FOSAD 2016

Privacy-preserving Information Sharing: Tools and Applications (Volume 2)

Emiliano De Cristofaro

University College London (UCL)

https://emilianodc.com

Prologue

Privacy-Enhancing Technologies (PETs):

- Increase privacy of users, groups, and/or organizations

PETs often respond to privacy threats

- Protect personally identifiable information
- Support anonymous communications
- Privacy-respecting data processing

Another angle: privacy as an enabler

- Actively enabling scenarios otherwise impossible w/o clear privacy guarantees

Sharing Information w/ Privacy

Needed when parties with limited mutual trust willing or required to share information

- Only the required minimum amount of information should be disclosed in the process

Private Set Intersection?

DHS (Terrorist Watch List) and **Airline** (Passenger List)

Find out whether any suspect is on a given flight

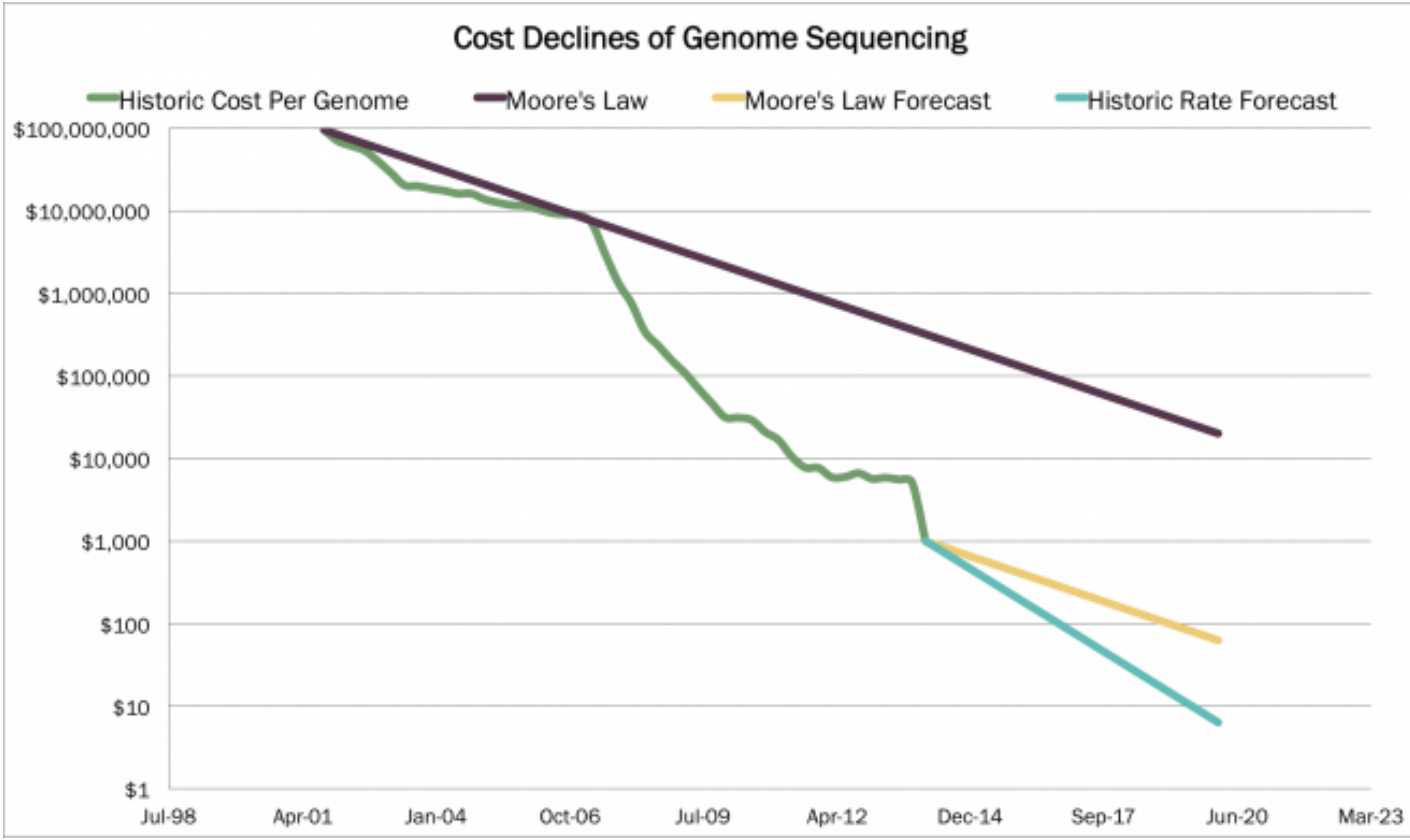
IRS (Tax Evaders) and **Swiss Bank** (Customers)

Discover if tax evaders have accounts at foreign banks

Hoag Hospital (Patients) and **SSA** (Social Security DB)

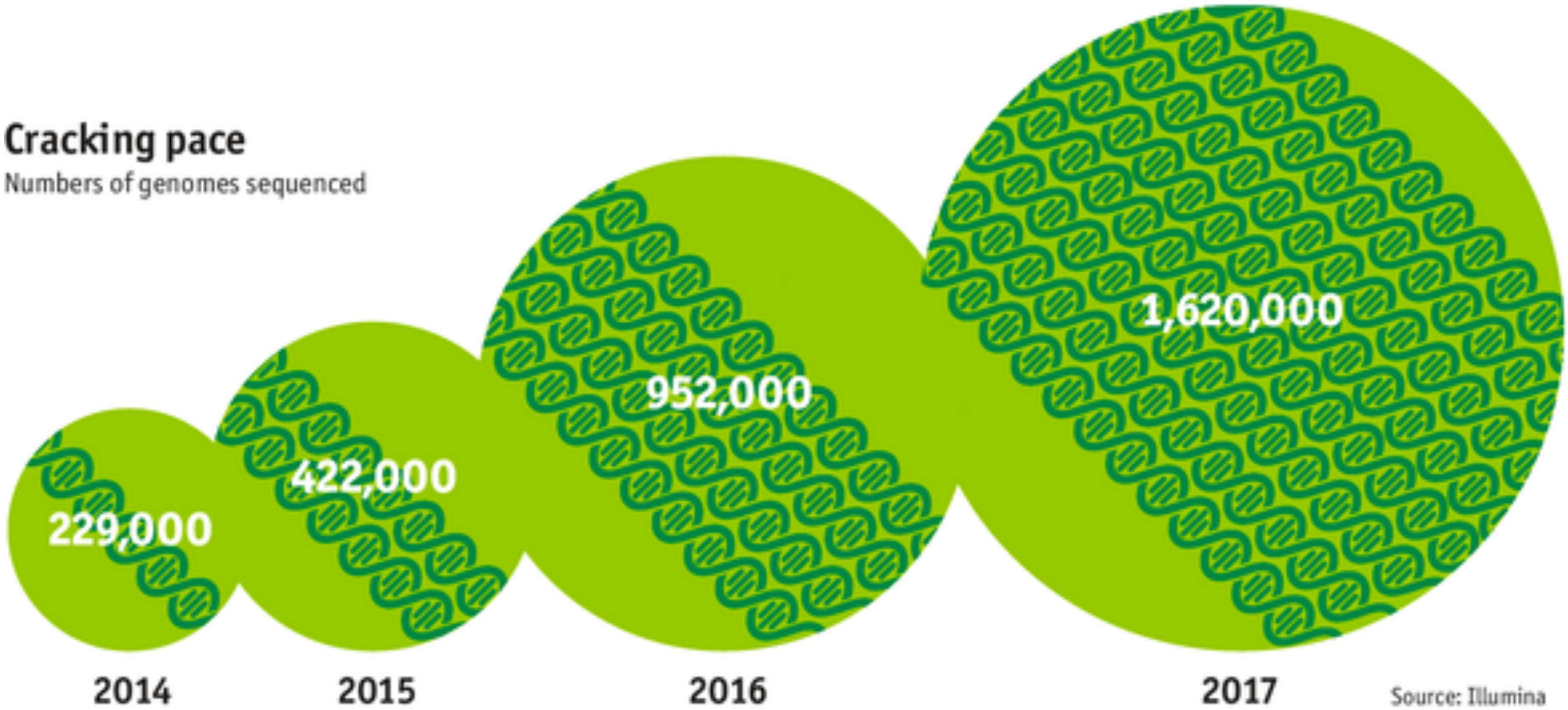
Patients with fake Social Security Number

Genomics



From: James Bannon, ARK

From: The Economist



1/05/2011 @ 4:57PM | 30,076 views

The First Child Saved By DNA Sequencing

+ Comment Now + Follow Comments

Genetic Gamble

New Approaches to Fighting Cancer

PART ONE
A Race to Leukemia's Source

PART TWO
Promise and Heartbreak

In Treatment for Leukemia, Glimpses of the Future

LETTER

doi:10.1038/nature13394

Genome sequencing identifies major causes of severe intellectual disability

Christian Gilissen[1]*, Jayne Y. Hehir-Kwa[1]*, Djie Tjwan Thung[1], Maartje van de Vorst[1], Bregje W. M. van Bon[1], Marjolein H. Willemsen[1], Michael Kwint[1], Irene M. Janssen[1], Alexander Hoischen[1], Annette Schenck[1], Richard Leach[2], Robert Klein[2], Rick Tearle[2], Tan Bo[1,3], Rolph Pfundt[1], Helger G. Yntema[1], Bert B. A. de Vries[1], Tjitske Kleefstra[1], Han G. Brunner[1,4]*, Lisenka E. L. M. Vissers[1]* & Joris A. Veltman[1,4]*


MAY 27, 2013
TIME
THE
ANGELINA
EFFECT
Angelina Jolie's double mastectomy puts genetic
testing in the spotlight. What her choice reveals
about calculating risk, cost and peace of mind
BY JEFFREY KLUGER & ALICE PARK
time.com
Time

Genetic Risk Factors (11)

REPORT	RESULT
Alpha-1 Antitrypsin Deficiency	Variant Absent; Typical Risk
Alzheimer's Disease (APOE Variants)	ε4 Variant Absent
Early-Onset Primary Dystonia (DYT1-TOR1A-Related)	Variant Absent; Typical Risk
Factor XI Deficiency	Variant Absent; Typical Risk
Familial Hypercholesterolemia Type B (APOB-Related)	Variant Absent; Typical Risk

See all 11 genetic risk factors...

Inherited Conditions (43)

REPORT	RESULT
Beta Thalassemia	Variant Present
ARSACS	Variant Absent
Agenesis of the Corpus Callosum with Peripheral Neuropathy (ACCPN)	Variant Absent
Autosomal Recessive Polycystic Kidney Disease	Variant Absent
Bloom's Syndrome	Variant Absent

See all 43 carrier status...

