Design and Implementation of Privacy-Preserving Surveillance

Aaron Segal

Yale University

May 11, 2016

Advisor: Joan Feigenbaum

Overview

- Introduction – Surveillance and Privacy
- Privacy Principles for **Open** Surveillance Processes
- Lawful Set Intersection and the High Country Bandits
- Contact Chaining
- Anonymity through Tor and Verdict

The Problem

- Open season on private personal data
- No accountability
- No guarantees
- The government is part of the problem

Motivation & Goals

Replace law enforcement's secretive, unprincipled treatment of citizens' big data with an open privacy policy.

- **Secret** processes for data collection
- Public is asked to **trust** the government
- Presumed **tradeoff** between *national security* and *personal privacy*
- Ideal world: **No surveillance**

Motivation & Goals

Replace law enforcement's secretive, unprincipled treatment of citizens' big data with an open privacy policy.

- **Secret** processes for data collection
- Public is asked to **trust** the government
- Presumed **tradeoff** between *national security* and *personal privacy*
- Ideal world: **No surveillance**
  - Realistic goal: **Surveillance with privacy preservation**

Motivation & Goals

Replace law enforcement's secretive, unprincipled treatment of citizens' big data with an open privacy policy.

- **Secret** processes for data collection

- Public is asked to **trust** the government

- Presumed **tradeoff** between *national security* and *personal privacy*
  - **No need** to abandon *personal privacy* to ensure *national security*
- Ideal world: **No surveillance**
  - Realistic goal: **Surveillance with privacy preservation**

Motivation & Goals

Replace law enforcement's secretive, unprincipled treatment of citizens' big data with an open privacy policy.

- **Secret** processes for data collection

- Public is asked to **trust** the government
  - Accountability guaranteed by existing **cryptographic technology**
- Presumed **tradeoff** between *national security* and *personal privacy*
  - **No need** to abandon *personal privacy* to ensure *national security*
- Ideal world: **No surveillance**
  - Realistic goal: **Surveillance with privacy preservation**

Motivation & Goals

Replace law enforcement's secretive, unprincipled treatment of citizens' big data with an open privacy policy.

- **Secret** processes for data collection
  - **Open** processes for data collection with a principled privacy policy
- Public is asked to **trust** the government
  - Accountability guaranteed by existing **cryptographic technology**
- Presumed **tradeoff** between *national security* and *personal privacy*
  - **No need** to abandon *personal privacy* to ensure *national security*
- Ideal world: **No surveillance**
  - Realistic goal: **Surveillance with privacy preservation**

Some Privacy Principles for Lawful Surveillance (1)

Open processes

- **Must** follow rules and procedures of public law
- **Need not** disclose targets and details of investigations

Two types of users:

- *Targeted users*
  - Under **suspicion**
  - Subject of a **warrant**
  - Can be *known* or *unknown*
- *Untargeted users*
  - No probable cause
  - Not targets of investigation
  - The vast majority of internet users

Some Privacy Principles for Lawful Surveillance (2)

- Distributed trust
  - No one agency can compromise privacy.
- Enforced scope limiting
  - No overly broad group of users' data are captured.
- Sealing time and notification
  - After a finite, reasonable time, surveilled users are notified.
- Accountability
  - Surveillance statistics are maintained and audited.

Case Study – High Country Bandits

2010 case – string of bank robberies in Arizona, Colorado

FBI Intersection attack compared 3 cell tower dumps totaling 150,000 users

- 1 number found in all 3 cell dumps – led to arrest
- 149,999 innocent users' information acquired

Intersecting Cell-Tower Dumps

- Law enforcement goal: Find *targeted*, *unknown* user whose phone number will appear in the intersection of cell-tower dumps
- Used in: High Country Bandits case, CO-TRAVELER program
  - Same principle for any collection of metadata

Cell Tower A Time t_1	Cell Tower B Time t_2	Cell Tower C Time t_3
• 650-555-4430	• 650-555-3222	• 650-555-7928
• 650-555-3435	• 650-555-3813	• 650-555-0599
• 650-555-2840	• 650-555-2786	• 650-555-6445
• 650-555-7691	• 650-555-7976	• 650-555-7511
• 650-555-1505	• 650-555-0392	• 650-555-2277
• 650-555-9589	• 650-555-5872	• 650-555-7976
• 650-555-7976	• 650-555-4891	• 650-555-2840
• 650-555-9266	• 650-555-9709	• 650-555-3222

Intersecting Cell-Tower Dumps

- Law enforcement goal: Find *targeted*, *unknown* user whose phone number will appear in the intersection of cell-tower dumps
- Used in: High Country Bandits case, CO-TRAVELER program
  - Same principle for any collection of metadata

Cell Tower A Time t_1	Cell Tower B Time t_2	Cell Tower C Time t_3
• 650-555-4430	• 650-555-3222	• 650-555-7928
• 650-555-3435	• 650-555-3813	• 650-555-0599
• 650-555-2840	• 650-555-2786	• 650-555-6445
• 650-555-7691	• **650-555-7976**	• 650-555-7511
• 650-555-1505	• 650-555-0392	• 650-555-2277
• 650-555-9589	• 650-555-5872	• **650-555-7976**
• **650-555-7976**	• 650-555-4891	• 650-555-2840
• 650-555-9266	• 650-555-9709	• 650-555-3222

Privacy-Preserving Solution [SFF, FOCI'14]

- A *private set intersection protocol* built to satisfy surveillance privacy principles (based on Vaidya-Clifton '05)

- Catching Bandits and *Only* Bandits: Privacy-Preserving Intersection Warrants for Lawful Surveillance
  - Presented at the 4th USENIX Workshop on Free and Open Communications on the Internet (FOCI '14)

Privacy-Preserving Cryptography

Probabilistic ElGamal encryption for secure storage of cell-tower records.

- Same records encrypt to **different** random-looking byte strings

```
a = "650-555-2840"
b = "650-555-2840"
print ElGamalEncrypt(a)
    > 0x00d07e08ec44712b
print ElGamalEncrypt(b)
    > 0x58c82a7f050f9683
```

Deterministic Pohlig-Hellman encryption for temporary, per-execution blinding of those records.

- Same records encrypt to **identical** random-looking byte strings

```
a = "650-555-2840"
b = "650-555-2840"
print PohligHellmanEncrypt(a)
    > 0x0cb508480f207ec5
print PohligHellmanEncrypt(b)
    > 0x0cb508480f207ec5
```

