Impact Tracing: Identifying the Culprit of Misinformation in Encrypted Messaging Systems

Zhongming Wang, Tao Xiang, Xiaoguo Li, Biwen Chen, Guomin Yang, Chuan Ma, and Robert H. Deng

NDSS Symposium 2025

重庆大学
CHONGQING UNIVERSITY

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Content Moderation for EEMSs

End-to-end encrypted messaging systems (EEMS):

Only the END users can read the messages.

Content Moderation for EEMSs

End-to-end encrypted messaging systems (EEMS):

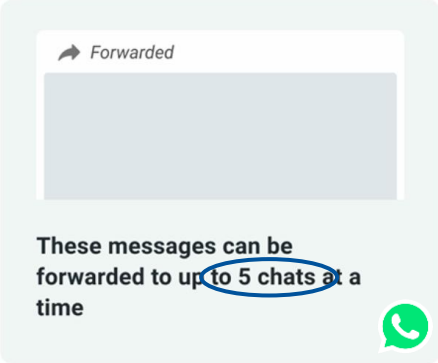
Only the END users can read the messages.

Problematic messages proliferates in EEMSs.

End-to-end encryption obstructs content moderation :(

Message Forwarding & Tracing

Misinformation propagate rapidly through forwarding.



Message Forwarding & Tracing

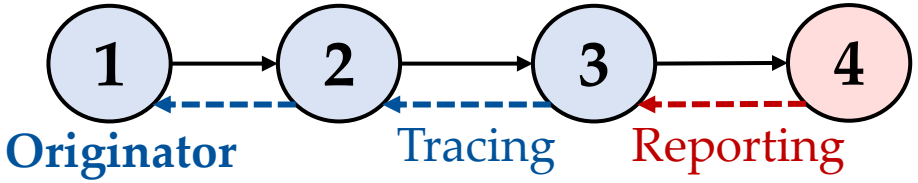
Misinformation propagate rapidly through forwarding.

Message Forwarding & Tracing

Misinformation propagate rapidly through forwarding.

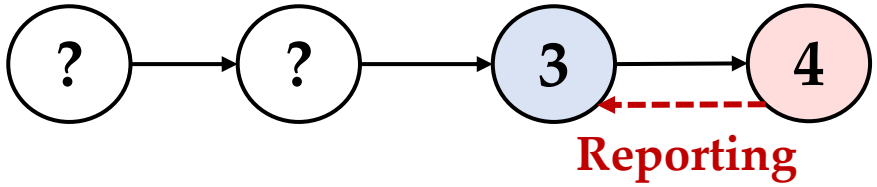
Traceability enables tracing after user reporting.

- The platform can disclose the dissemination path.



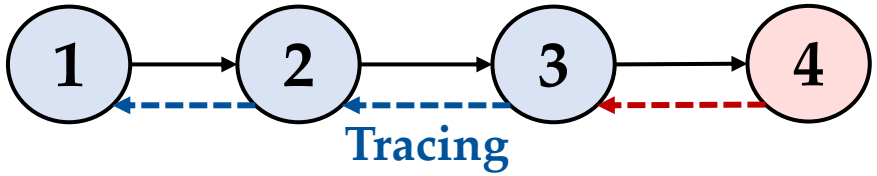
https://faq.whatsapp.com/1053543185312573/?helpref=uf_share

Tracing Policies



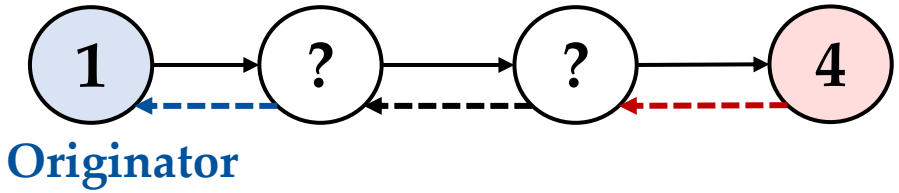
Message Franking

[Facebook17,RWC], [GLR17,CRYPTO], [TGL+19,CRYPTO], [GPE25,NDSS], etc.



Message Traceback

[TMR19,CCS], [KTW22,ESORICS]