Traits (41)

REPORT	RESULT
Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Can Taste
Blond Hair	28% Chance
Earwax Type	Wet
Eye Color	Likely Brown

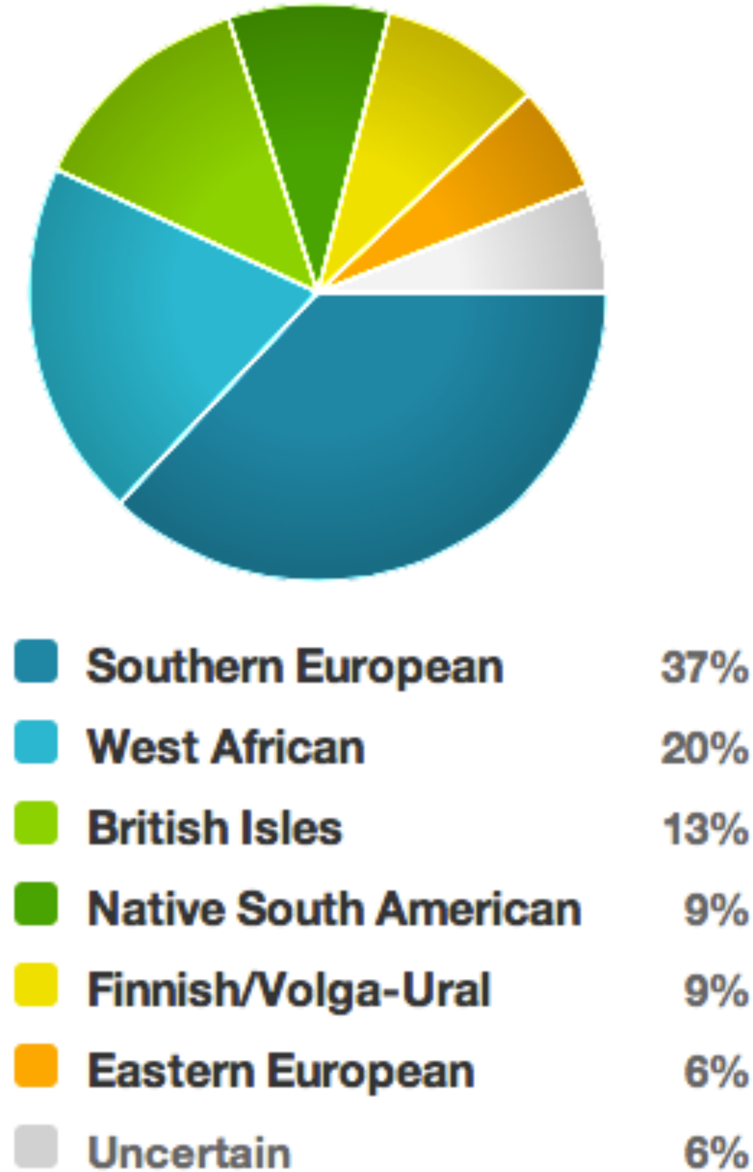
See all 41 traits...

Drug Response (12)

REPORT	RESULT
Proton Pump Inhibitor (PPI) Metabolism (CYP2C19-related)	Rapid
Warfarin (Coumadin®) Sensitivity	Increased
Phenytoin Sensitivity (Epilepsy Drug)	Increased
Sulfonylurea Metabolism	Greatly reduced
Abacavir Hypersensitivity	Typical

See all 12 drug response...

Genetic Ethnicity



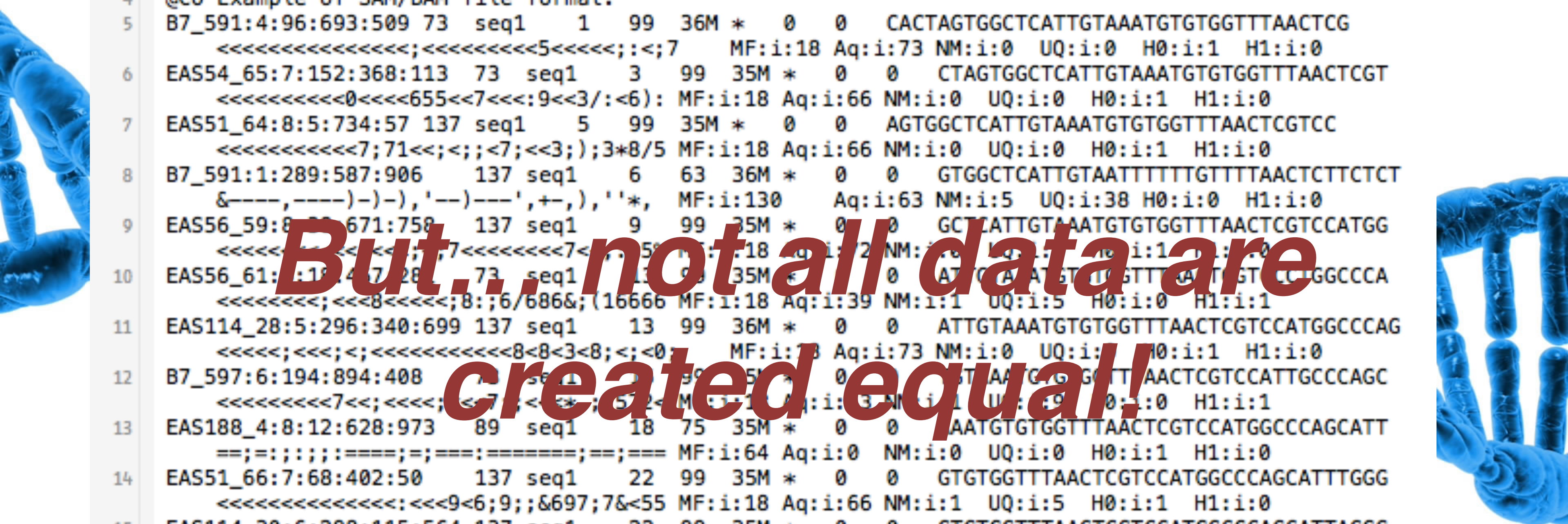
DNA RELATIVES DOWNLO

List View | Map View | Surname View

search matches | Show: both sides | Sort: relationship | 25 per page | 1 - 25 of 424

[redacted] Male	You	[redacted]	UPDATE YOUR PROFILE
Female	2nd to 3rd Cousin 1.68% shared, 5 segments	J2a2	Send an Introduction
[redacted] Female	3rd to 4th Cousin 1.30% shared, 3 segments	United States · Alsace-Lorraine (Strasbourg), Fr... · Paternal · [redacted] · Senape · 5 more · U5b2	Public Match Send a Message
Male	3rd to 4th Cousin 1.03% shared, 2 segments	H13a1a · R1b1b2	Send an Introduction
Female	3rd to 5th Cousin 0.45% shared, 2 segments	H7	Send an Introduction
Female	3rd to 5th Cousin 0.42% shared, 2 segments	H1	Send an Introduction
[redacted] Male	3rd to 5th Cousin 0.40% shared, 2 segments	United States · Reno, Nevada · San Diego, California · Tucker · Littlefield · Warga · 4 more · H1c · G2a	Public Match Send a Message
[redacted] Male	3rd to 5th Cousin 0.37% shared, 2 segments	United States · fathers father prince Edward isla... · [redacted] · K1a1b · R1b1b2a1a	Public Match Send a Message
Male, b. 1978	3rd to 6th Cousin 0.40% shared, 1 segment	United States · New Jersey · Utah · California · Northern Europe · U3b1 · T	Send an Introduction

But… not all data are created equal!



ex1.sam

ex1.sam 〉 No Selection

```
@HD VN:1.0 SO:coordinate
@SQ SN:seq1 LN:5000
@SQ SN:seq2 LN:5000
@CO Example of SAM/BAM file format.
B7_591:4:96:693:509 73  seq1    1   99  36M *   0   0   CACTAGTGGCTCATTGTAAATGTGTGGTTTAACTCG
    <<<<<<<<<<<<<<<;<<<<<<<<<5<<<<<;:<;7    MF:i:18 Aq:i:73 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS54_65:7:152:368:113  73  seq1    3   99  35M *   0   0   CTAGTGGCTCATTGTAAATGTGTGGTTTAACTCGT
    <<<<<<<<<<0<<<<655<<7<<<:9<<3/:<6): MF:i:18 Aq:i:66 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS51_64:8:5:734:57 137 seq1    5   99  35M *   0   0   AGTGGCTCATTGTAAATGTGTGGTTTAACTCGTCC
    <<<<<<<<<<<7;71<<;<;;<7;<<3;);3*8/5 MF:i:18 Aq:i:66 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
B7_591:1:289:587:906    137 seq1    6   63  36M *   0   0   GTGGCTCATTGTAATTTTTTGTTTTAACTCTTCTCT
    &----,----)-)-),'--)---',+-,),''*,  MF:i:130    Aq:i:63 NM:i:5  UQ:i:38 H0:i:0  H1:i:0
EAS56_59:[illegible]671:758 137 seq1    9   99  35M *   [illegible] GC[illegible]ATTGT[illegible]AATGTGTGGTTTAACTCGTCCATGG
    <<<<<[illegible]<<<;[illegible]7<<<<<<<<7<[illegible]
EAS56_61:[illegible]  73  seq1    [illegible] 35M [illegible] 0   [illegible]GGCCCA
    <<<<<<<<;<<<8<<<<<;8:;6/686&;(16666 MF:i:18 Aq:i:39 NM:i:1  UQ:i:5  H0:i:0  H1:i:1
EAS114_28:5:296:340:699 137 seq1    13  99  36M *   0   0   ATTGTAAATGTGTGGTTTAACTCGTCCATGGCCCAG
    <<<<<;<<<;<;<<<<<<<<<<<8<8<3<8;<;<0;    MF:i:[illegible] Aq:i:73 NM:i:0  UQ:i:[illegible]0:i:1  H1:i:0
B7_597:6:194:894:408    [illegible] AACTCGTCCATTGCCCAGC
    <<<<<<<<<7<<;<<<<;[illegible]H1:i:1
EAS188_4:8:12:628:973   89  seq1    18  75  35M *   0   0   [illegible]AATGTGTGGTTTAACTCGTCCATGGCCCAGCATT
    ==;=:;:;;:====;=;===:=======;==;=== MF:i:64 Aq:i:0  NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS51_66:7:68:402:50    137 seq1    22  99  35M *   0   0   GTGTGGTTTAACTCGTCCATGGCCCAGCATTTGGG
    <<<<<<<<<<<<<<:<<<9<6;9;;&697;7&<55 MF:i:18 Aq:i:66 NM:i:1  UQ:i:5  H0:i:1  H1:i:0
EAS114_30:6:298:115:564 137 seq1    22  99  35M *   0   0   GTGTGGTTTAACTCGTCCATGGCCCAGCATTAGGG
    <<<<<<<<<<<<<;<<<;<<<;<<<;<:<<;;5;; MF:i:18 Aq:i:72 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
B7_591:3:188:662:155    73  seq1    24  99  36M *   0   0   GTGGTTTAACTCGTCCATGGCCCAGCATTAGGGAGC
    <<<<<<<<<<<<<<<<<<:;<<<4<<+<<14991;4    MF:i:18 Aq:i:71 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS56_59:2:225:608:291  73  seq1    28  99  35M *   0   0   TTTAACTCGTCCATGGCCCAGCATTAGGGATCTGT
    <<<<<<<<<<<<<<8&<<<;6<9;;+2++(%59(< MF:i:18 Aq:i:58 NM:i:1  UQ:i:4  H0:i:1  H1:i:0
EAS51_66:7:328:397:316  73  seq1    29  99  35M *   0   0   TTAACTCGTCCATGGCCCAGCATTAGGGAGCTGTG
    <<<<<<<<<<<<<6=<<<;<<5<<<+<15:'<;;4 MF:i:18 Aq:i:69 NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS51_64:5:257:960:682  73  seq1    31  75  35M *   0   0   AACTCGTCCATGGCCCAGCATTAGGGAGCTGTGGA
    <<<<<<<<<<<;<<<<<<<<<<<<<<<<<9;;9<  MF:i:64 Aq:i:0  NM:i:0  UQ:i:0  H0:i:1  H1:i:0
EAS54_61:4:143:69:578   99  seq1    36  98  35M =   185 184 GTACATGGCCCAGCATTAGGGAGCTGTGGACCCCG
    ===;=====48=844;=;+=5==*57,2+5&,5+5 MF:i:18 Aq:i:35 NM:i:2  UQ:i:38 H0:i:0  H1:i:1
```

Privacy Researcher’s Perspective

Treasure trove of sensitive information

Ethnic heritage, predisposition to diseases

Genome = the ultimate identifier

Hard to anonymize / de-identify

Sensitivity is perpetual

Cannot be “revoked”

Leaking one’s genome ≈ leaking relatives’ genome

Secure Genomics?