Private Set Intersection Setup

- ElGamal encryption and Pohlig-Hellman encryption are ***mutually commutative*** with one another

$$D_2(D_3(D_1(E_3(E_2(E_1(x)))))) = x$$

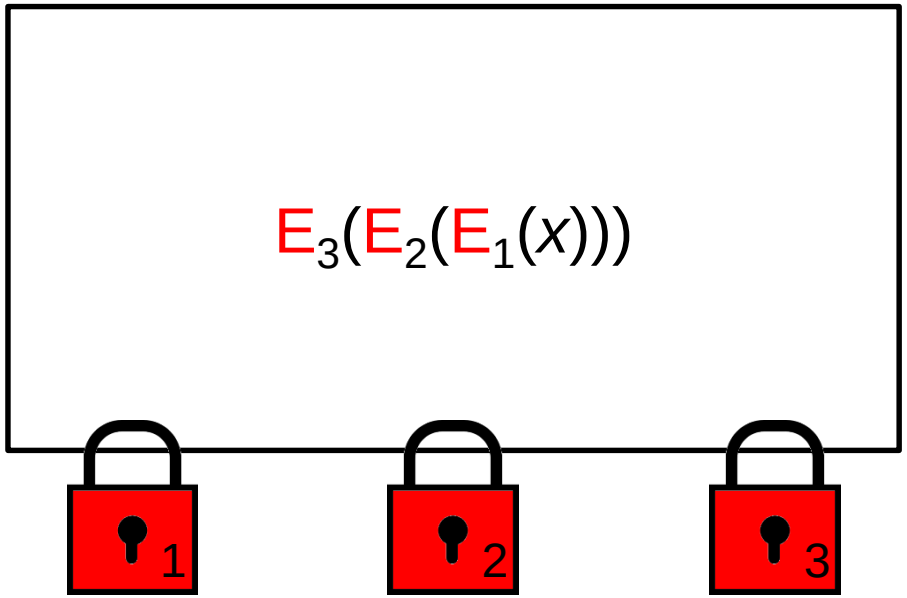
$$D_3(D_2(E_3(D_1(E_2(E_1(x)))))) = x$$

- Relies on **multiple, independent agencies** to execute protocol, providing distributed trust and accountability, e.g.:
  - Executive agency (FBI, NSA)
  - Judicial agency (warrant-issuing court)
  - Legislative agency (oversight committee established by law)

- Each agency must participate at each step or else no one can decrypt!

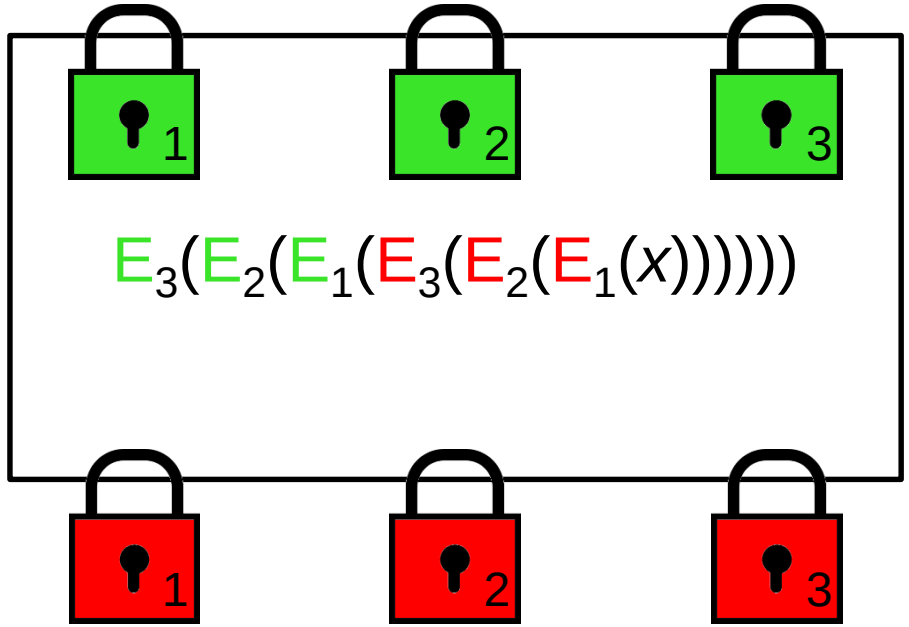
Private Set Intersection Protocol (Step 1)

- Repository serves data encrypted with ElGamal encryption
  - Uses agencies' long-term public (encryption) keys



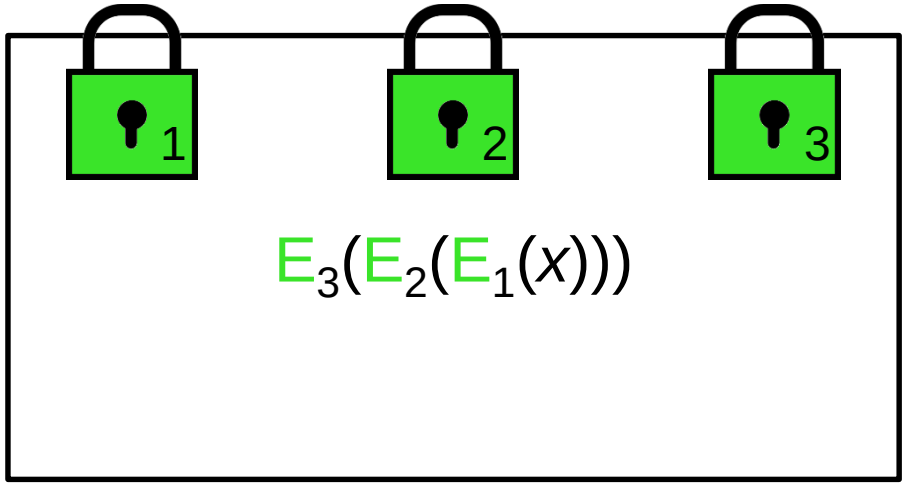
Private Set Intersection Protocol (Step 1)

- Repository serves data encrypted with ElGamal encryption
  - Uses agencies' long-term public (encryption) keys
- Agencies encrypt the encryptions with Pohlig-Hellman encryption
  - Uses agencies' ephemeral encryption keys



Private Set Intersection Protocol (Step 1)

- Repository serves data encrypted with ElGamal encryption
  - Uses agencies' long-term public (encryption) keys
- Agencies encrypt the encryptions with Pohlig-Hellman encryption
  - Uses agencies' ephemeral encryption keys
- Agencies decrypt the encrypted encryptions with ElGamal decryption
  - Uses agencies' long-term private (decryption) keys
- Can now inspect data, which is encrypted under Pohlig-Hellman



Private Set Intersection Protocol (Step 2)

- Accomplished: Moved from an ElGamal state to a Pohlig-Hellman state without ever fully decrypting the private data!
- Agencies can now inspect encrypted data to find matching records
- Last step: decrypt *only* those records with Pohlig-Hellman

```
a = "650-555-2840"
b = "650-555-2840"
print ElGamalEncrypt(a)
    > 0x00d07e08ec44712b
print ElGamalEncrypt(b)
    > 0x58c82a7f050f9683
```

```
a = "650-555-2840"
b = "650-555-2840"
print PohligHellmanEncrypt(a)
    > 0x0cb508480f207ec5
print PohligHellmanEncrypt(b)
    > 0x0cb508480f207ec5
```

Protocol Satisfies Privacy Principles

- Open Process
  - Can openly standardize the protocol and the crypto *without* compromising investigative power
- Distributed trust
  - No one agency can decrypt or perform intersection.
- Enforced scope limiting
  - Any agency can stop an execution if sets or intersection are too large.
- Sealing time and notification
  - Implementable by policy – all agencies get final data set
- Accountability
  - Because every agency must participate, no agency can perform illegitimate surveillance without the other agencies' learning and getting statistics.