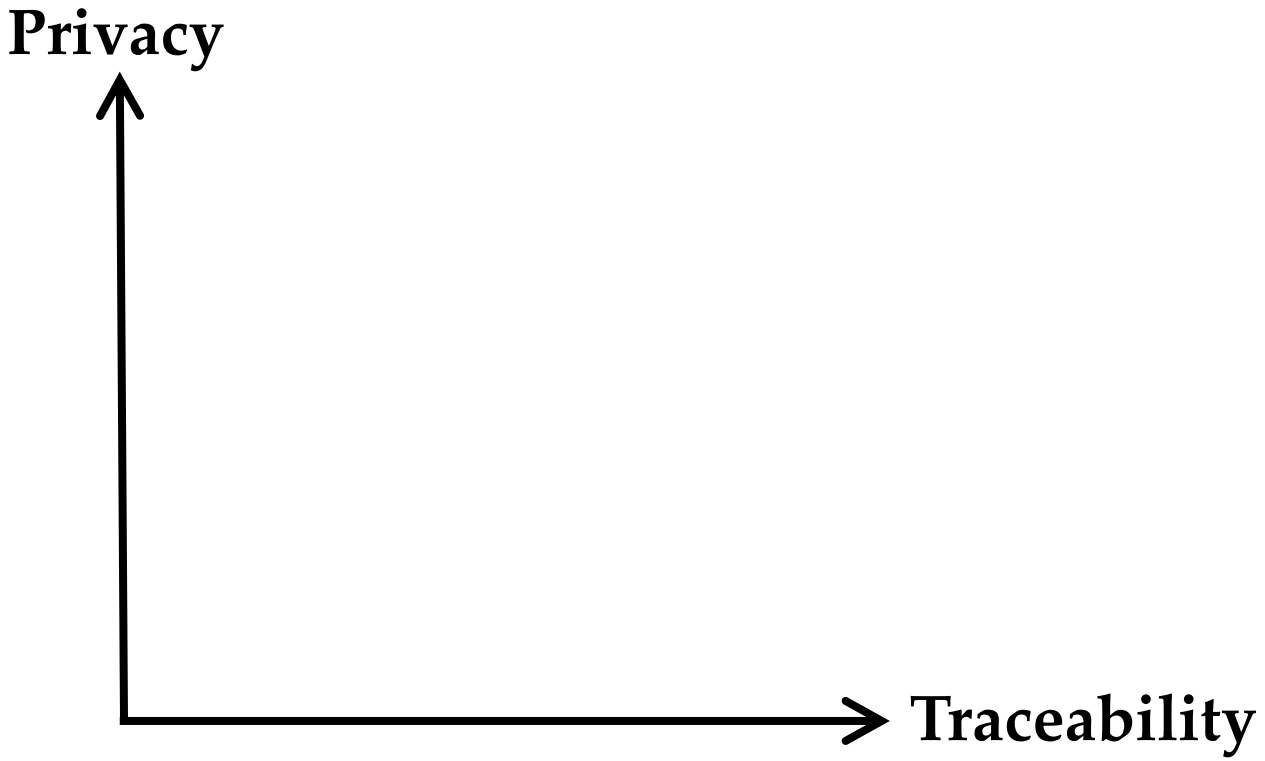
Source Tracing

[PEB21,CCS], [IAV22,Usenix Sec.], [LRTY22,NDSS], [BGJP23,EUROCRYPT], [BE24,PETS], etc.

Which part of the dissemination path is tracked during tracing?

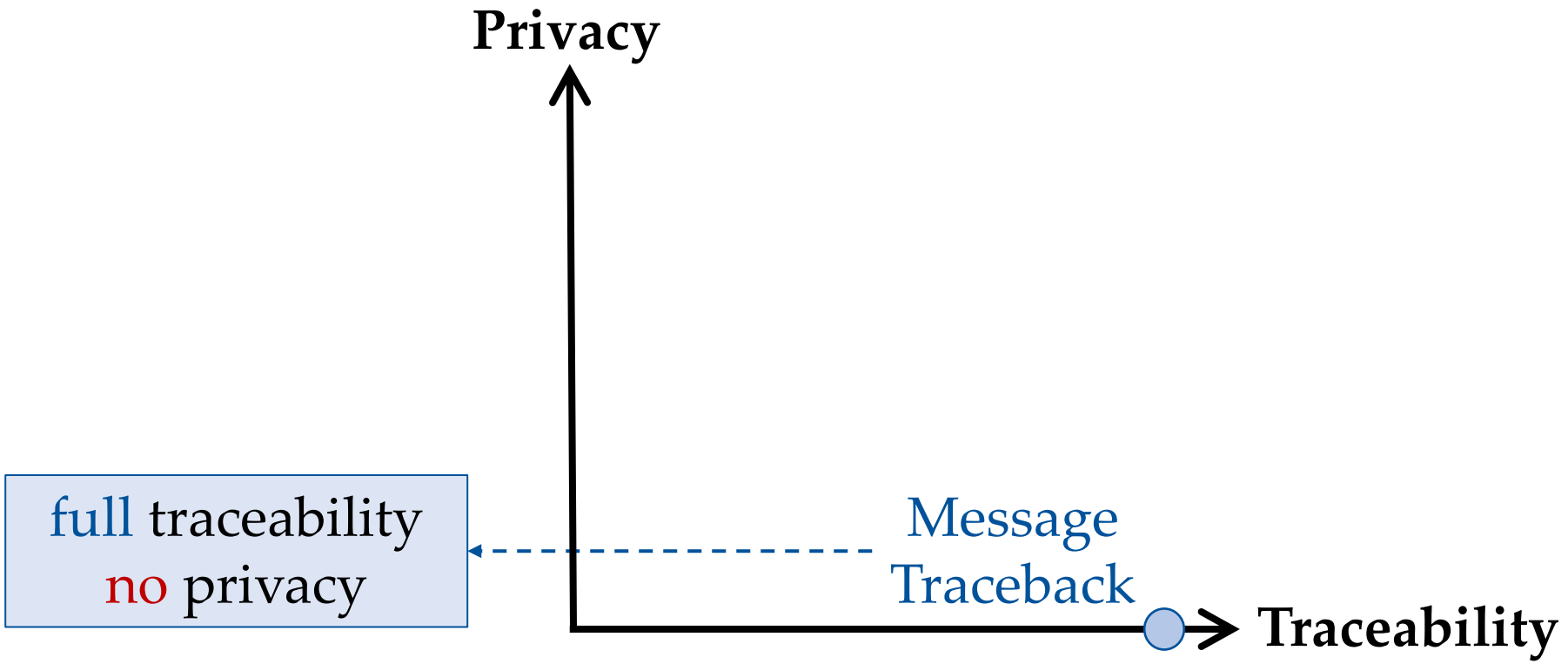
Traceability vs. Privacy

- **Traceability:** Reveal the culprits of spreading misinformation.
- **Privacy:** Reveal nothing about forwarding path.