Privacy:

Individuals remain in control of their genome

Allow doctors/clinicians/labs to run genomic tests, while disclosing the required minimum amount of information, i.e.:

(1) Individuals don’t disclose their entire genome

(2) Testing facilities keep test specifics (“secret sauce”) confidential

[BBDGT11]: Secure genomics via PSI

Most personalized medicine tests in < 1 second

Works on Android too

Genetic Paternity Test

A Strawman Approach for Paternity Test:

On average, ~99.5% of any two human genomes are identical

Parents and children have even more similar genomes

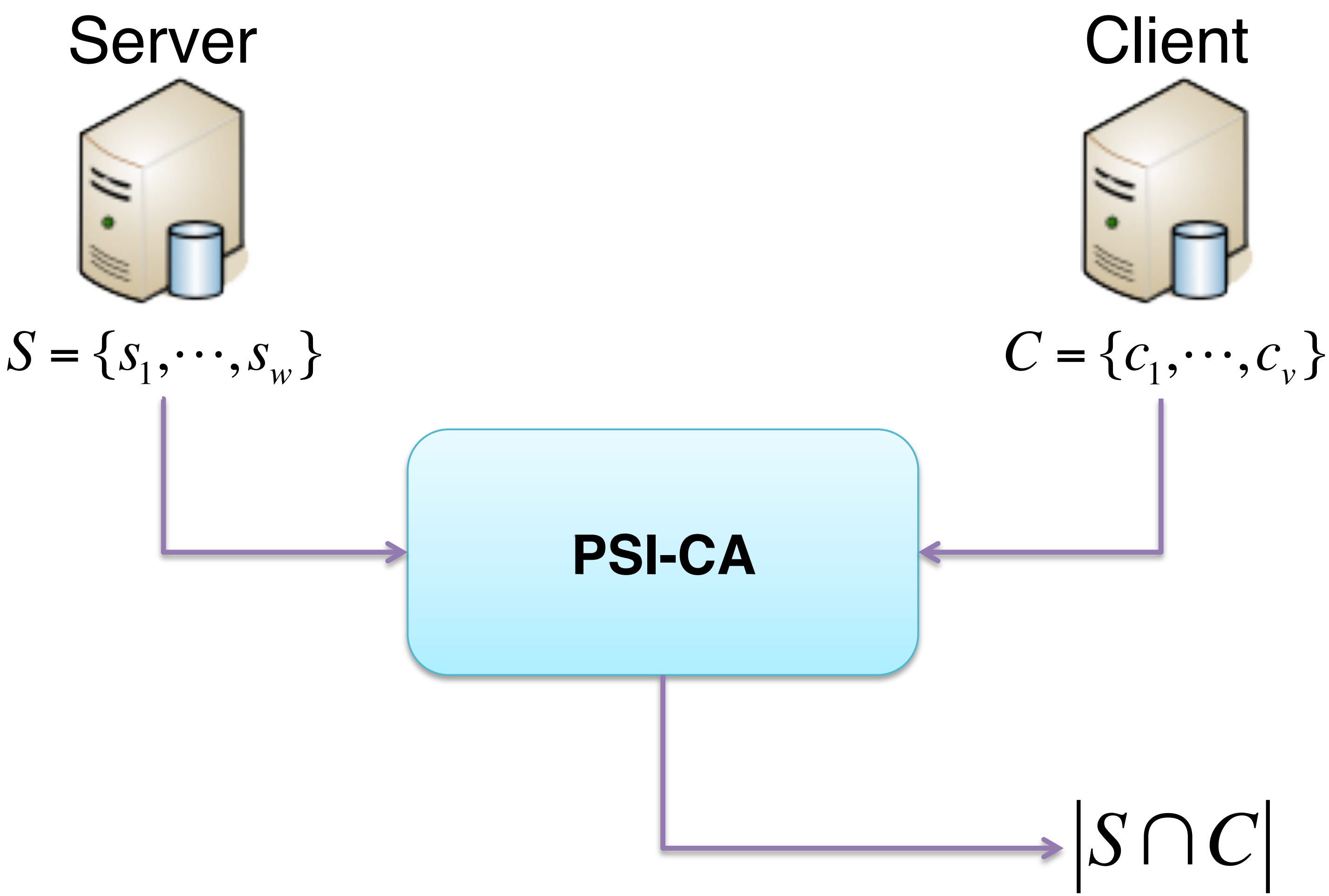
Compare candidate's genome with that of the alleged child:

Test positive if percentage of matching nucleotides is > 99.5 + τ

First-Attempt Privacy-Preserving Protocol:

Use an appropriate secure two-party protocol for the comparison

Private Set Intersection Cardinality (PSI-CA)



Genetic Paternity Test

A Strawman Approach for Paternity Test:

On average, ~99.5% of any two human genomes are identical

Parents and children have even more similar genomes

Compare candidate's genome with that of the alleged child:

Test positive if percentage of matching nucleotides is > 99.5 + τ

First-Attempt Privacy-Preserving Protocol:

Use an appropriate secure two-party protocol for the comparison

PROs: High-accuracy and error resilience

CONs: Performance not promising (3 billion symbols in input)

In our experiments, computation takes a few days

Genetic Paternity Test

Wait a minute!

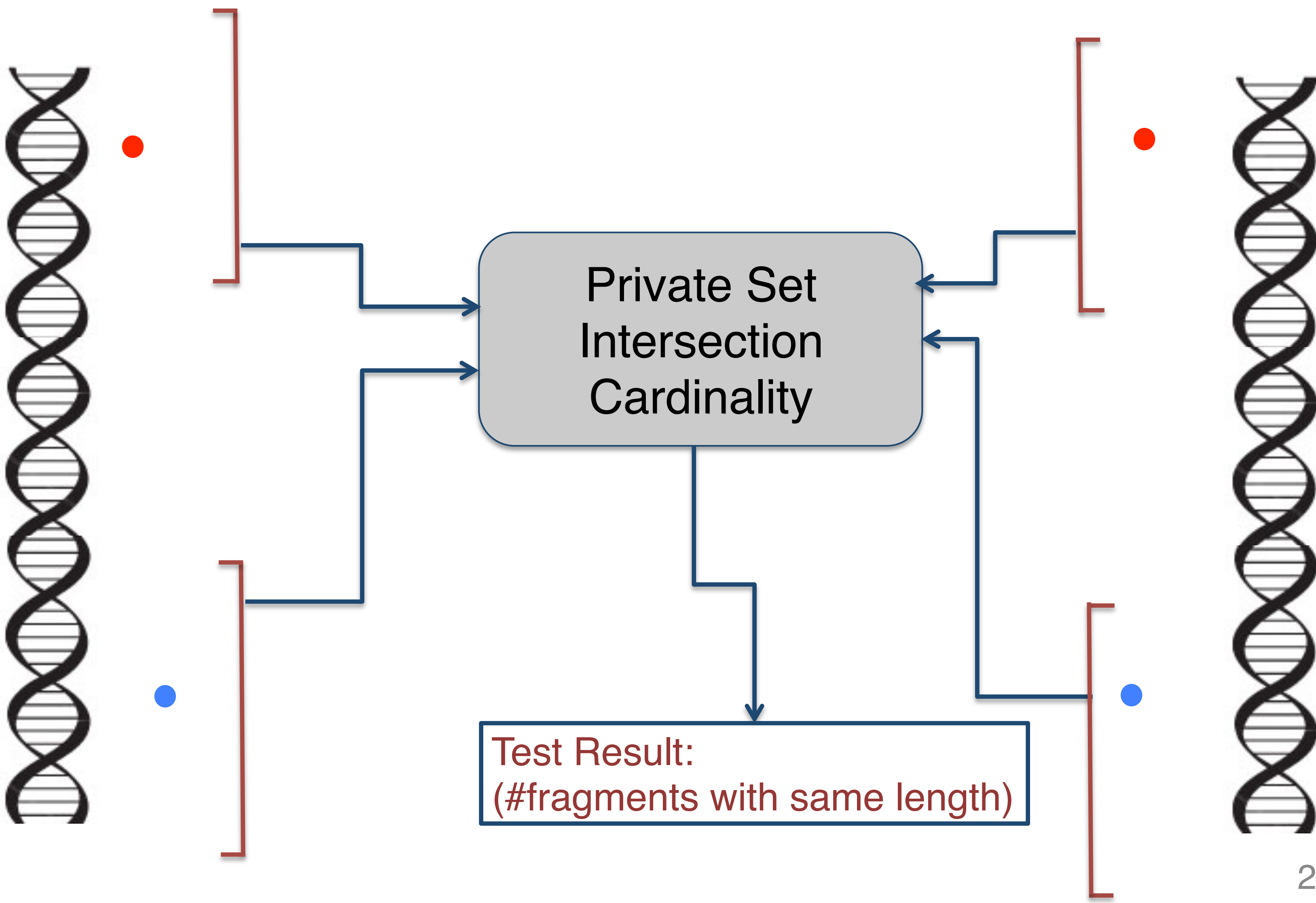
~99.5% of any two human genomes are identical

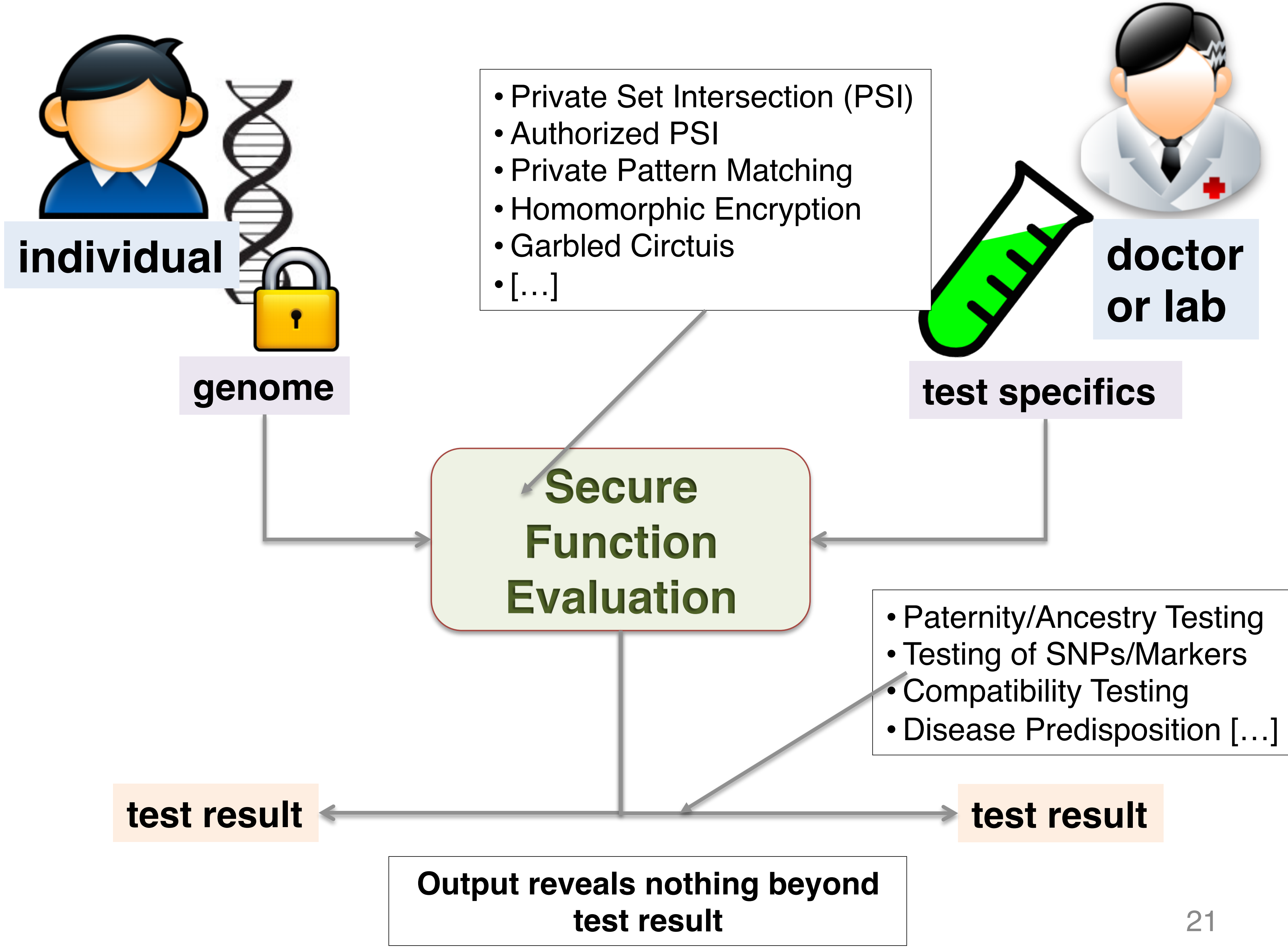
Why don't we compare *only* the remaining 0.5%?