Evaluation of Implementation

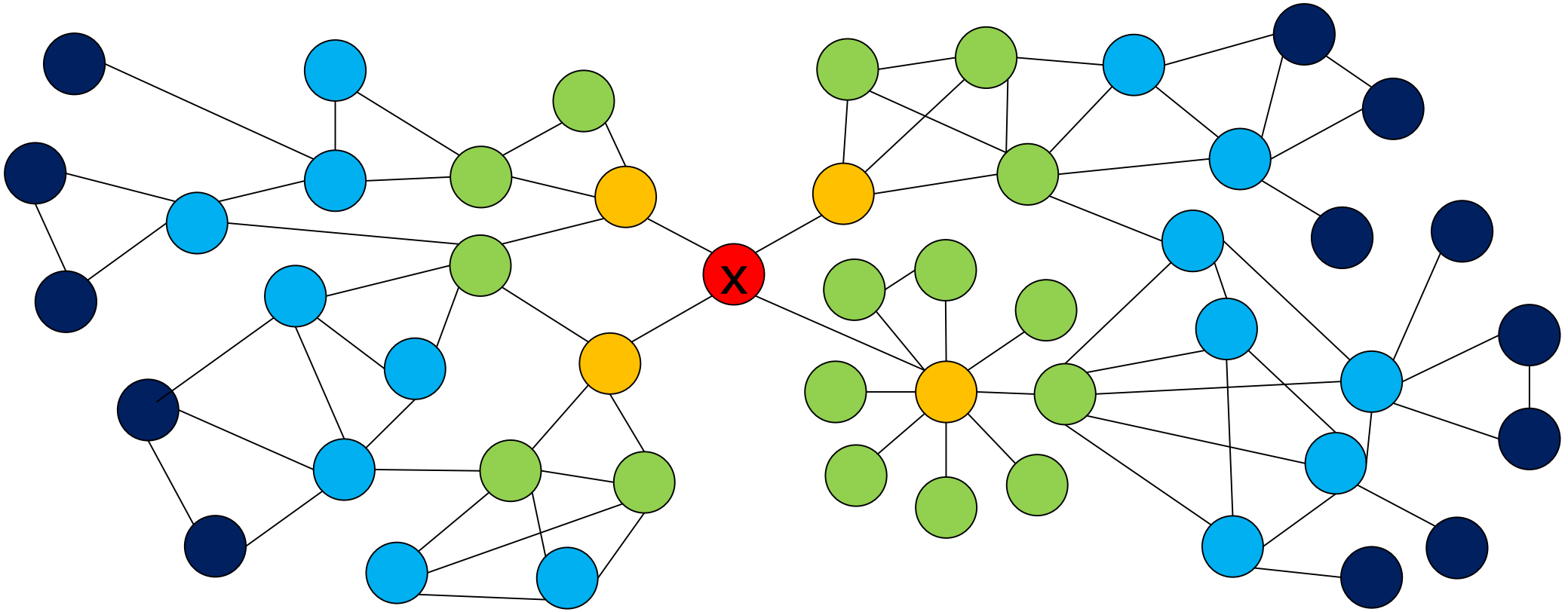
- Java implementation of protocol run in parallel on Yale CS Cloud
- High Country Bandits example with 50,000 items per set takes less than 11 minutes to complete.
- Note that this is an *offline* process.

Items	Data sent per node (KB)	End-to-End runtime (s)
10	21	1.0
25	46	1.1
50	86	1.3
75	127	1.6
100	167	1.7
250	410	2.9
500	815	4.9
750	1220	6.8
1000	1625	8.2
2500	4055	18.5
5000	8106	36.7
7500	12156	53.6
10000	16206	71.8
25000	40507	229.4
50000	81009	629.4

Table 1: Experimental Results

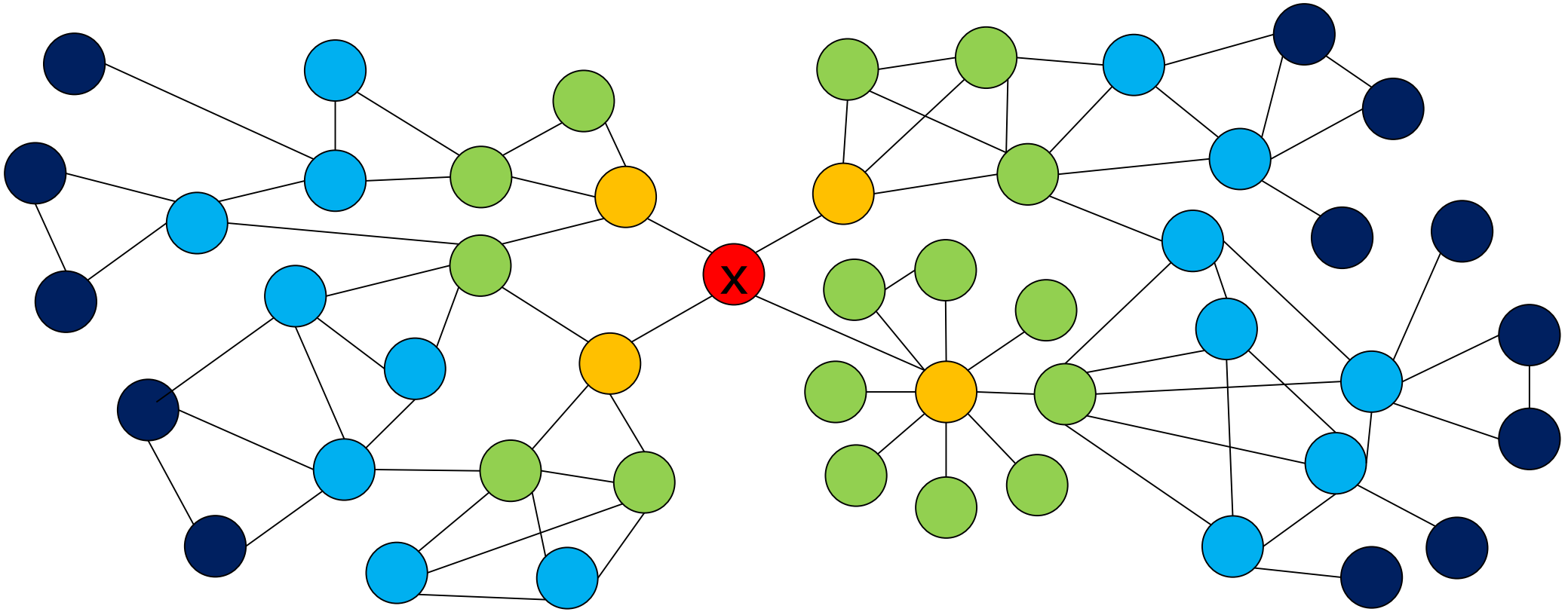
Contact Chaining

- Government knows phone number of target X.
- Goal: Consider the “k-contacts” of X (nodes within distance k).