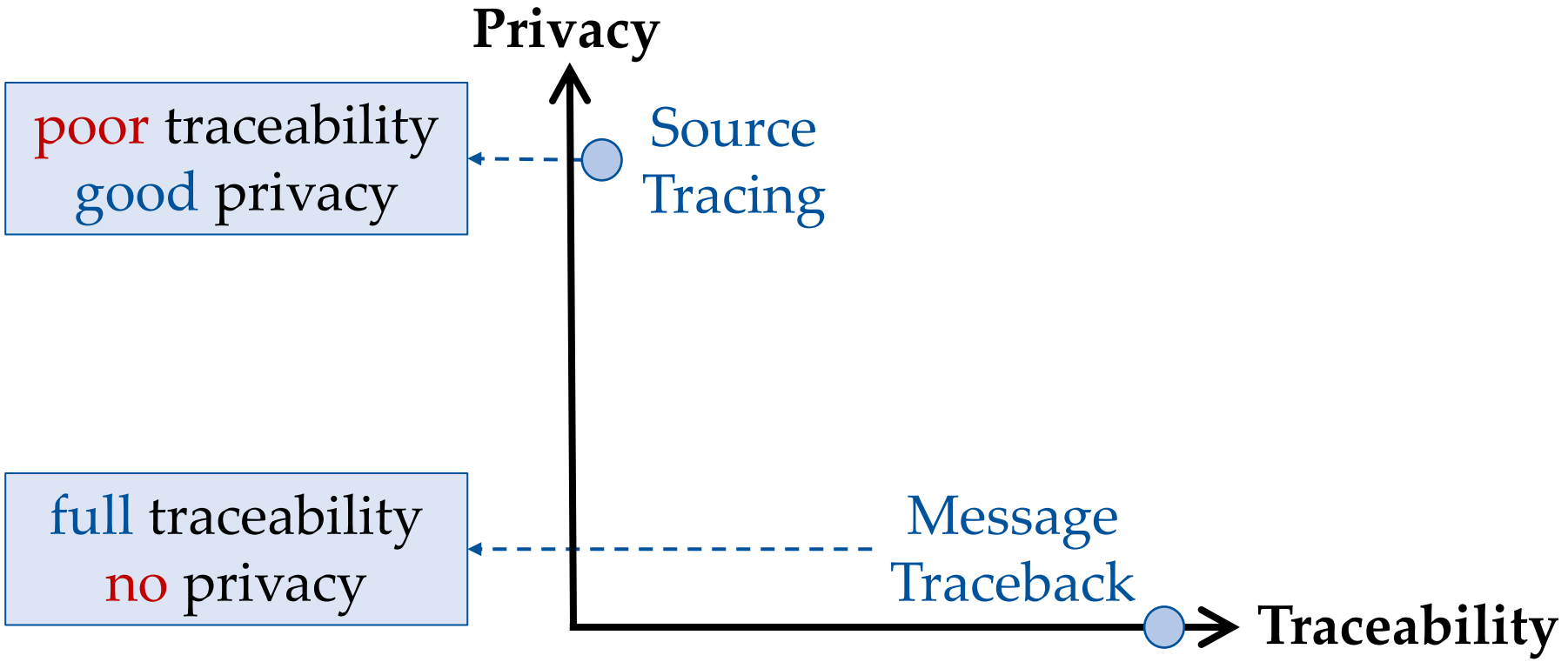
Traceability vs. Privacy

- **Traceability:** Reveal the culprits of spreading misinformation.
- **Privacy:** Reveal nothing about forwarding path.



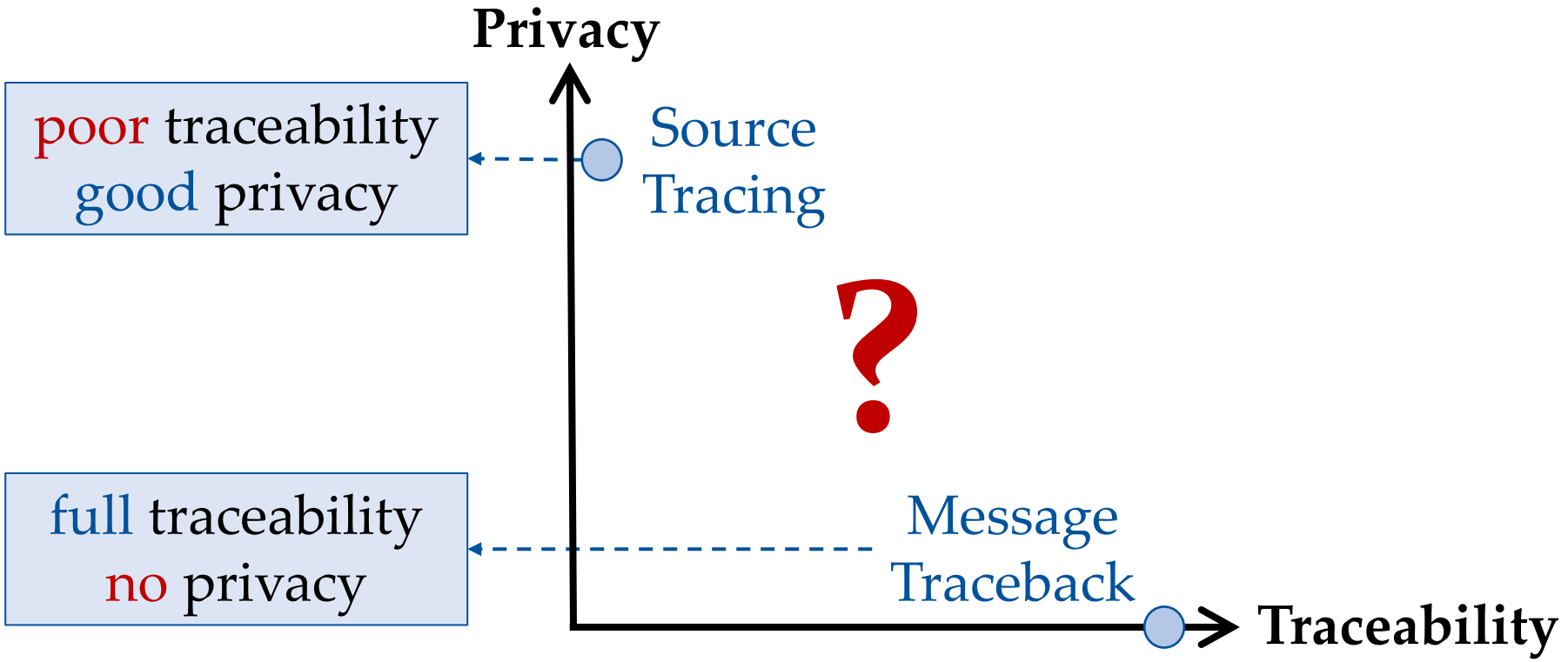
Traceability vs. Privacy

- **Traceability:** Reveal the culprits of spreading misinformation.
- **Privacy:** Reveal nothing about forwarding path.