We can compare by counting how many

But… We don't know (yet) where *exactly* this 0.5% occur!

Private RFLP-based Paternity Test




individual
genome
• Private Set Intersection (PSI)
• Authorized PSI
• Private Pattern Matching
• Homomorphic Encryption
• Garbled Circtuis
• […]
doctor or lab
test specifics
Secure Function Evaluation
• Paternity/Ancestry Testing
• Testing of SNPs/Markers
• Compatibility Testing
• Disease Predisposition […]
test result
test result
Output reveals nothing beyond test result

Personalized Medicine (PM)

Drugs designed for patients' genetic features

- Associating drugs with a unique genetic fingerprint
- Max effectiveness for patients with matching genome
- **Test drug's "genetic fingerprint" against patient's genome**

Examples:

- *tmpt* gene – relevant to leukemia
  - (1) G->C mutation in pos. 238 of gene's c-DNA, or (2) G->A mutation in pos. 460 and one A->G is pos. 419 cause the *tpmt* disorder (relevant for leukemia patients)
- *hla-B* gene – relevant to HIV treatment
  - One G->T mutation (known as *hla-B*5701* allelic variant) is associated with extreme sensitivity to abacavir (HIV drug)

Privacy-preserving PM Testing (P^3MT)

Challenges:

Patients may refuse to unconditionally release their genomes

Or may be sued by their relatives…

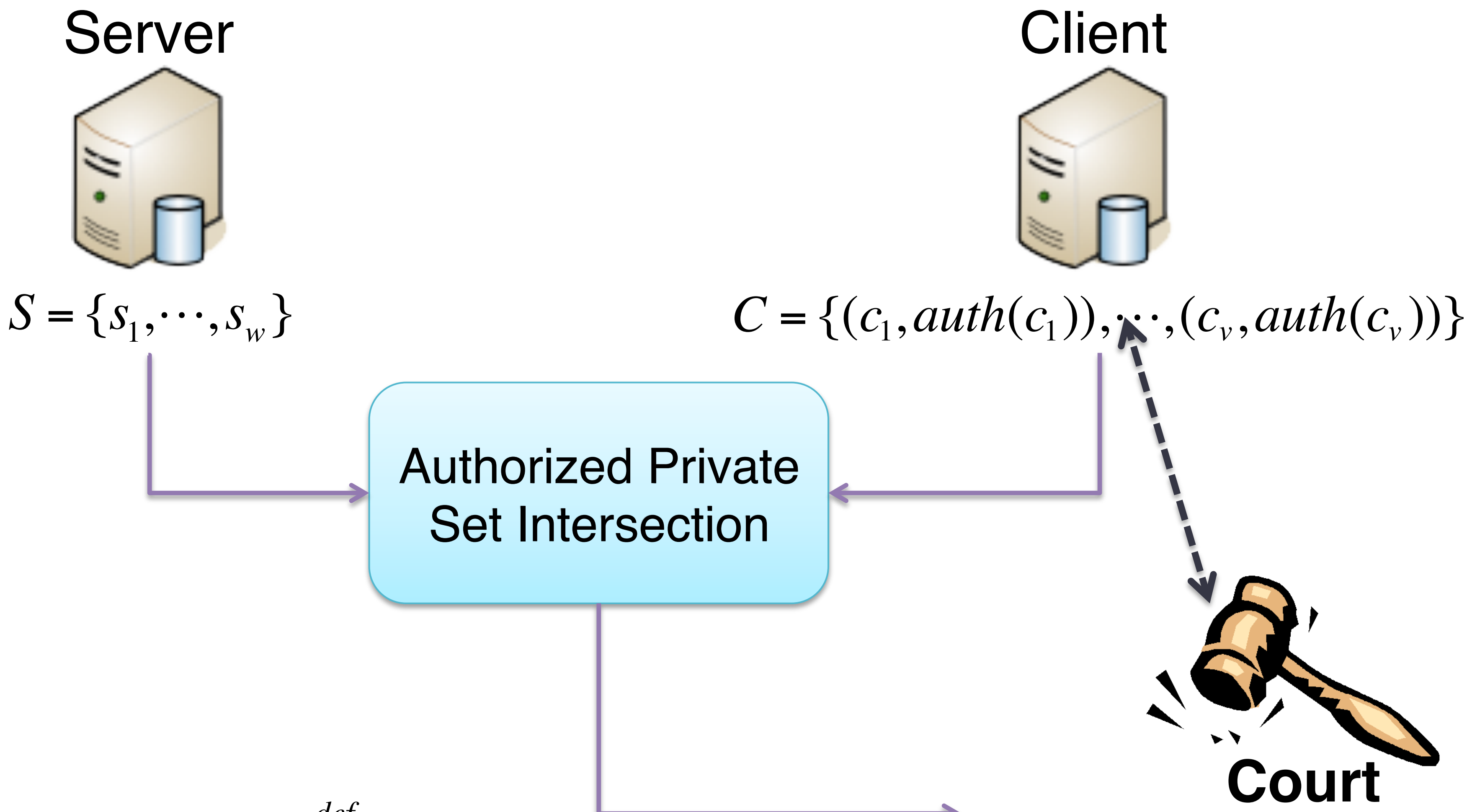
DNA fingerprint corresponding to a drug may be proprietary:

✓ **We need privacy-protecting fingerprint matching**

But we also need to enable FDA approval on the drug/fingerprint

✓ **We reduce P^3MT to Authorized Private Set Intersection (APSI)**

Authorized Private Set Intersection (APSI)



$$S \cap C \stackrel{def}{=} \{ s_j \in S \mid \exists c_i \in C : c_i = s_j \wedge auth(c_i) \; is \; valid \}$$

Reducing P³MT to APSI

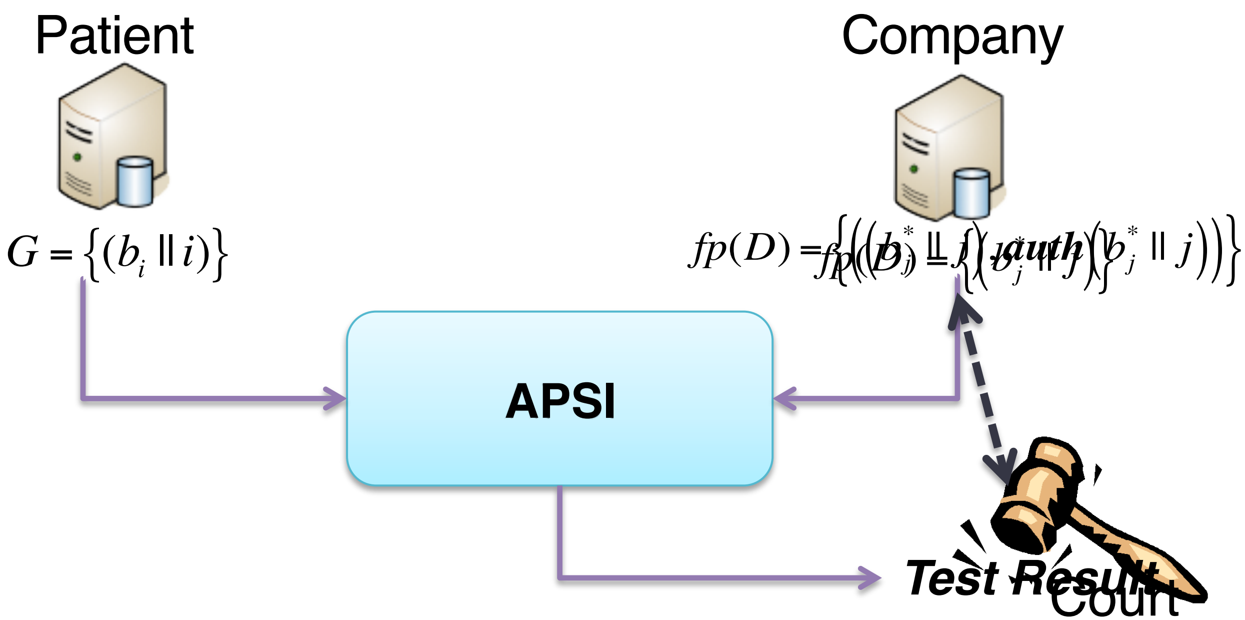
Intuition:

FDA = Court, Pharma = *Client*, Patient = *Server*

Patient's private input set: $G = \left\{ (b_i \| i) \middle| b_i \in \{A, C, G, T\} \right\}_{i=1}^{3 \cdot 10^9}$

Pharmaceutical company's input set: $fp(D) = \left\{ \left(b_j^* \| j \right) \right\}$

Each item in $fp(D)$ needs to be authorized by FDA



P^3MT – Performance Evaluation

Pre-Computation

Patient's pre-processing of the genome: a few days

Optimization*:*

Patient applies reference-based compression techniques

Input all *differences* with *"reference"* genome (0.5%)

Online Computation

Depend (linearly) on fingerprint size – typically a few nucleotides, <1s for most tests

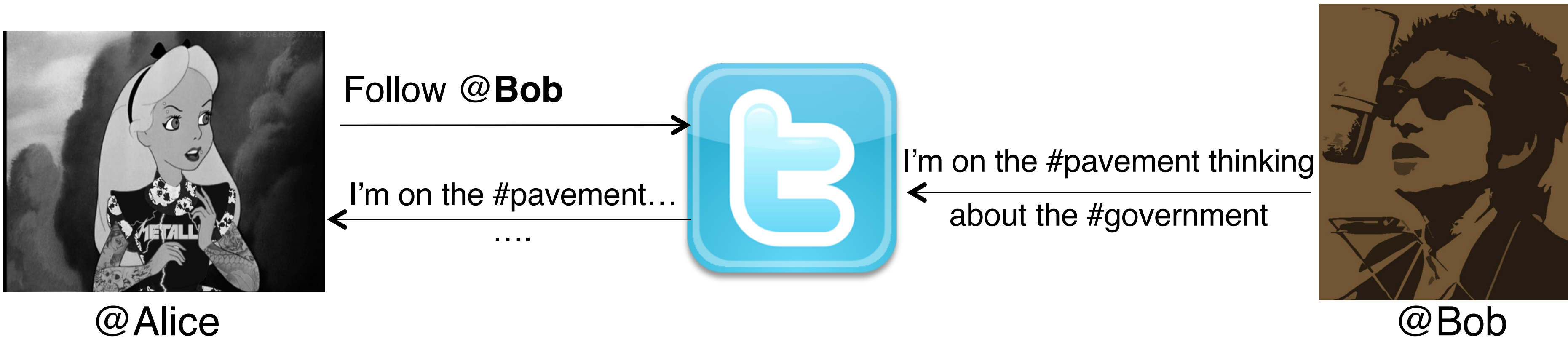
Communication

Depends on the size of encrypted genome (about 4GB)

Open Problems?

Micro-blogging

@Alice and @Bob – Twitter edition



There might be no mutual knowledge/trust between Alice and Bob

Follow requests are approved by default (opt-out)

Tweets are public by default

- Streamed into www.twitter.com/public_timeline, available through API
- But Bob can restrict his tweets to followers

All public tweets are searchable by hashtag

#Privacy and Twitter

Twitter.com is "trusted" to

- Get all tweets
- Enforce coarse-grained access control (follower-only)
- Monitor relations between users

Privacy and Twitter

- Targeted advertisement, PII collected and shared with third parties
- Trending topics, real-time "news"

I don't care about #privacy on @Twitter… but

- Remember @Wikileaks? Snowden?