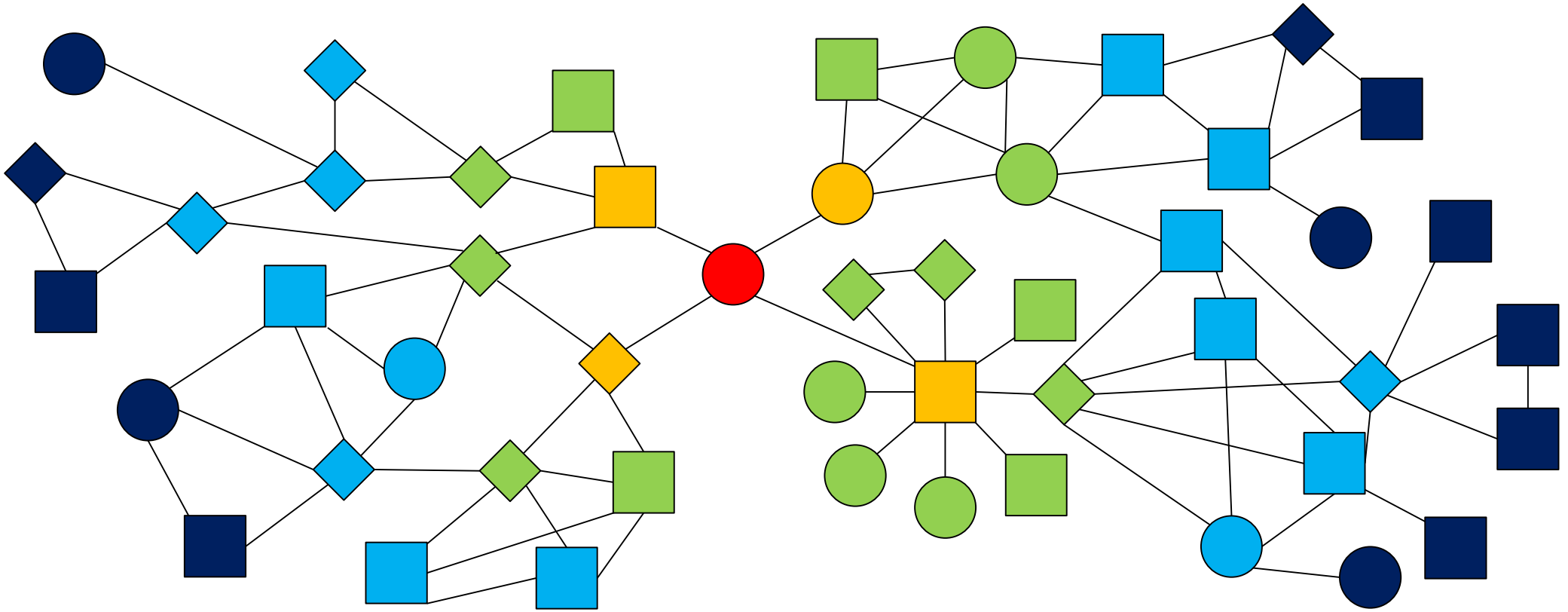
Privacy-Preserving Contact Chaining Goals

- Government learns actionable, relevant intelligence
- Telecommunications companies learn nothing more about other companies' clients



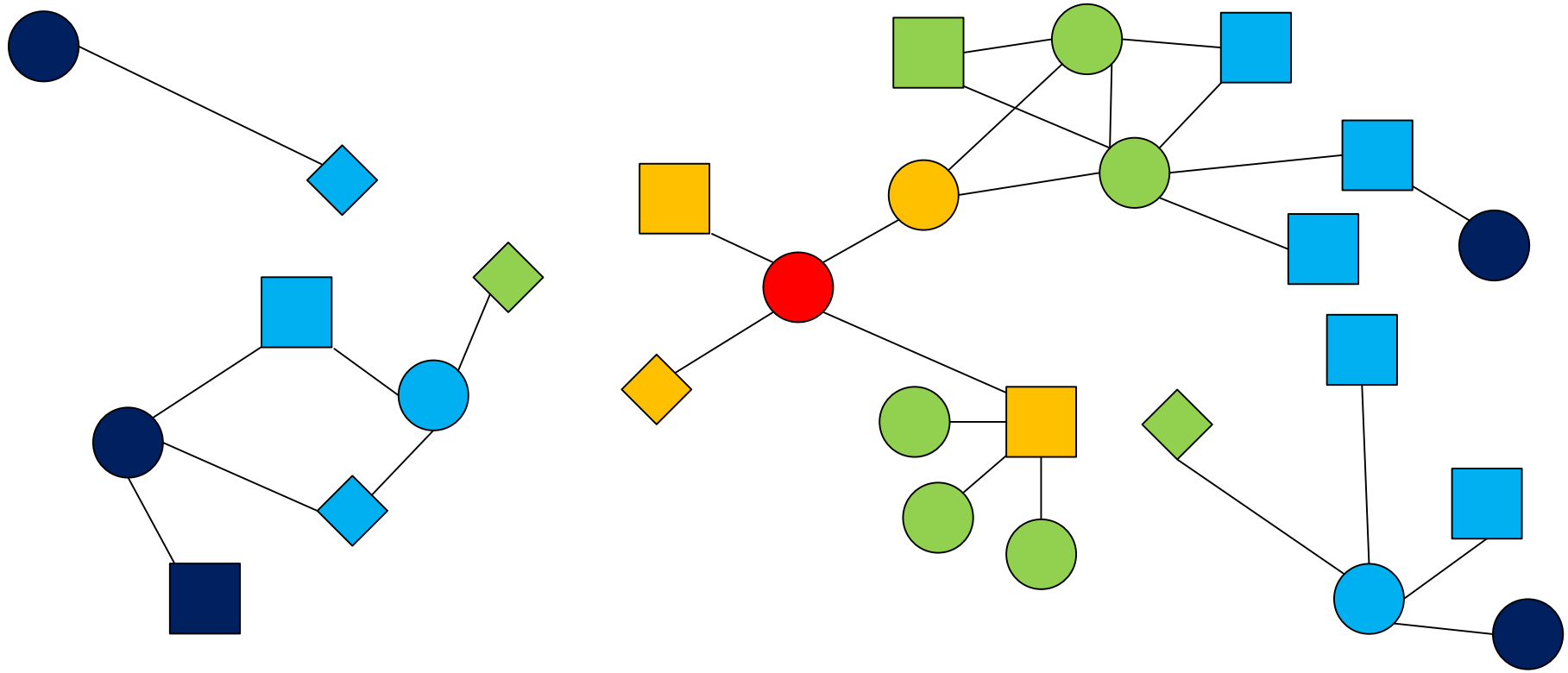
Privacy-Preserving Contact Chaining Goals

- Government learns actionable, relevant intelligence
- Telecommunications companies learn nothing more about other companies' clients