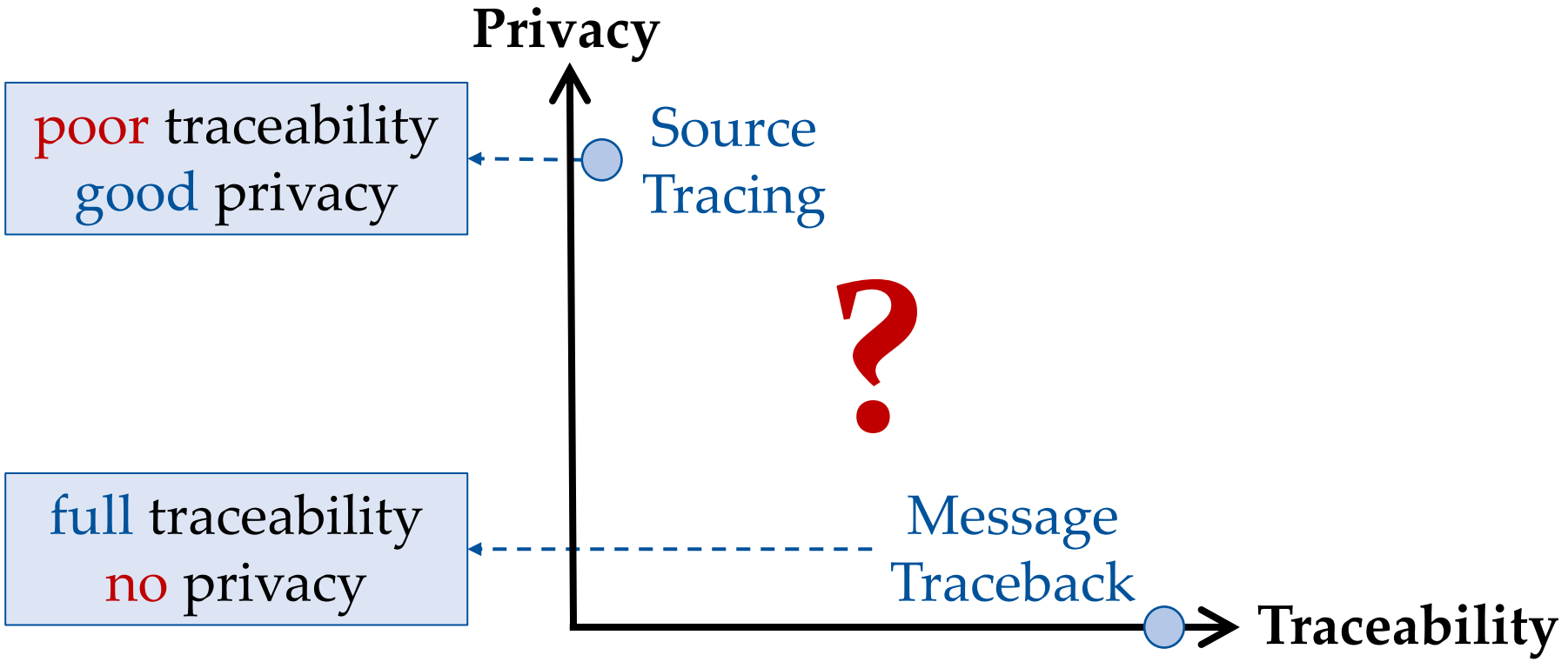
Traceability vs. Privacy

- **Traceability:** Reveal the culprits of spreading misinformation.
- **Privacy:** Reveal nothing about forwarding path.



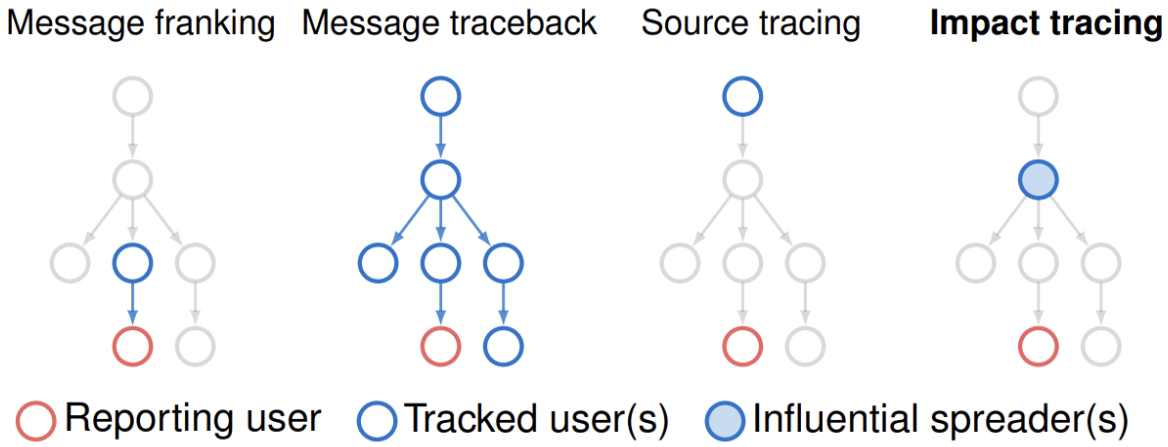
Traceability vs. Privacy

Is there a tracing policy that balances traceability and privacy, but also provides practical values to EEMSs?



Our Solution: Impact Tracing!

A small group of users (called influential spreaders) *significantly* contribute to spreading misinformation.



Traceability: The platform *can* identify influential spreaders.

Privacy: The platform *cannot* uncover non-influential users.