Our proposal: Hummingbird

Follow by hashtag:

- E.g., @Alice follows @Bob only on hashtag #privacy

Tweeter (@Bob)

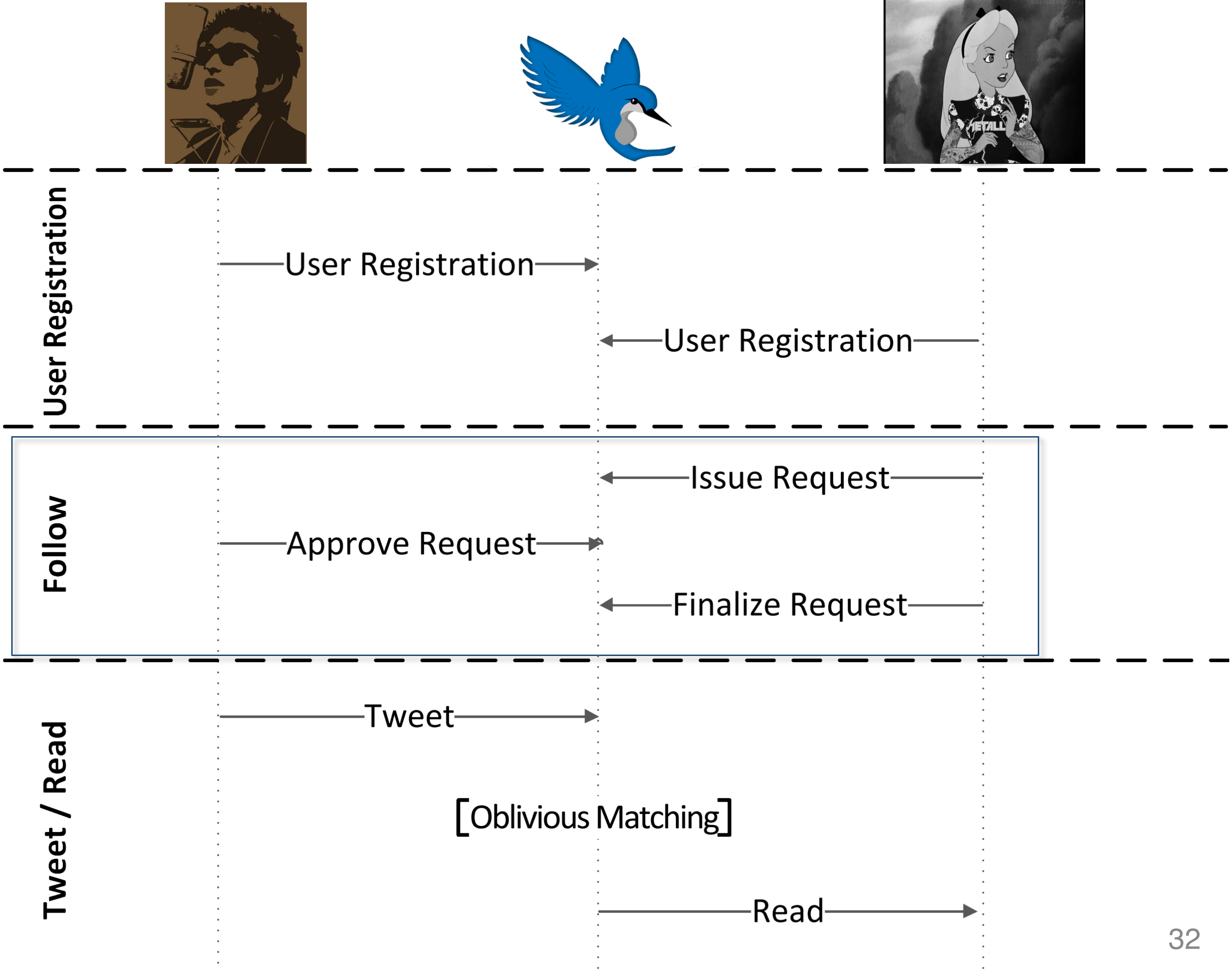
- Learns who follows him but not which hashtags have been subscribed to

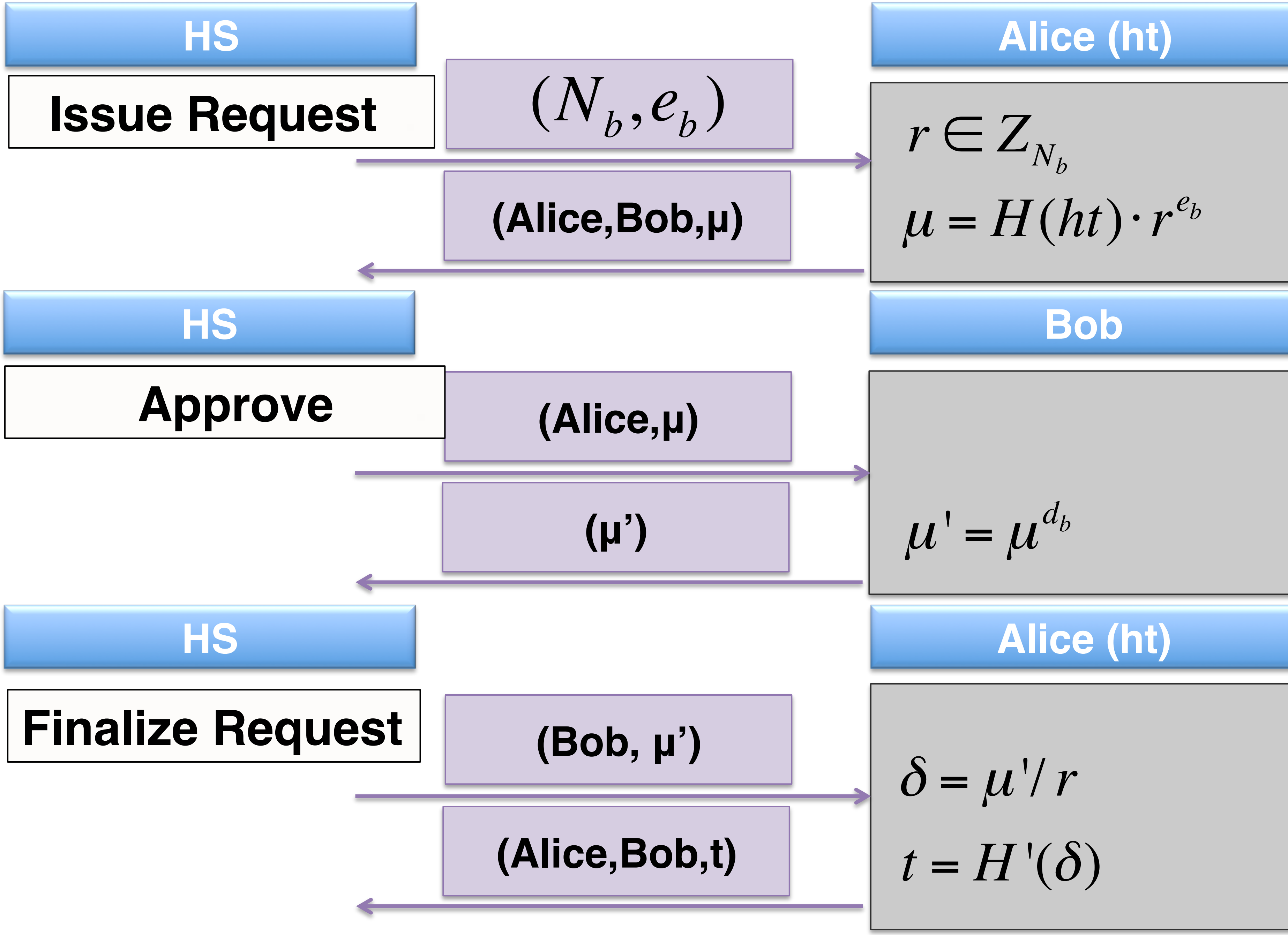
Follower (@Alice)

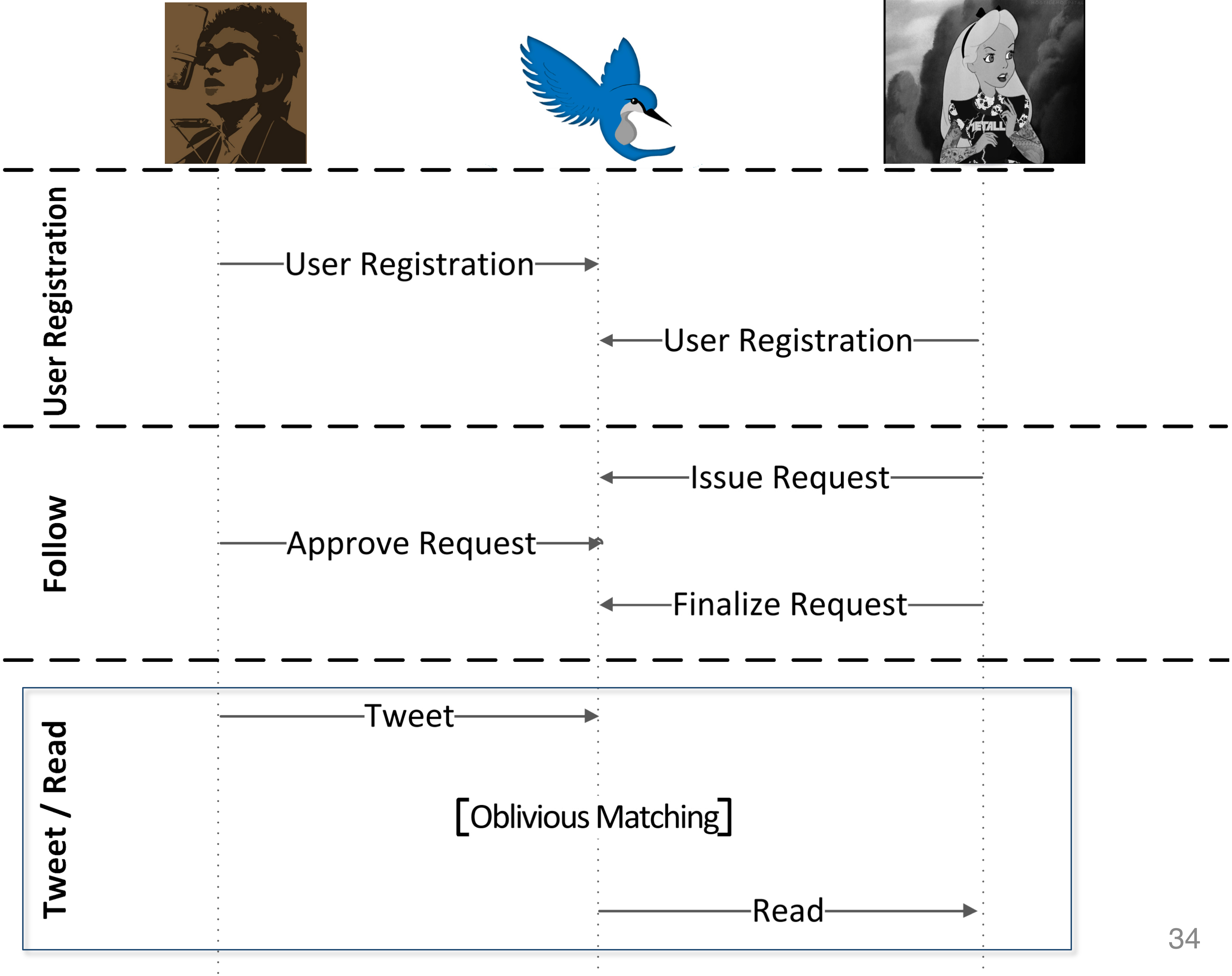
- Learns nothing beyond her own subscriptions

Server (HS)

- Doesn’t learn tweets’ content or hashtags of in
  - (But can scale to million of tweets/users)


User Registration
User Registration
User Registration
Follow
Issue Request
Approve Request
Finalize Request
Tweet / Read
Tweet
[Oblivious Matching]
Read


HS
Alice (ht)
Issue Request
(N_b, e_b)
$r \in Z_{N_b}$
$\mu = H(ht) \cdot r^{e_b}$
(Alice,Bob,μ)
HS
Bob
Approve
(Alice,μ)
$\mu' = \mu^{d_b}$
(μ')
HS
Alice (ht)
Finalize Request
(Bob, μ')
$\delta = \mu'/r$
$t = H'(\delta)$
(Alice,Bob,t)



User Registration
User Registration
User Registration
Follow
Issue Request
Approve Request
Finalize Request
Tweet / Read
Tweet
[Oblivious Matching]
Read

HS

Bob(d_b,M,ht*)

Tweet

$$\delta = H(ht^*)^{d_b}$$

$$t^* = H'(\delta) \quad k^* = H''(\delta)$$

$$ct^* = Enc_{k^*}(M)$$

$$(t^*, ct^*)$$

HS

For all (U,V,t) s.t. V='Bob' and t=t*:

Store and mark (Bob,t*,ct*) for delivering (t*,ct*) to Alice

Oblivious Matching

HS

Alice(δ,t)

Read

(Bob, t*, ct*)

$$k = H''(\delta)$$

$$M = Dec_k(ct^*)$$

Overhead

Follow protocol: Alice wants to follow Bob on #privacy

- Bob's computation: 1 CRT-RSA signature (<1ms) per hashtag
- Alice's computation: 2 mod multiplications per hashtag
- Communication: 2 RSA group elements (<1KB)

Tweet: Bob tweets "I'm at #fosad!"