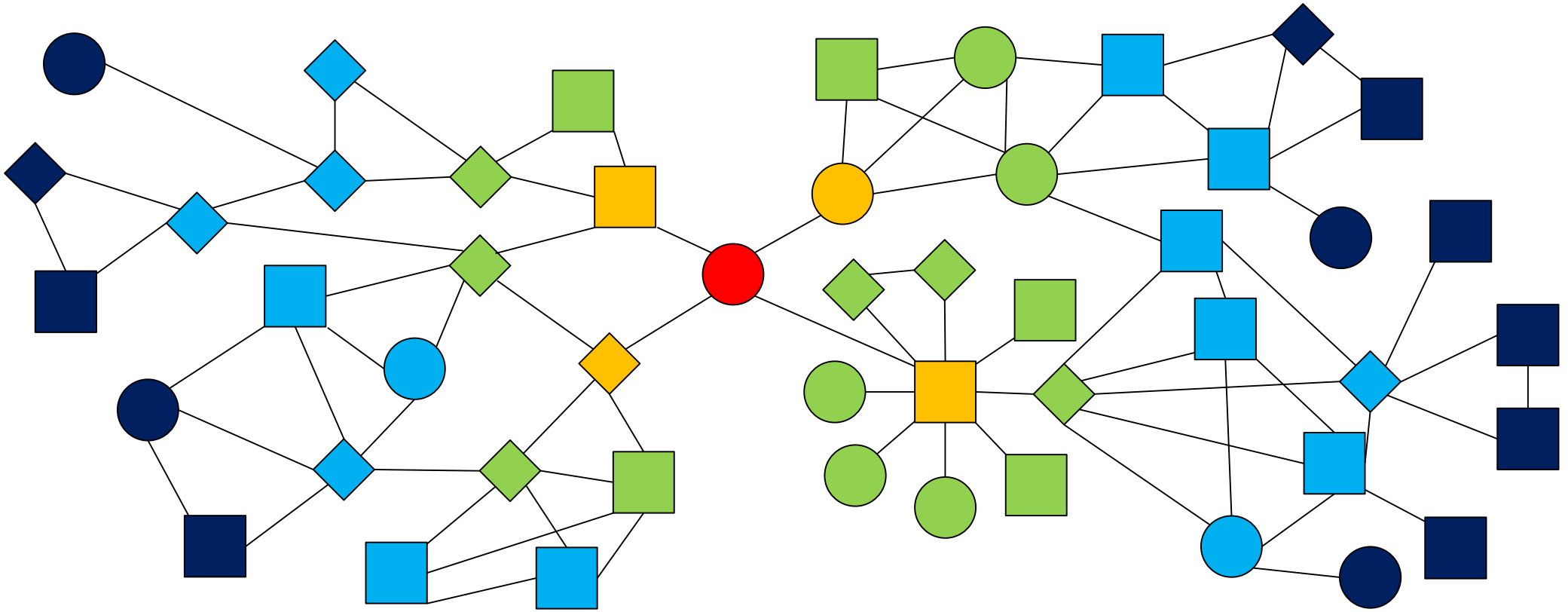
Privacy-Preserving Contact Chaining Goals

- Government learns actionable, relevant intelligence
- Telecommunications companies learn nothing more about other companies' clients



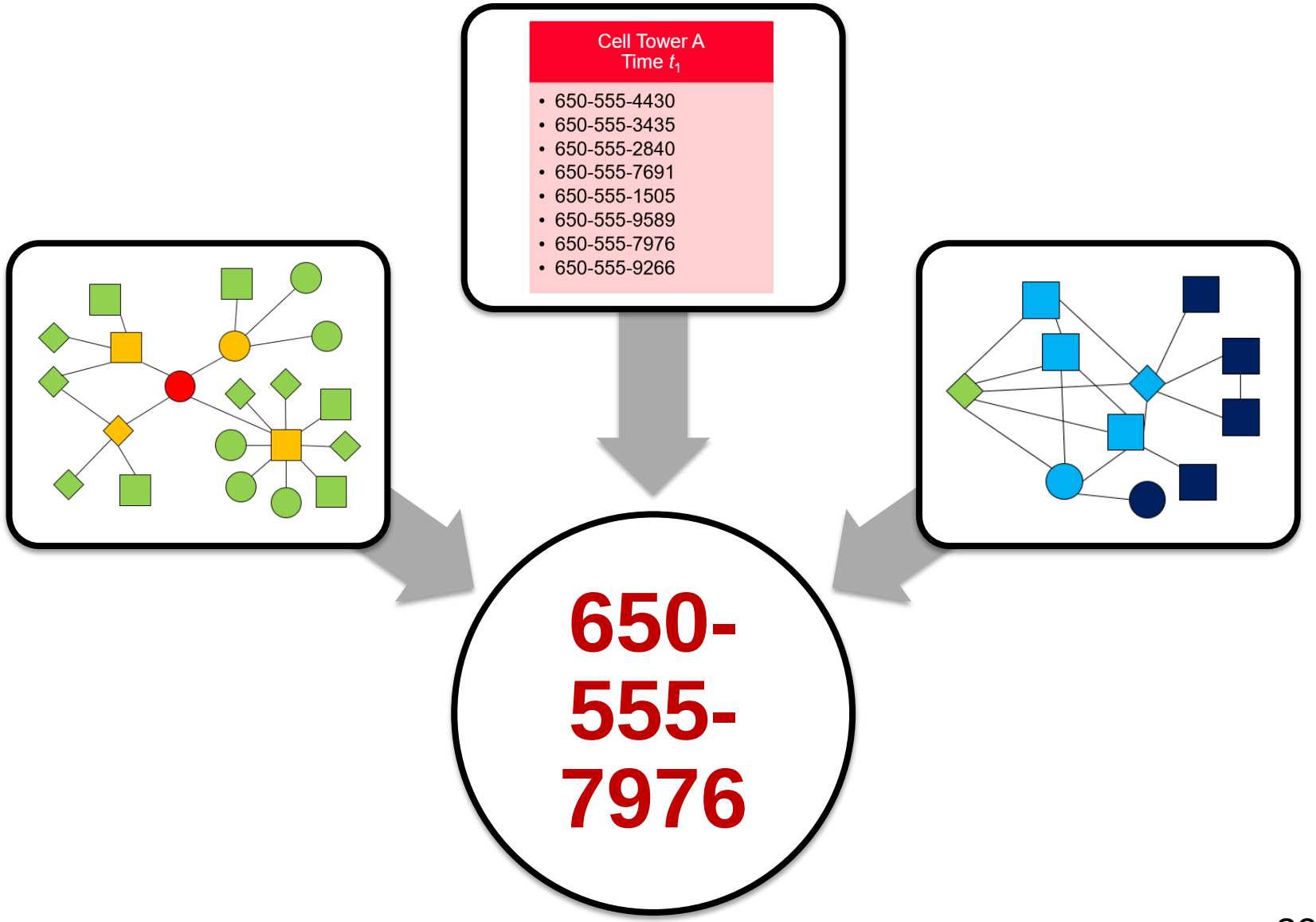
Restrictions on Contact Chaining

- Respect the distinction between targeted and untargeted users
- Enforce scope limiting
- Enforce division of trust between authorities



Using Contact Chaining - Main Idea

- Use privacy-preserving contact chaining protocol to get encryptions of *k*-contacts of target
- Use privacy-preserving set intersection to filter *k*-contacts and decrypt only new targets



Privacy-Preserving Contact Chaining Protocol

- Government agencies agree on a warrant:
  - Initial target id *X*
  - Maximum chaining length *k*
  - Scope-limiting parameter *d* : Maximum degree

- Each telecom has:
  - List of client identities served
  - Contact list for each client

- Agencies repeatedly query telecoms about their data

Privacy-Preserving Contact Chaining Protocol Setup

- Agencies perform a modified parallel breadth-first search by querying telecoms
- $\text{Enc}_{T(a)}(a)$ is a public-key encryption of a under the encryption key of $T(a)$, the telecom that serves user a
- $\text{Enc}_{Agencies}(a)$ is an ElGamal encryption of a under the keys of all agencies