The Design: Enabling Reporting

Sender

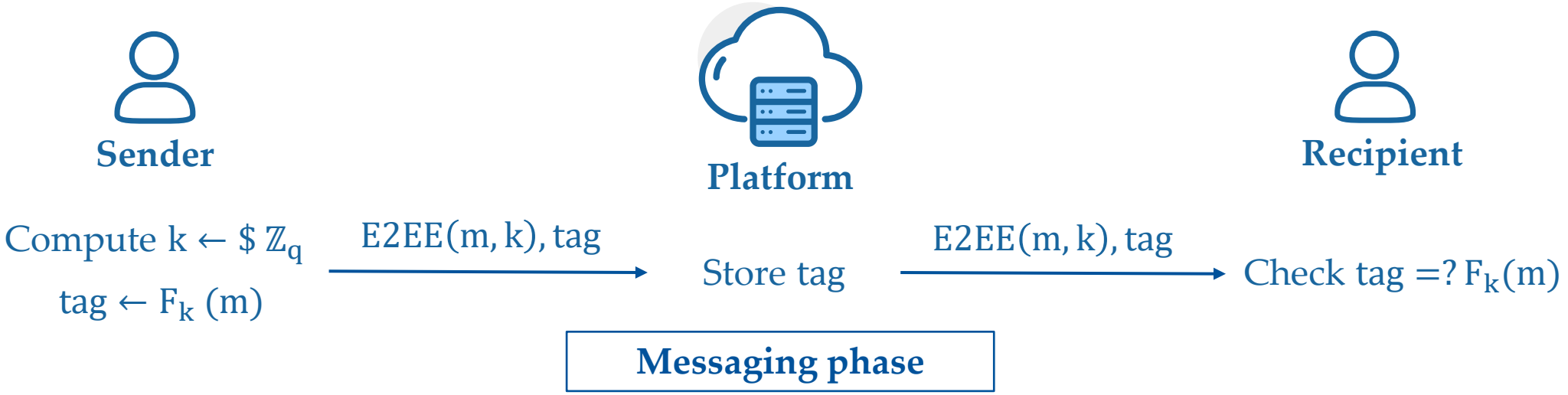
Platform

Recipient

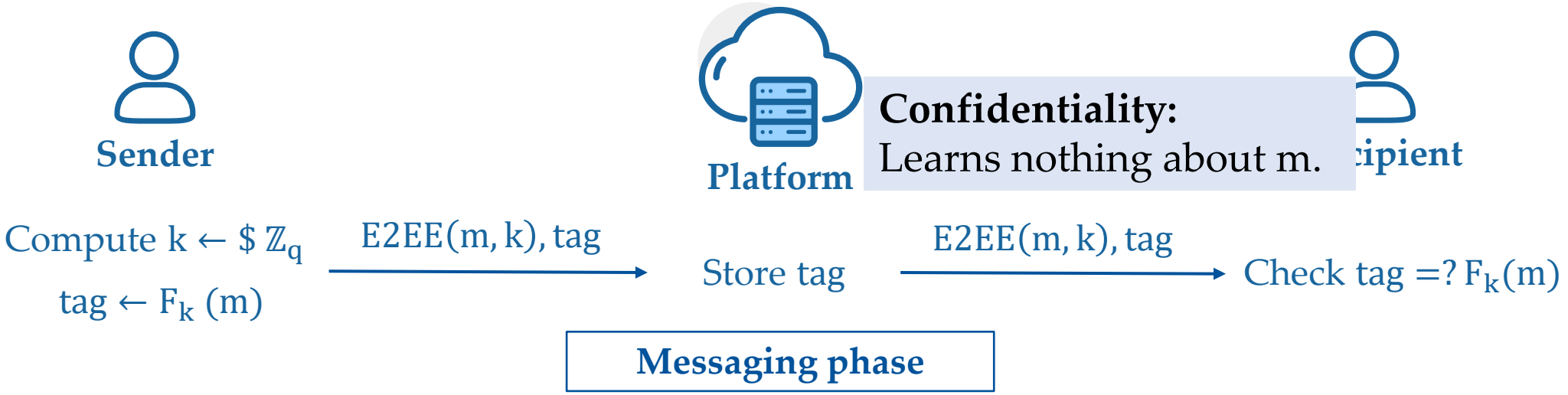
Messaging phase

Reporting phase

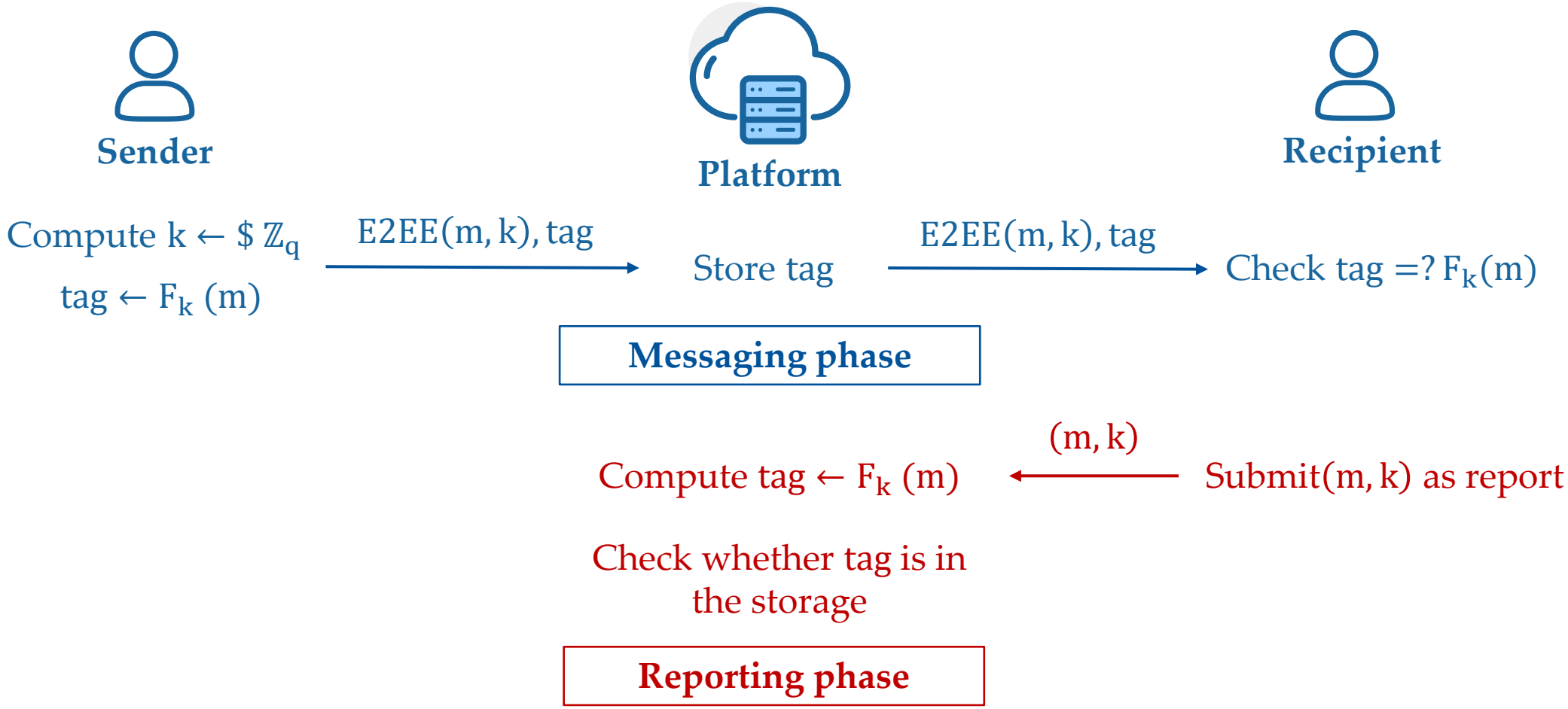
The Design: Enabling Reporting



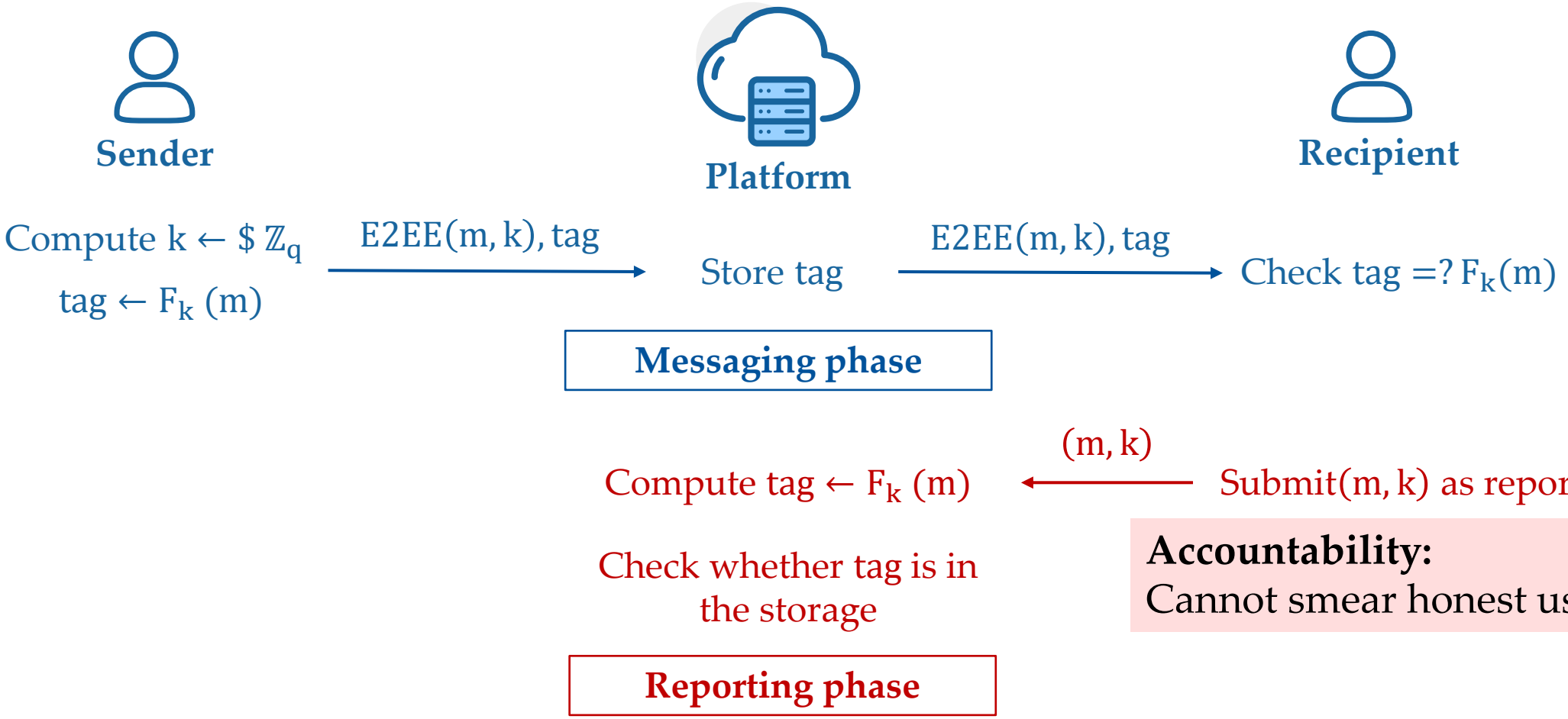