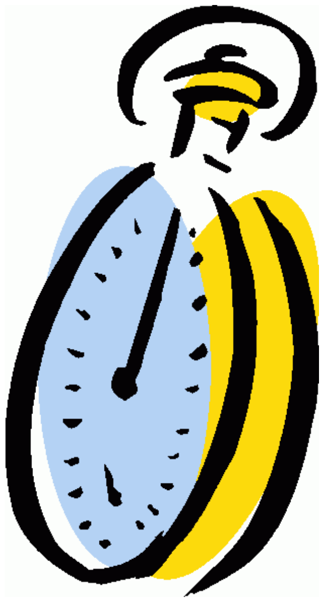
- Computation: 1 CRT-RSA signature (<1ms) per hashtag, 1 AES enc
- Communication: 1 hash output (160-bit)

Read

- Computation: 1 AES decryption
- Communication: 1 hash output (160-bit)

Server

- No crypto!
- Overhead: matching of PRF outputs, 160-bit
  - Can do efficiently, just like for cleartexts



Collecting Statistics Privately?

Collaboratively Train Machine Learning Models, Privately?

Why are statistics important?

Examples:

1. Recommender systems for online streaming services
2. Statistics about mass transport movements
3. Traffic statistics for the Tor Network

How about privacy?

Private Recommendations

BBC keeps 500-1000 free programs on iPlayer

No account, no tracking, no ads

Still, BBC wants to collect statistics, offer recommendations to its users

- E.g., you have watched Dr Who, maybe you'll like Sherlock Homes too!

Item-KNN Recommendation

Predict favorite items for users based on their own ratings and those of “similar” users

Consider ***N*** users, ***M*** TV programs and binary ratings (viewed/not viewed)

Build a co-views matrix ***C***, where $\boldsymbol{C_{ab}}$ is the number of views for the pair of programs (a,b)

Compute the **Similarity Matrix**

$$\{Sim\}_{ab} = \frac{C_{ab}}{\sqrt{C_a \cdot C_b}}$$

Identify K-Neighbours (***KNN***) based on matrix

Privacy-Preserving Aggregation

Goal: aggregator collects matrix, s.t.

Can only learn aggregate counts (e.g., 237 users have watched both a and b)

Not who has watched what

Use additively homomorphic encryption?

$Enc_{PK}(a)*Enc_{PK}(b) = Enc_{PK}(a+b)$

How can I used it to collect statistics?

Keys summing up to zero

Users $U_1, U_2, \ldots, U_N$, each has $k_1, k_2, \ldots, k_N$ s.t.

$k_1+k_2+\ldots+k_N=0$

Now how can I use this?

User $\mathcal{U}_i$ $(i \in [1, N])$ Tally

$$x_i \in_r \mathbb{G}, y_i := g^{x_i} \bmod q \quad \xrightarrow{\quad y_i \quad}$$

$$k_{i_\ell} := \sum_{j \neq i} \mathrm{H}(y_j^{x_i} || \ell || s) \cdot (-1)^{i>j} \bmod 2^{32} \quad \xleftarrow{\quad \{y_j\}_{j \in [1,N]} \quad}$$

$$b_{i_\ell} := X_{i_\ell} + k_{i_\ell} \bmod 2^{32} \quad \xrightarrow{\quad \{b_{i_\ell}\}_{\ell=1}^{L} \quad} \quad \text{Fault recovery (if needed)}$$

$$\xleftarrow{\quad \mathcal{U}^{on} \quad}$$

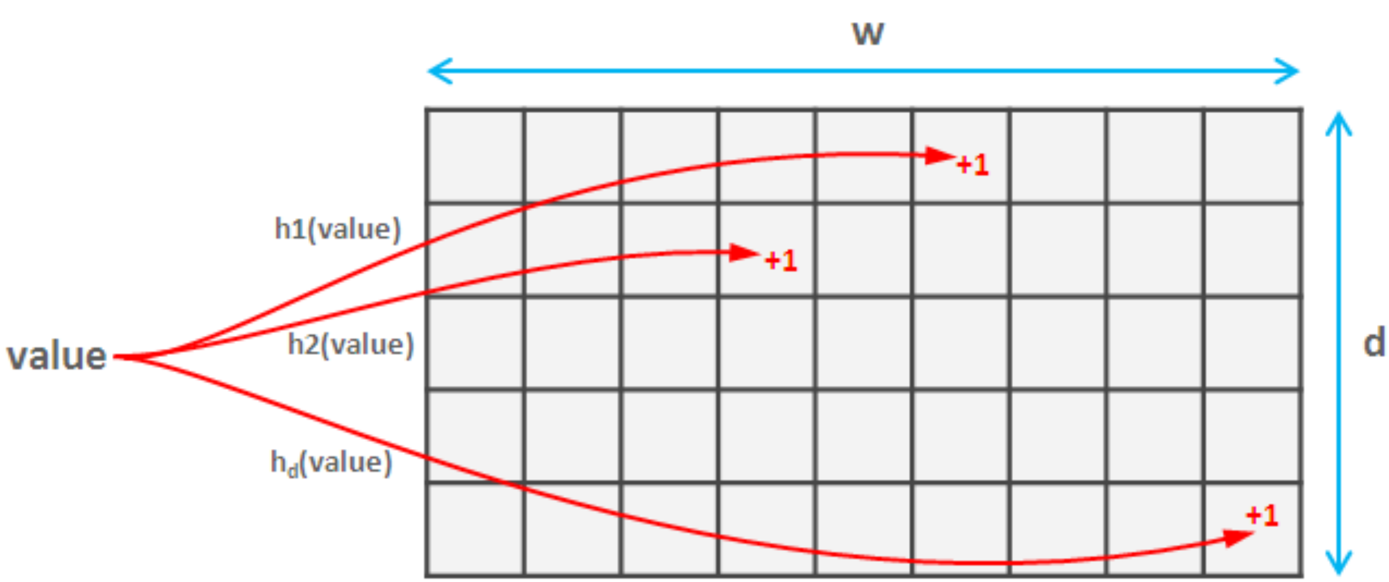
$$k'_{i_\ell} := \sum_{\substack{j \neq i, \\ j \notin \mathcal{U}^{on}}} \mathrm{H}(y_j^{x_i} || \ell || s) \cdot (-1)^{i>j} \bmod 2^{32} \quad \xrightarrow{\quad \{k'_{i_\ell}\}_{\ell=1}^{L} \quad} \quad C'_\ell := \left(\sum_{i \in \mathcal{U}^{on}} b_{i_\ell} - \sum_{i \in \mathcal{U}^{on}} k'_{i_\ell} \right) \bmod 2^{32}$$

Is this efficient?

Preliminaries: Count-Min Sketch

An estimate of an item's frequency in a stream

- Mapping a stream of values (of length T) into a matrix of size O(logT)
- The sum of two sketches results in the sketch of the union of the two data streams



Security & Implementation

Security

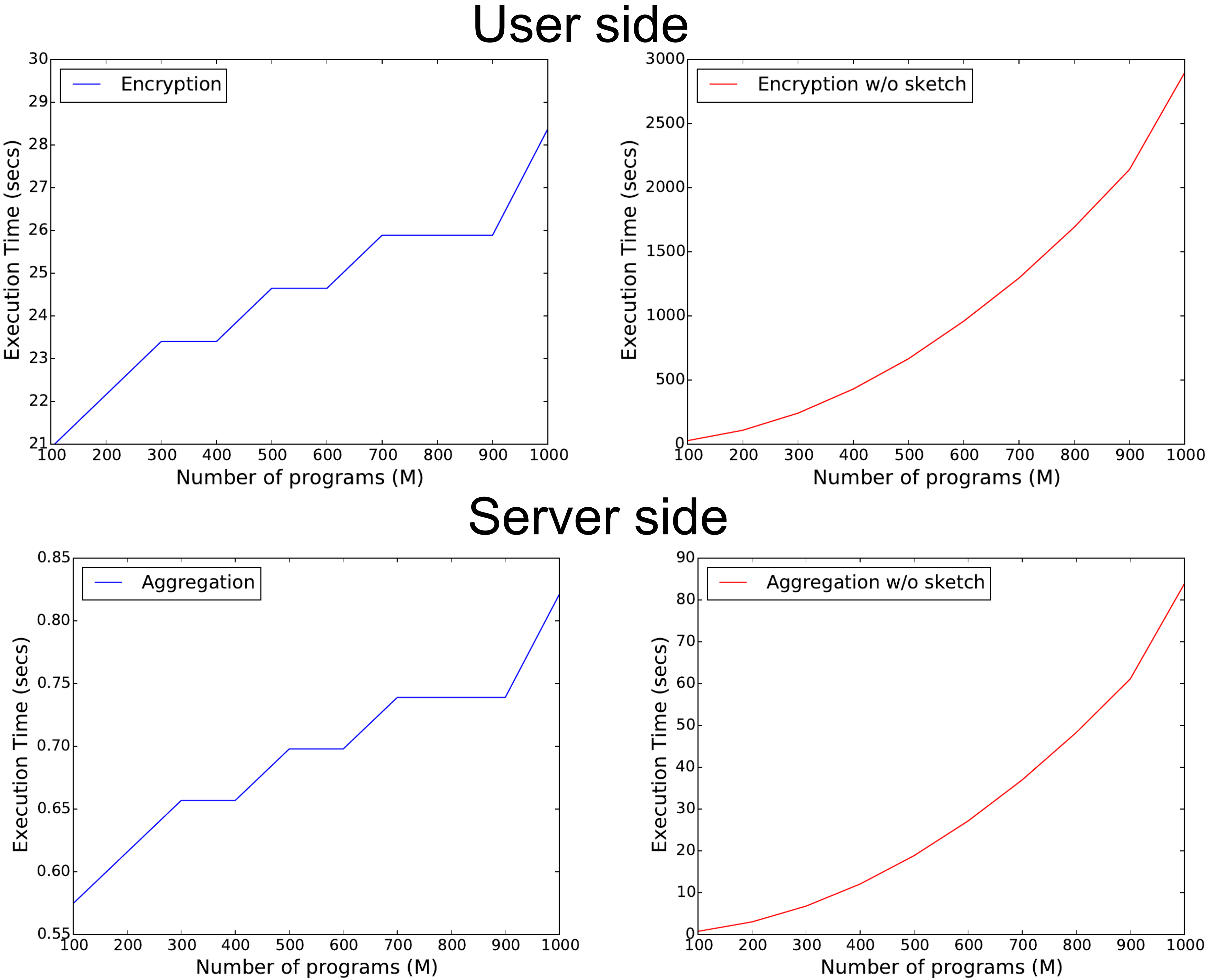
In the honest-but-curious model under the CDH assumption

Prototype implementation:

Tally as a Node.js web server

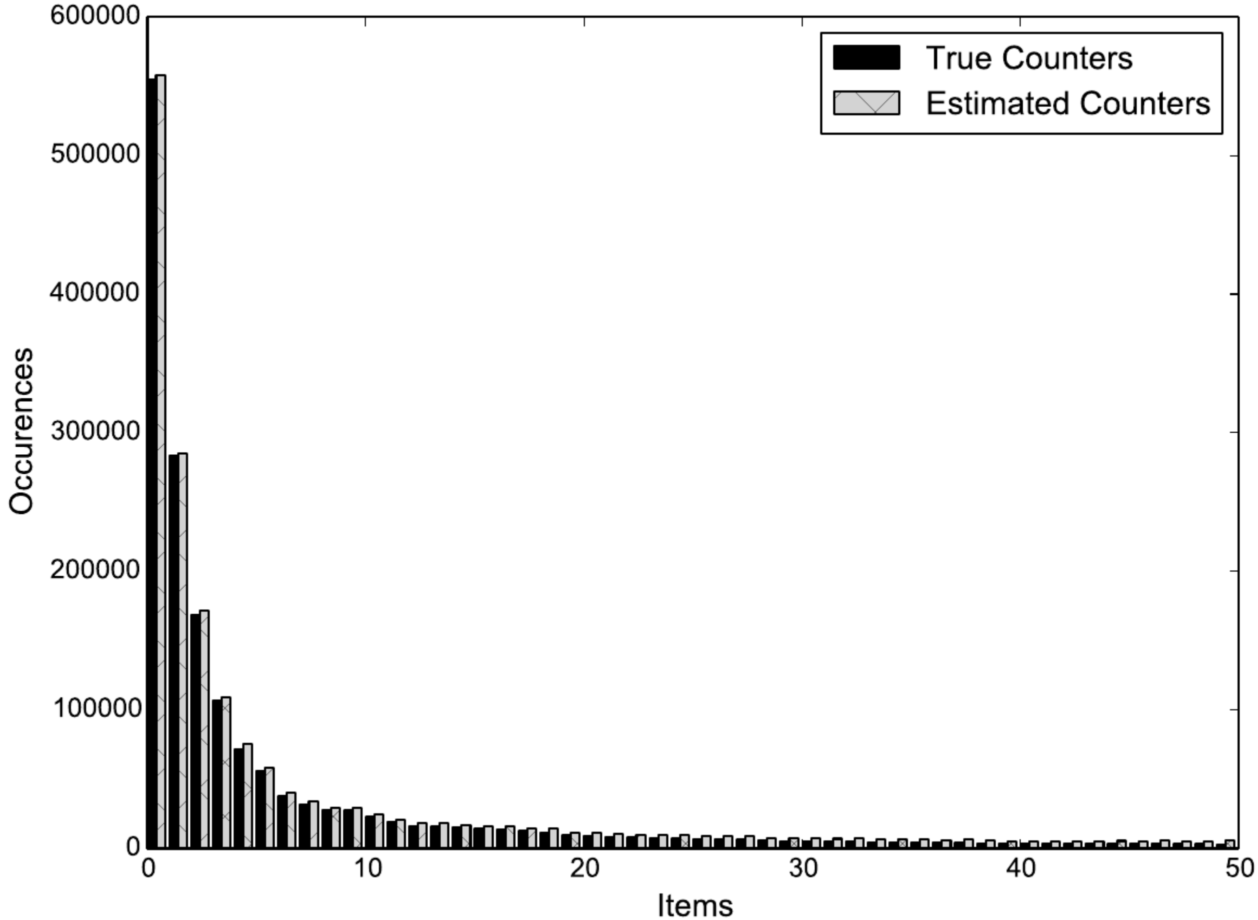
Users run in the browser or as a mobile cross-platform application (Apache Cordova)

Transparency, ease of use, ease of deployment



User side
Encryption
Encryption w/o sketch
Execution Time (secs)
Number of programs (M)
Server side
Aggregation
Aggregation w/o sketch
Execution Time (secs)
Number of programs (M)

Accuracy



True Counters
Estimated Counters
Occurences
Items
600000
500000
400000
300000
200000
100000
0
0
10
20
30
40
50

Tor Hidden Services

Aggregate statistics about the number of hidden service descriptors from multiple HSDirs

Median statistics to ensure robustness

Problem: Computation of statistics from collected data can potentially de-anonymize individual Tor users or hidden services

Private Tor Statistics?

We rely on:

- A set of authorities
- A homomorphic public-key scheme (AH-ECC)
- Count-Sketch (a variant of CMS)

Setup phase

- Each authority generates their public and private key
- A group public key is computed

Private Tor Statistics?

Each HSDir (router) builds a Count-Sketch, inserts its values, encrypts it, sends it to a set of authorities

The authorities:

- Add the encrypted sketches element-wise to generate one sketch characterizing the overall network traffic
- Execute a divide and conquer algorithm on this sketch to estimate the median

How we do it (1/2)

The range of the possible values is known

On each iteration, the range is halved and the sum of all the elements on each half is computed

Depending on which half the median falls in, the range is updated and again halved

Process stops once the range is a single element

How we do it (2/2)

Output privacy:

- Volume of reported values within each step is leaked
- Provide *differential privacy* by adding Laplacian noise to each intermediate value

Evaluating

Experimental setup:

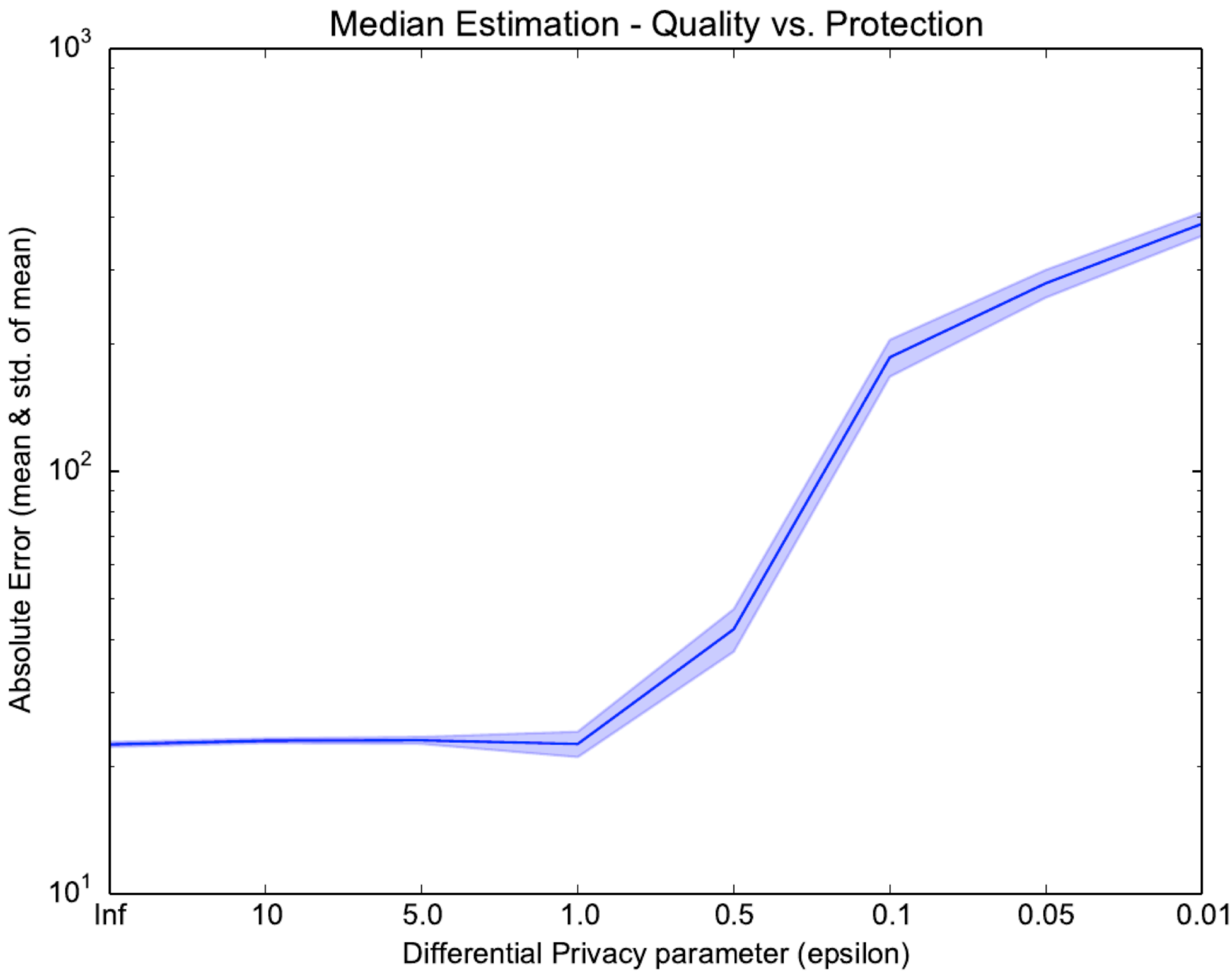
1200 samples from a mixture distribution

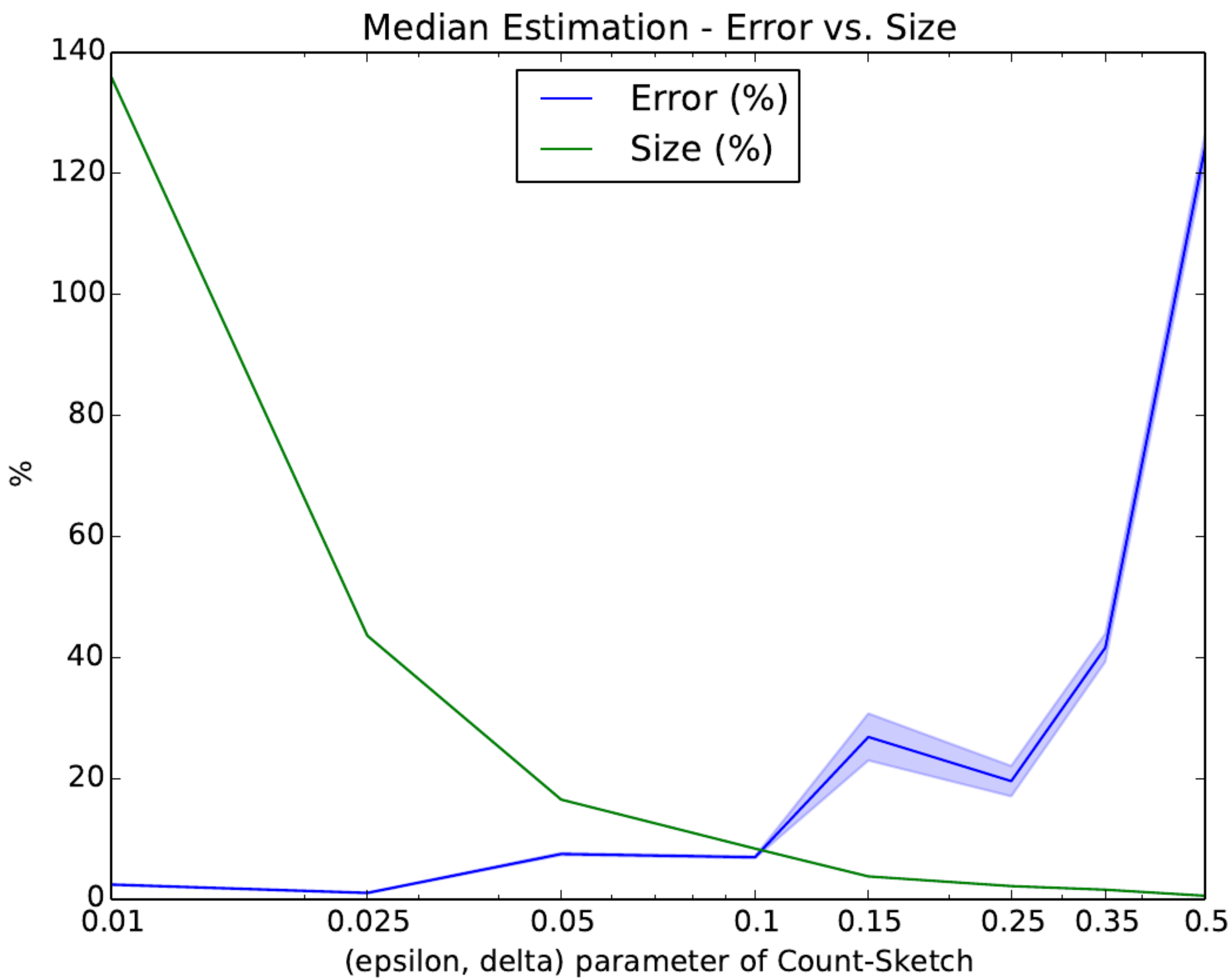
Range of values in [0,1000]

Performance evaluation:

Python implementation (*petlib*)

1 ms to encrypt a sketch (of size 165) for each HSDir and 1.5 sec to aggregate 1200 sketches


Median Estimation - Quality vs. Protection
Absolute Error (mean & std. of mean)
10^3
10^2
10^1
Inf
10
5.0
1.0
0.5
0.1
0.05
0.01
Differential Privacy parameter (epsilon)



Median Estimation - Error vs. Size
Error (%)
Size (%)
%
140
120
100
80
60
40
20
0
0.01
0.025
0.05
0.1
0.15
0.25
0.35
0.5
(epsilon, delta) parameter of Count-Sketch

Collaborative Threat Mitigation

Collaborative Anomaly Detection

Anomaly detection is hard

- Suspicious activities deliberately mimic normal behavior
- But, malevolent actors often use same resources

Wouldn't it be better if organizations collabo

- It's a w

> **"It is the policy of the United States Government to increase the volume, timelines, and quality of cyber threat information shared with U.S. private sector entities so that these entities may better protect and defend themselves against cyber attacks."**
>
> **Barack Obama**
> **2013 State of the Union Address**

Problems with Collaborations

Trust

Will others leak my data?