Query to $T(a)$

- $\text{Enc}_{T(a)}(a)$
- Signatures from all agencies

Response from $T(a)$

- $\text{Enc}_{Agencies}(a)$
- $\text{Enc}_{T(b)}(b)$ for all b in a's set of neighbors
- Signature from $T(a)$

Privacy-Preserving Contact Chaining Protocol

- Step 0:
  - Query $T(x)$ on original target x
- Step 1 through k:
  - Query appropriate telecom on all ciphertexts received during previous step
  - Exception: If a single response has more than d contacts, do not query them
- Output: Agency ciphertexts received

Query to $T(a)$

- $\text{Enc}_{T(a)}(a)$
- Signatures from all agencies

Response from $T(a)$

- $\text{Enc}_{Agencies}(a)$
- $\text{Enc}_{T(b)}(b)$ for all b in a's set of neighbors
- Signature from $T(a)$

Protecting Private Data

- Agencies see *no* cleartext identities from this contact chaining protocol
- Telecoms learn no information about other telecoms' users by responding to queries
- Signatures ensure validity of all messages

Query to $T(a)$

- $\text{Enc}_{T(a)}(a)$
- Signatures from all agencies

Response from $T(a)$

- $\text{Enc}_{Agencies}(a)$
- $\text{Enc}_{T(b)}(b)$ for all b in a's set of neighbors
- Signature from $T(a)$

Protocol Satisfies Privacy Principles

- Open Process
  - Can openly standardize the protocol and the crypto *without* compromising investigative power
- Distributed trust
  - Telecoms disregard queries unless signed by all agencies
  - No one agency can decrypt responses
- Enforced scope limiting
  - Any agency can block paths through high-degree vertices
- Sealing time and notification
  - Agencies can notify targeted users after intersection step
- Accountability
  - Surveillance statistics collected by any or all agencies

Contact Chaining Experimental Setup

- Java implementation of protocol run in parallel on Yale CS Cloud
- Used actual network data from a Slovakian social network as “realistic” stand-in for a telephone network

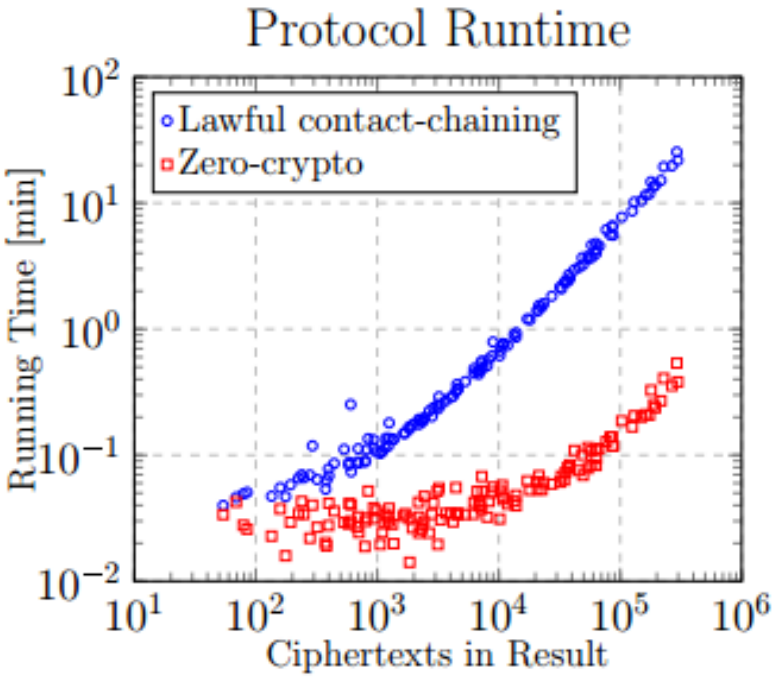
Ciphertexts in result	Degree of Target *x*	Maximum Path Length *k*	Large Vertex Degree Cutoff ***d***
582	40	2	50
1061	47	2	75
5301	128	2	150
10188	123	2	500
27338	32	3	200
49446	40	3	150
102899	230	3	100
149535	159	3	150
194231	128	3	500
297474	123	3	500

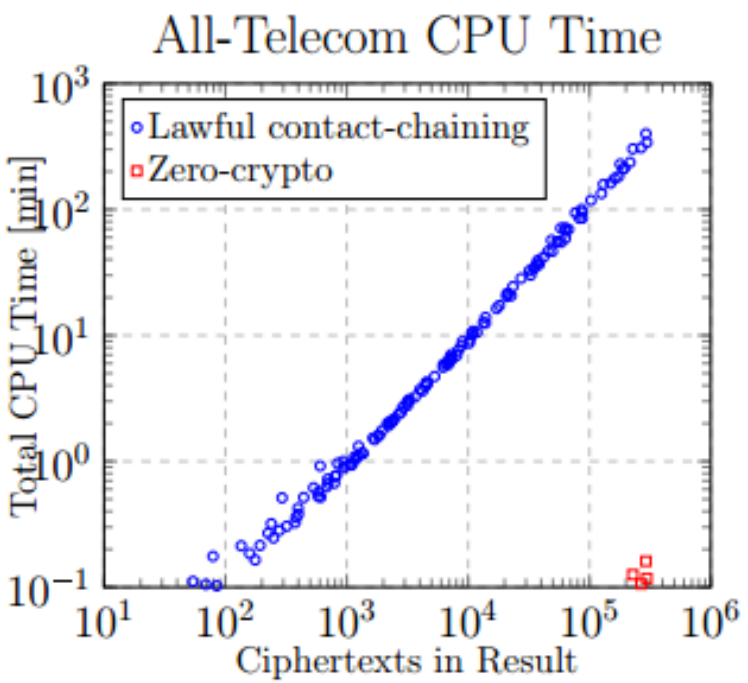
Contact Chaining Experimental Results

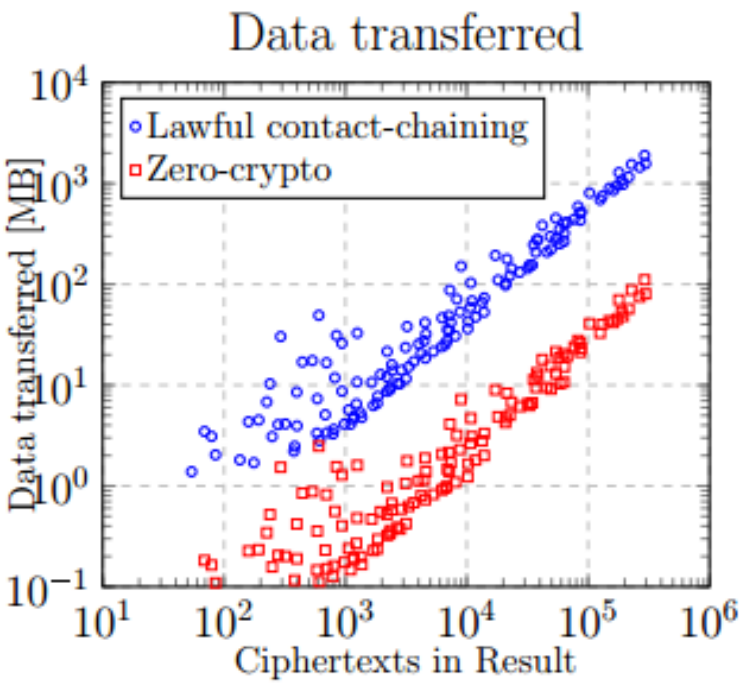
- Varied starting position, *k*, and *d* to examine a variety of neighborhood sizes
- Measured
  - End-to-end running time
  - CPU time used by all telecoms
  - Total bandwidth sent over network

