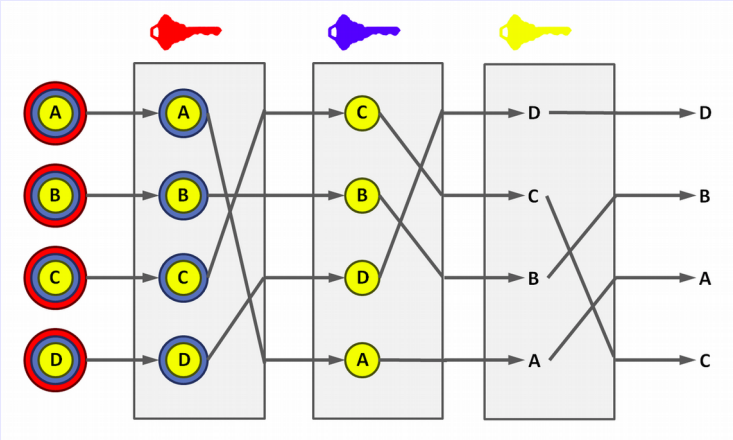
- Strengths
  - Compromise of more than one is needed, so more coercion resistant than trusted-relay approaches
  - “any single mix is able to provide the secrecy of the correspondence between the input and the outputs of the entire cascade” (Chaum)
- Weaknesses
  - Tagging attacks violate unlinkability (blind signing attack)
  - replay attacks
  - slow (public-key cryptography)

Semi-trusted relays

- What are the problems with a mixnet with only one node? (Chaum)

Mixnets are anonymity systems with semi-trusted relays

- Routing protocol with a cascade of cryptographic relays called ‘mixes’
- Mixes only know their neighbors
- User-specifiable routing (Chaum’s “new kind of mix”)



wikipedia.org

Mix-nets are anonymity systems with semi-trusted relays

- Suppose we are at a mix A1, which receives message m.
- m is split into a fixed number blocks, ℓ

$$A_1\colon [K_{A_1}(R_{A_1}, A_2)], [R_{A_1}^{-1}(K_{A_2}(R_{A_2}, A_3))], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)] \rightarrow.$$

$$A_2\colon [K_{A_2}(R_{A_2}, A_3)], [R_{A_2}^{-1}(K_{A_3}(R_{A_3}, A_4))], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots)], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)], [R_{A_1}(J_{A_1})] \rightarrow,$$

$$A\colon [M_1], [M_2], \ldots, [M_{l-n}],$$
$$[R_{A_n}(R_{A_{n-1}} \cdots R_{A_1}(J_{A_1}) \cdots)], \ldots, [R_{A_n}(J_{A_n})].$$

Mix-nets are anonymity systems with semi-trusted relays

- The first block is like a header: it contains the key R_{A1} and address A2 for the next hop. This is stripped off of the message, and a padding ("junk") block is added to the end

$$A_1\colon [K_{A_1}(R_{A_1}, A_2)], [R_{A_1}^{-1}(K_{A_2}(R_{A_2}, A_3))], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)] \rightarrow.$$

$$A_2\colon [K_{A_2}(R_{A_2}, A_3)], [R_{A_2}^{-1}(K_{A_3}(R_{A_3}, A_4))], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots)], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)], [R_{A_1}(J_{A_1})] \rightarrow,$$

$$A\colon [M_1], [M_2], \ldots, [M_{l-n}],$$
$$[R_{A_n}(R_{A_{n-1}} \cdots R_{A_1}(J_{A_1}) \cdots)], \ldots, [R_{A_n}(J_{A_n})].$$

-

Mix-nets are anonymity systems with semi-trusted relays

- The rest of the blocks are, first, the header blocks for all remaining routers in the cascade, and next, the message. All of these are encoded using .

$$A_1\colon [K_{A_1}(R_{A_1}, A_2)], [R_{A_1}^{-1}(K_{A_2}(R_{A_2}, A_3))], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)] \rightarrow.$$

$$A_2\colon [K_{A_2}(R_{A_2}, A_3)], [R_{A_2}^{-1}(K_{A_3}(R_{A_3}, A_4))], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots)], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)], [R_{A_1}(J_{A_1})] \rightarrow,$$

$$A\colon [M_1], [M_2], \ldots, [M_{l-n}],$$
$$[R_{A_n}(R_{A_{n-1}} \cdots R_{A_1}(J_{A_1}) \cdots)], \ldots, [R_{A_n}(J_{A_n})].$$

Mix-nets are anonymity systems with semi-trusted relays

- A1 uses the R_{A1} it now has to decode the (ℓ-1) blocks after the header in the original message: these are the first part of the message sent out from A1, they contain the headers for A2, the encoded headers for A3,...An, and then the encoded message
- The blocks are passed to the next node, which could be another mix

$$A_1\colon [K_{A_1}(R_{A_1}, A_2)], [R_{A_1}^{-1}(K_{A_2}(R_{A_2}, A_3))], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)] \rightarrow.$$

$$A_2\colon [K_{A_2}(R_{A_2}, A_3)], [R_{A_2}^{-1}(K_{A_3}(R_{A_3}, A_4))], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots)], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)], [R_{A_1}(J_{A_1})] \rightarrow,$$

$$A\colon [M_1], [M_2], \ldots, [M_{l-n}],$$
$$[R_{A_n}(R_{A_{n-1}} \cdots R_{A_1}(J_{A_1}) \cdots)], \ldots, [R_{A_n}(J_{A_n})].$$

Mix-nets are anonymity systems with semi-trusted relays

- Mixes only know their neighbors. (Question: Why?)
- All nodes have a public key
- Weaknesses
  - active attacks: tagging attacks (blind signing attack), replay attacks
  - slow (public-key cryptography, latency in anonymous remailers).

$$A_1: [K_{A_1}(R_{A_1}, A_2)], [R_{A_1}^{-1}(K_{A_2}(R_{A_2}, A_3))], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots], \ldots,$$
$$[R_{A_1}^{-1}(R_{A_2}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)] \rightarrow.$$

$$A_2: [K_{A_2}(R_{A_2}, A_3)], [R_{A_2}^{-1}(K_{A_3}(R_{A_3}, A_4))], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_{n-1}}^{-1}(K_{A_n}(R_{A_n}, A)) \cdots)],$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_1) \cdots)], \ldots,$$
$$[R_{A_2}^{-1}(R_{A_3}^{-1} \cdots R_{A_n}^{-1}(M_{l-n}) \cdots)], [R_{A_1}(J_{A_1})] \rightarrow,$$

$$A: [M_1], [M_2], \ldots, [M_{l-n}],$$
$$[R_{A_n}(R_{A_{n-1}} \cdots R_{A_1}(J_{A_1}) \cdots)], \ldots, [R_{A_n}(J_{A_n})].$$

Mixing techniques for Mixnets

- Cascading: All nodes are always used, in the same order
- Scalability is a problem, requires setting up a fixed route with all nodes
- Only requires one honest node to preserve anonymity

Mixing techniques for Mixnets

- User specified: user arbitrarily picks its route through the network
- Scalable, does not require initial configuration of a route
- Not anonymous if only one node is honest (nodes can figure out their positions)