Legal Liability

Will I be sued for sharing customer data?

Will others find me negligible?

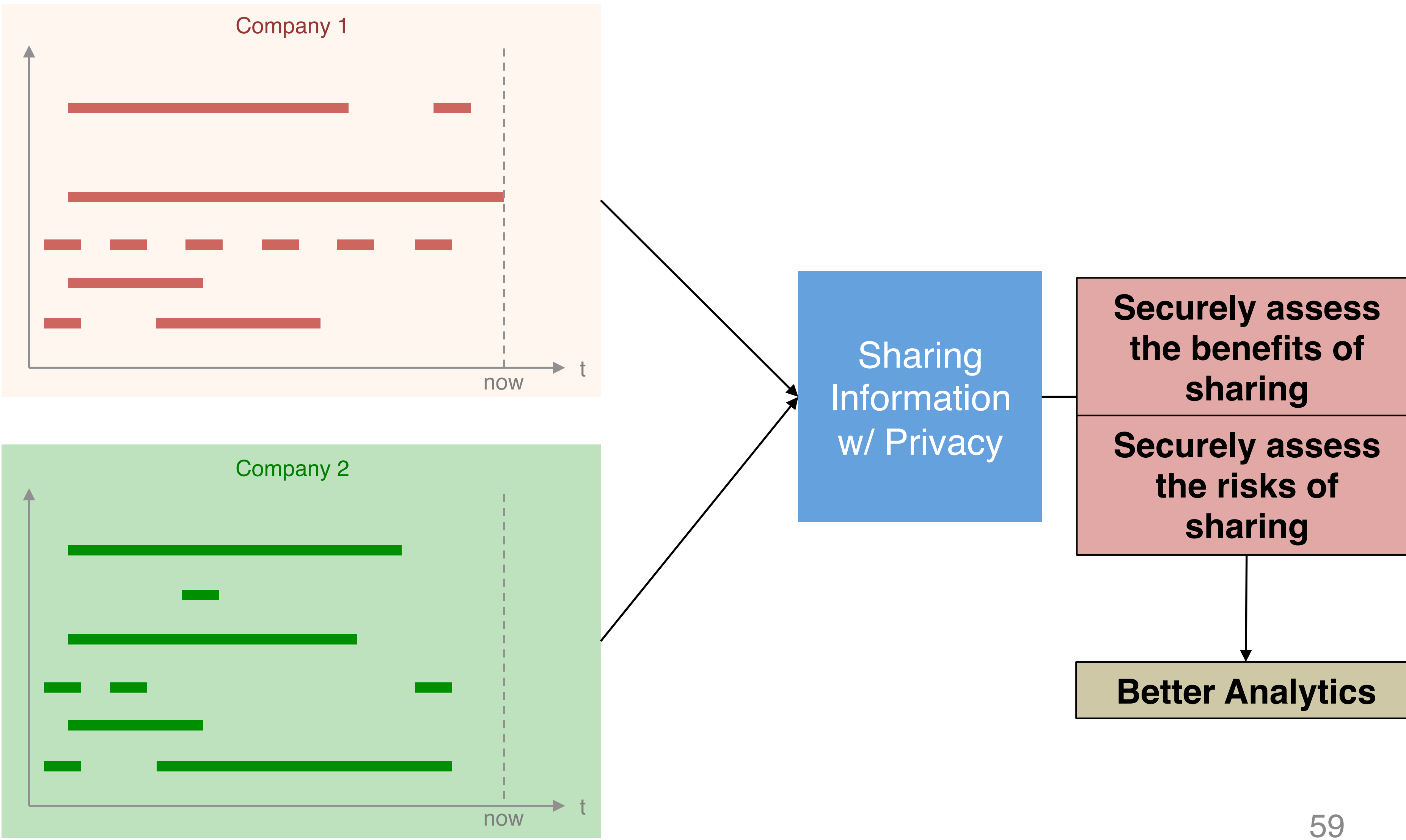
Competitive concerns

Will my competitors outperform me?

Shared data quality

Will data be reliable?

Solution Intuition [FDB15]



1. Estimate Benefits

What are good **indicators** of the fact that sharing will be beneficial?

- Many attackers in common?
- Many similar attacks in common?
- Many correlated attacks in common?

2. Select Partners

How do I **choose** who to collaborate with?

- Collaborate with the top-k?
- Collaborate if benefit above threshold?
- Hybrid?

3. Merge

Once we partnered up, **what** do we share?

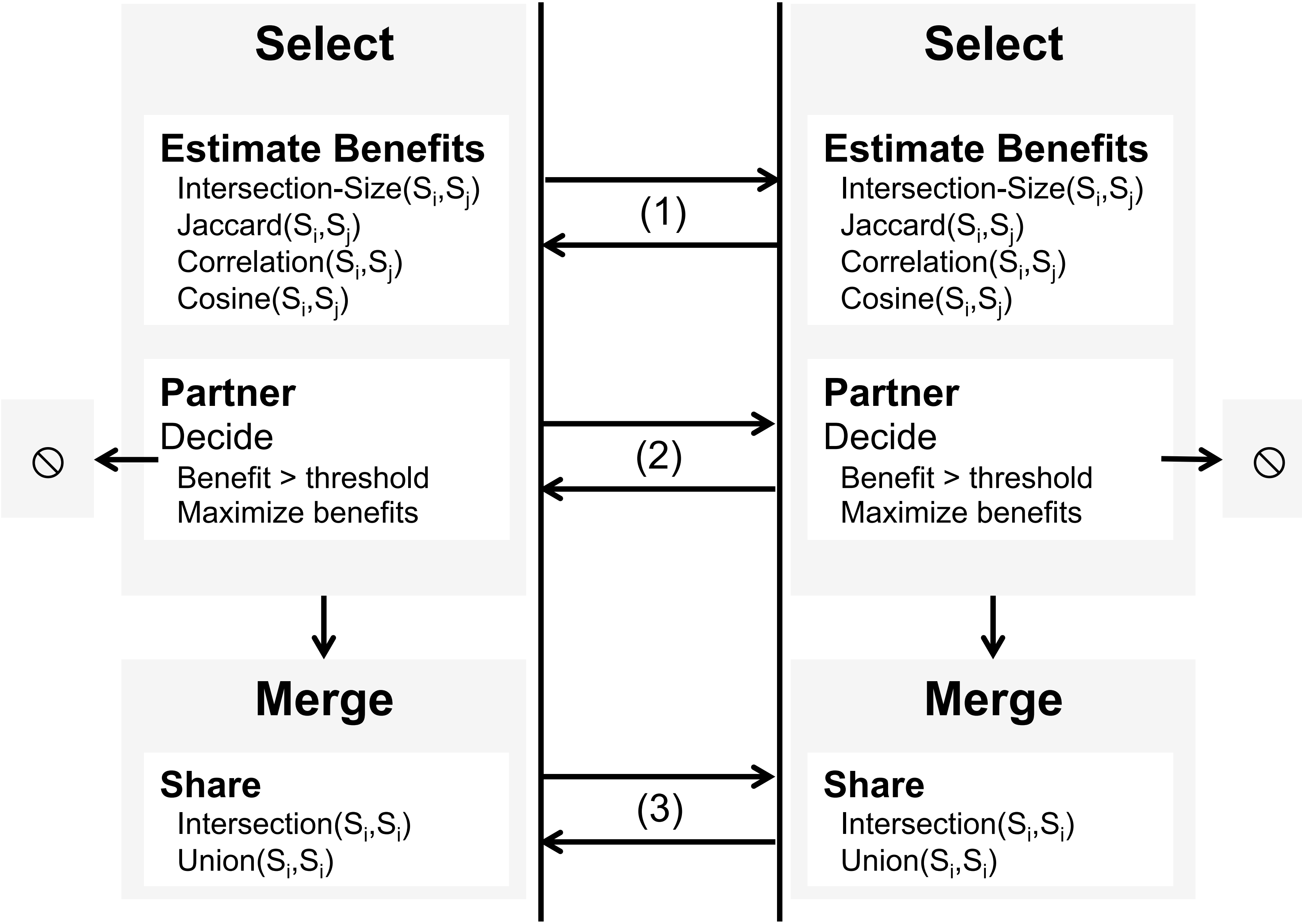
- Everything?
- Just what we have in common?
- Just information about attacks or also metadata?

System Model

Network of n entities $\{V_i\}$ (for $i=1,\ldots,n$)

Each V_i holds a dataset S_i of suspicious events

- E.g., events in the form $\langle$IP, time, port$\rangle$ as observed by a firewall or an IDS


Select
Estimate Benefits
Intersection-Size(S_i,S_j)
Jaccard(S_i,S_j)
Correlation(S_i,S_j)
Cosine(S_i,S_j)
Partner
Decide
Benefit > threshold
Maximize benefits
Merge
Share
Intersection(S_i,S_i)
Union(S_i,S_i)
(1)
(2)
(3)
Select
Estimate Benefits
Intersection-Size(S_i,S_j)
Jaccard(S_i,S_j)
Correlation(S_i,S_j)
Cosine(S_i,S_j)
Partner
Decide
Benefit > threshold
Maximize benefits
Merge
Share
Intersection(S_i,S_i)
Union(S_i,S_i)

Privacy-preserving benefit estimation

Metric	Operation	Private Protocol
Intersection-Size	$\lvert S_i \cap S_j \rvert$	Private Set Intersection Cardinality (PSI-CA)
Jaccard	$\frac{\lvert S_i \cap S_j \rvert}{\lvert S_i \cup S_j \rvert}$	Private Jaccard Similarity (PJS)
Pearson	$\sum_{l=1}^{N} \frac{(s_{i_l} - \mu_i)(s_{j_l} - \mu_j)}{N\sigma_i\sigma_j}$	Garbled Circuits (2PC)
Cosine	$\frac{\vec{S}_i \vec{S}_j}{\lVert\vec{S}_i\rVert \, \lVert\vec{S}_j\rVert}$	Private Cosine Similarity (PCS)

Privacy-preserving data sharing

Metric	Operation	Private Protocol
Intersection	$\lvert S_i \cap S_j \rvert$	Private Set Intersection (PSI)
Intersection with Associated Data	$\{\langle \mathrm{IP}, \mathrm{time}, \mathrm{port} \rangle \mid \mathrm{IP} \in S_i \cap S_j\}$	Private Set Intersection w/ Data Transfer (PSI-DT)
Union with Associated Data	$\{\langle \mathrm{IP}, \mathrm{time}, \mathrm{port} \rangle \mid \mathrm{IP} \in S_i \cup S_j\}$	-

The Road Ahead…

This slide is intentionally left blank