Ciphertexts in result	End-to-end runtime MM:SS	Telecom CPU Time H:MM:SS	Bytes transferred MB
582	00:05	0:00:32	18
1061	00:06	0:00:57	6
5301	00:23	0:04:43	22
10188	00:37	0:08:41	36
27338	01:50	0:28:23	132
49446	03:15	0:46:28	222
102899	07:43	1:58:16	804
149535	10:25	2:42:49	896
194231	13:57	3:34:48	978
297474	21:51	5:41:43	1570

Contact Chaining Experimental Results





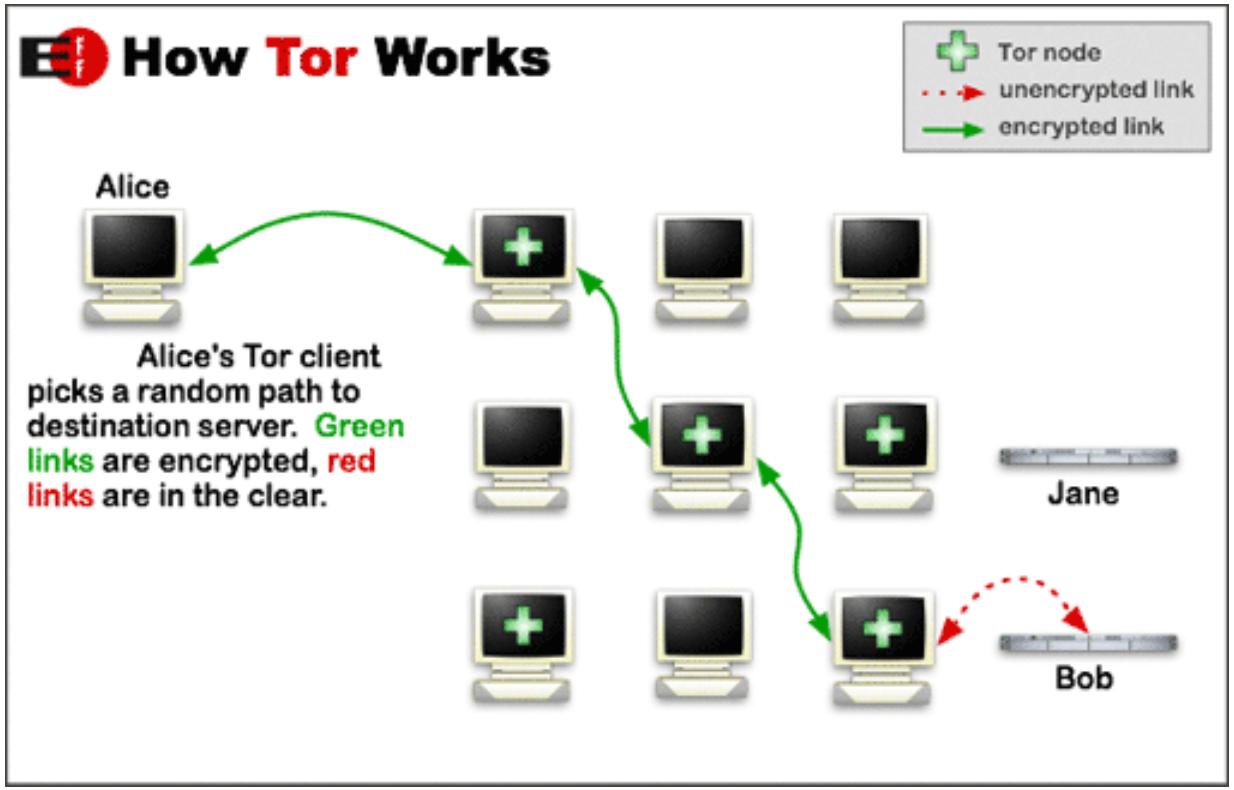


Privacy-Preserving Contact Chaining and Intersection

- Privacy-preserving contact chaining & set intersection together
- Our principles apply to other surveillance of private data
- No need for new cryptographic tools, “backdoors,” or secret processes

Anonymity: Users Protecting Themselves With Tor

- Anonymous communication dissociates network activity from user identity
- Tor: The Second-Generation Onion Router [DMS 2004]
  - 2 million Tor users daily
  - 7000+ volunteer relays in the Tor network
- Connections made through three relays: *guard*, *middle*, *exit*
- Vulnerability: Adversary who can view *guard* and *exit* traffic together



TorFlow: Critical but Vulnerable

- TorFlow conducts *bandwidth scans* to measure all 7000+ relays
- Relays can determine when they're being scanned
  - Exploit: Give better service to measurement authorities
- Bandwidth scans use only *two* relays, not *three*
  - Exploit: Launch DoS on another relay by blocking traffic only when on a circuit with that relay
- Results of scans are used only to proportionally adjust *self-reported* measurements
  - Exploit: Lie

PeerFlow: Secure Load Balancing Alternative

- Periodically estimate relay bandwidth and use estimates to calculate selection weight

- Three estimates of relay bandwidth:
  1. Measurements collected from relays about other relays
     - Use natural traffic to generate measurements
     - Ignore measurements made by smaller relays
     - Add random noise to measurements before sending
  2. Self-reports from relays
     - Relays report estimate of own capacity
     - Reports are not trusted
  3. Expected traffic carried
     - Based on selection weight in last measurement period

PeerFlow: High-level Idea

- Use estimates to choose relay selection weight
  - Selection weight ~= fraction of traffic carried

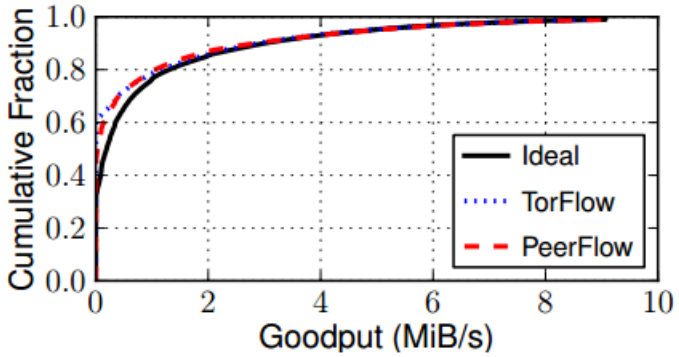
If measured bandwidth ≥ expected bandwidth and self-reported bandwidth > measured bandwidth:

Increase selection weight

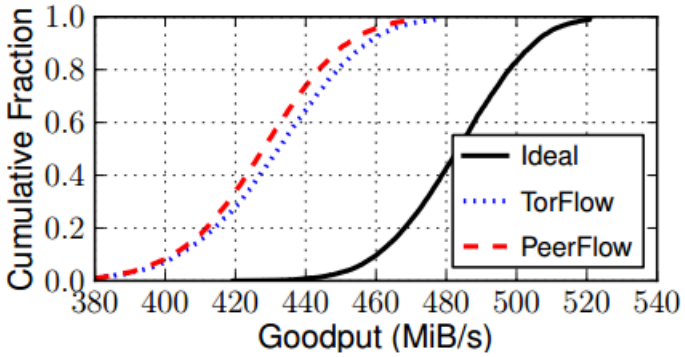
If measured bandwidth < expected bandwidth and self-reported bandwidth > measured bandwidth:

Decrease selection weight in next period to be equal to measured bandwidth in that period

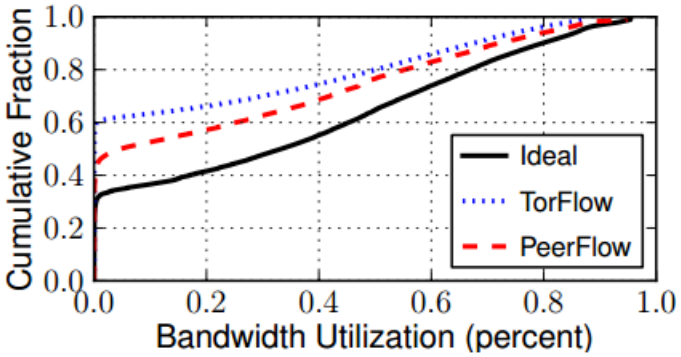
Performance of Peerflow



(a) Relay goodput per second



(b) Aggregate relay goodput per second



(c) Relay utilization per second

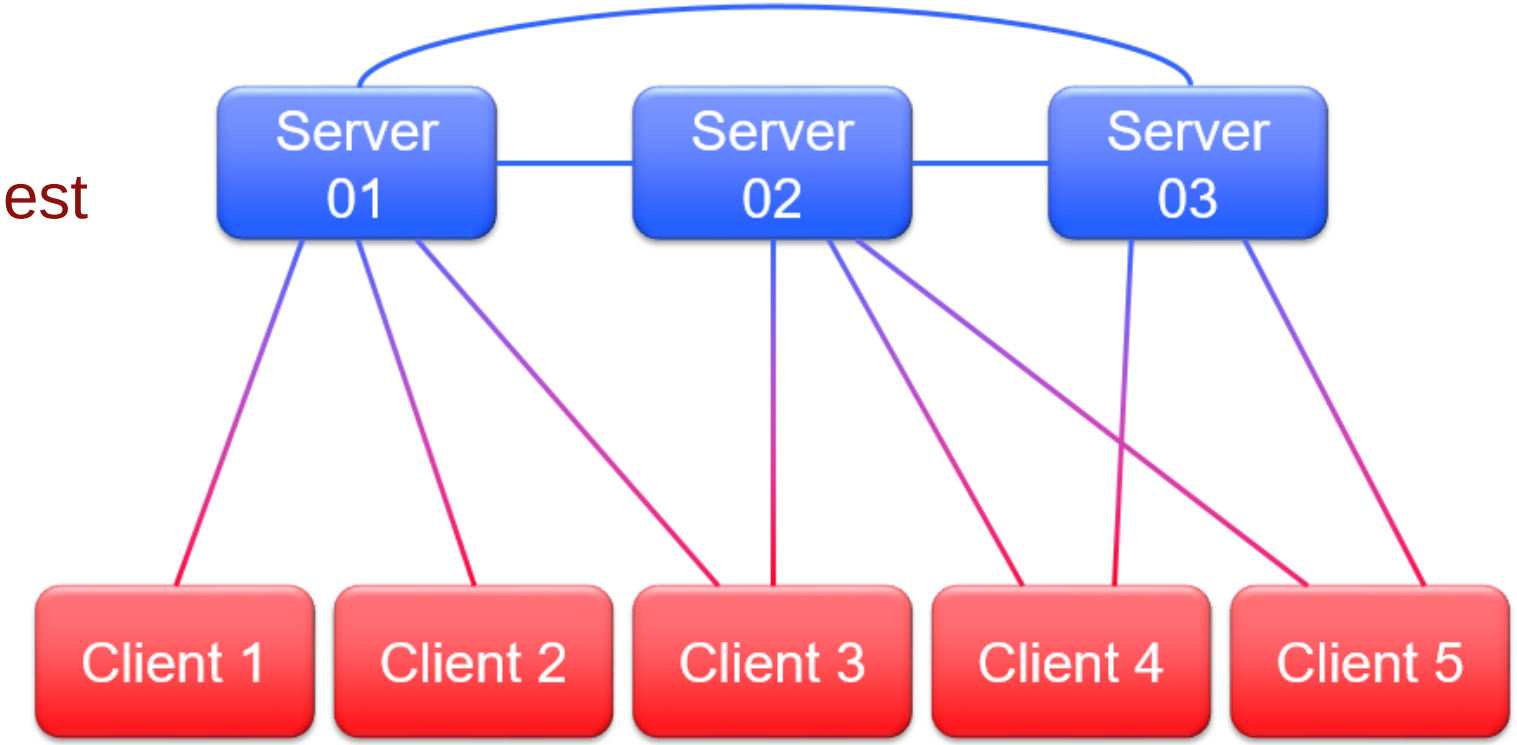
Verdict: Alternative to Tor

- Verdict: accountable anonymity through Dining-Cryptographers Networks (DC-Nets)
  - Original paper: Henry Corrigan-Gibbs, David Isaac Wolinsky, Bryan Ford (USENIX 2013)
- Not vulnerable to an adversary, even if they can view all messages
- Trade-off: Users take turns sending messages over network, increasing latency
- Proof of security!

Verdict Architecture

- Multi-provider cloud
  - Each client connected with one or more servers
  - Each server connected with all other servers

- Anytrust
  - At least one server is honest



Verdict Properties Proven

- **Accountability**
  - Whenever the protocol fails, an honest node can produce a proof that shows a deviation from the protocol on the part of one other participant
  - A dishonest participant can’t produce a proof blaming an honest participant
    - With every message, each participant sends a non-interactive zero-knowledge proof that the sender is following the protocol
- **Anonymity**
- **Integrity**

Verdict Properties Proven

- **Accountability**
- **Anonymity**
  - As long as there are two honest clients, no other participant can tell which client sends which message, even if they can see all messages being sent over the wire
    - Adversary can’t distinguish between encryptions of messages without breaking security of underlying encryption scheme or zero-knowledge property of proof scheme
- **Integrity**

Verdict Properties Proven

- **Accountability**
- **Anonymity**
- **Integrity**
  - Either all clients receive accurate messages from all other clients, or all clients know that the protocol failed
  - Forging or altering messages is impossible
    - Straightforward as long as E(*m*)+E(0)+E(0)+E(0)+... = E(m) and proofs of knowledge can’t be forged

Conclusions

- Privacy-preserving surveillance *is* technologically feasible
- Privacy-preserving set intersection and contact chaining can accomplish law-enforcement goals with open processes and without users losing control of their data

- Anonymity through Tor is practical and can be secured against bandwidth-inflation attacks using PeerFlow
- Verdict offers *provably* accountable anonymity as alternative to Tor

Thank you!