The Design: Enabling Reporting



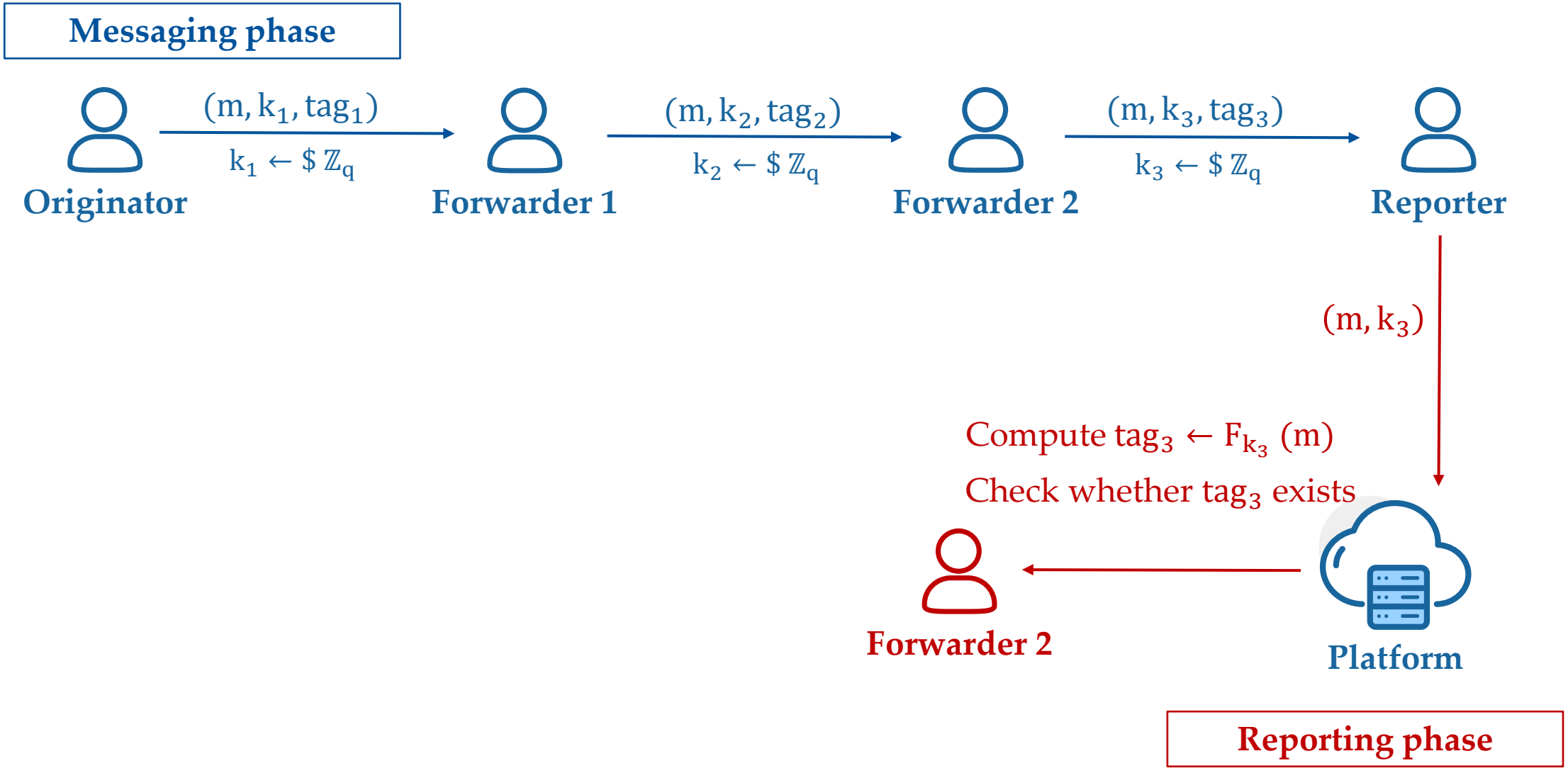
The Design: Enabling Reporting



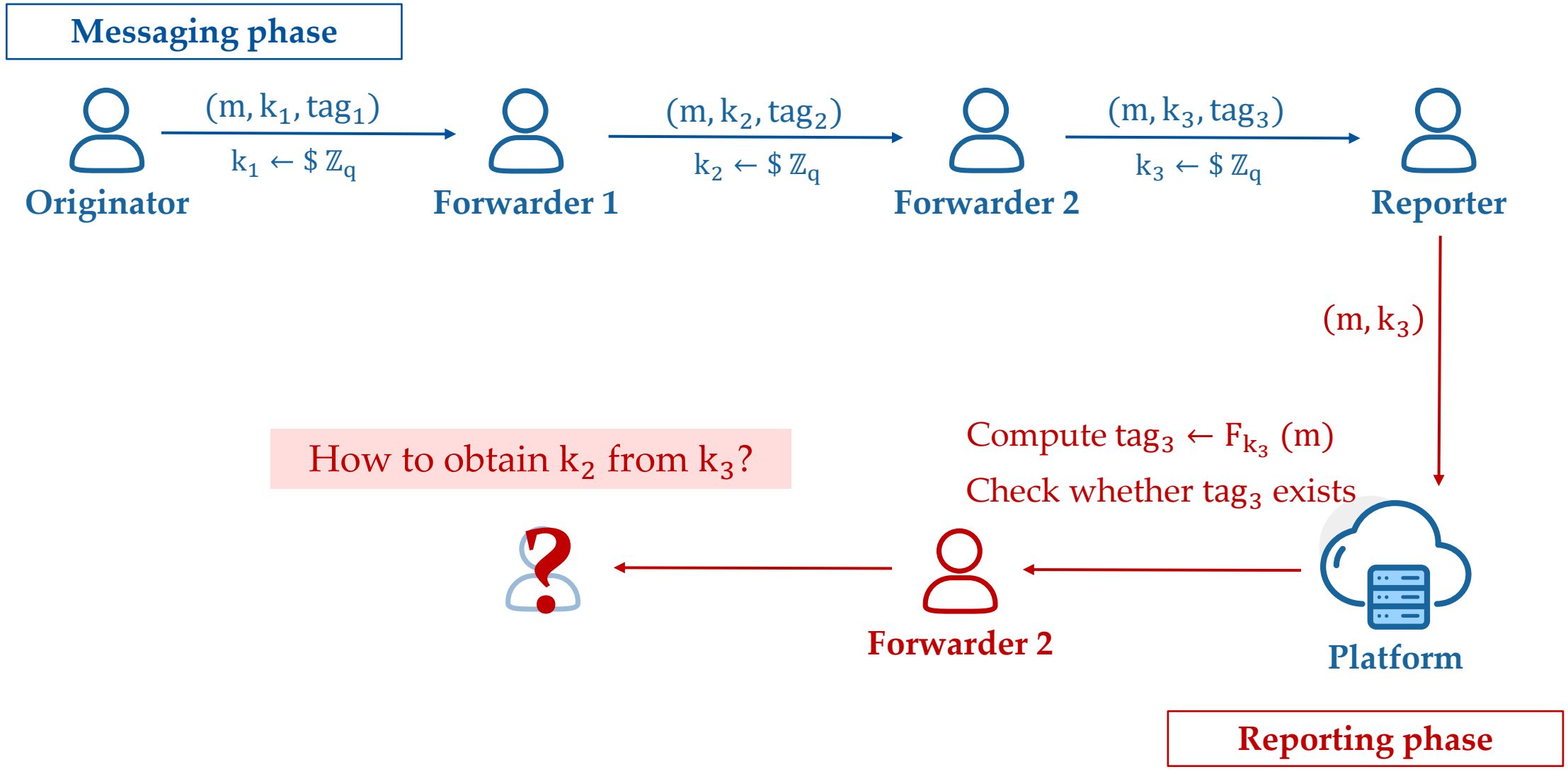
The Design: Enabling Reporting



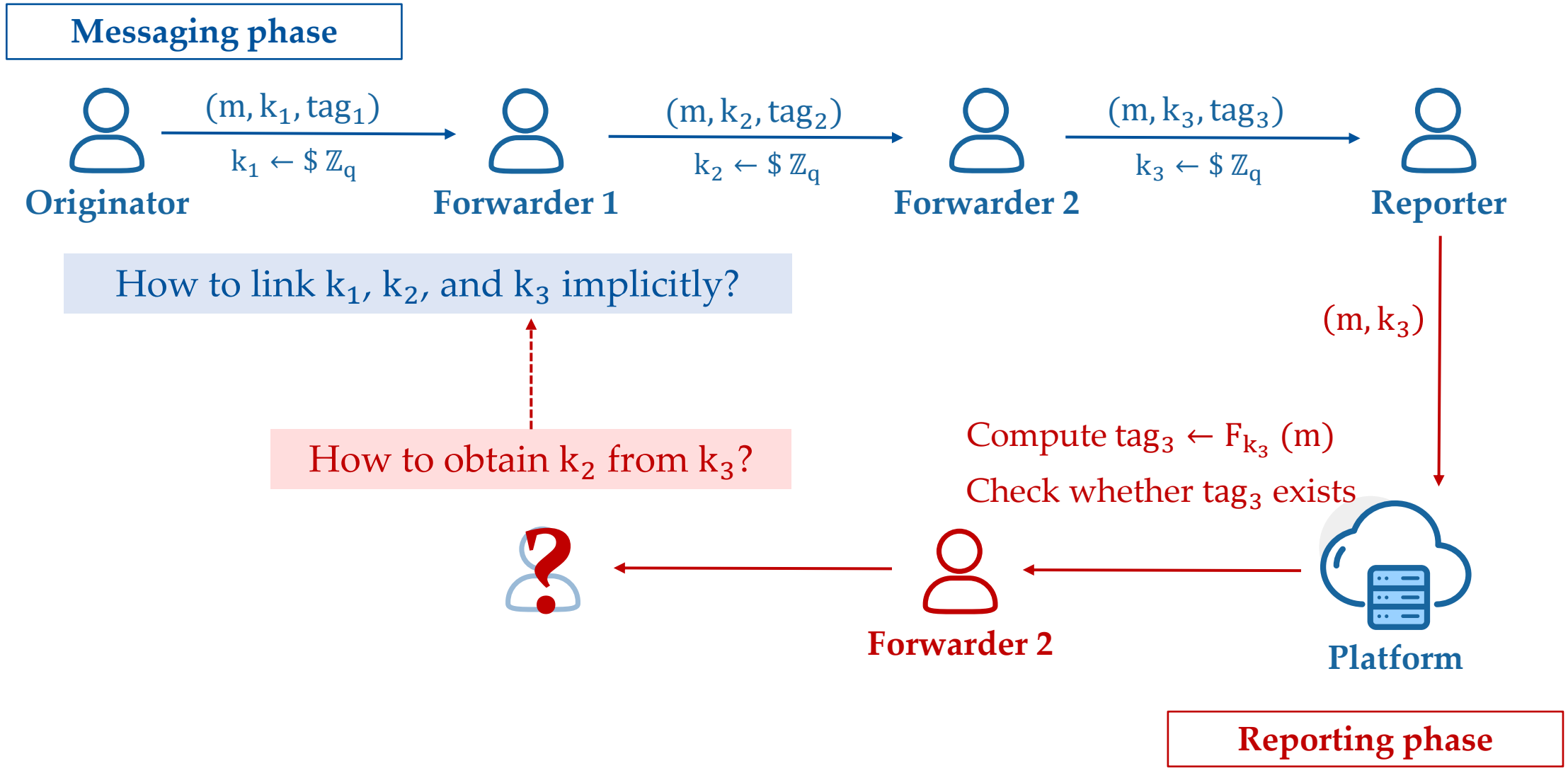
The Design: Enabling Reporting



The Design: Enabling Reporting



The Design: Enabling Reporting



The Design: Reporting then Tracing

Messaging phase

The Design: Reporting then Tracing

Messaging phase

ik: Users' secret identity key; **U:** Users' public identity; **Enc:** Block cipher

Originator

$tk_{01} \leftarrow H(ik_0||U_1)$

$k_1 \leftarrow Enc_{tk_{01}}(k_0)$

$k_0 \leftarrow \$\ \mathbb{Z}_q$

Forwarder 1

$tk_{12} \leftarrow H(ik_1||U_2)$

$k_2 \leftarrow Enc_{tk_{12}}(k_1)$

Forwarder 2

$tk_{23} \leftarrow H(ik_2||U_3)$

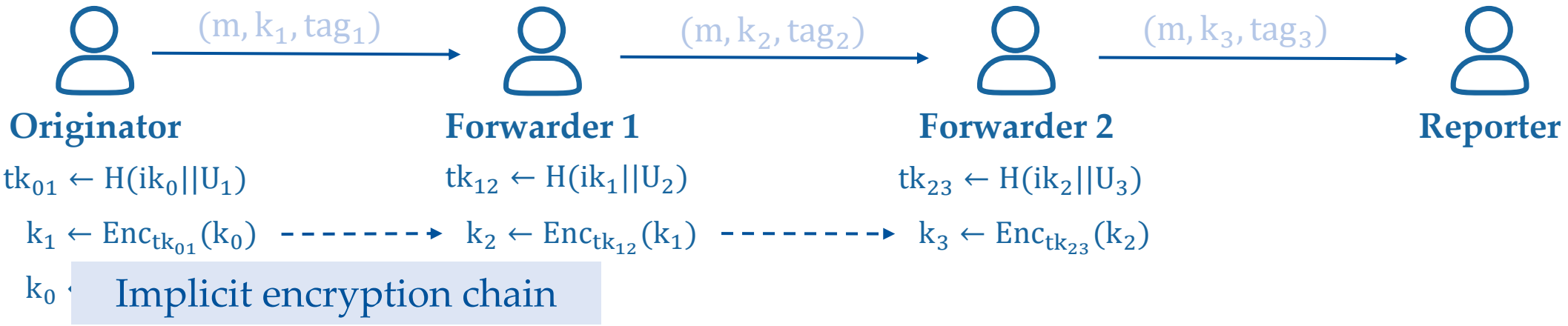
$k_3 \leftarrow Enc_{tk_{23}}(k_2)$

Reporter

The Design: Reporting then Tracing

Messaging phase

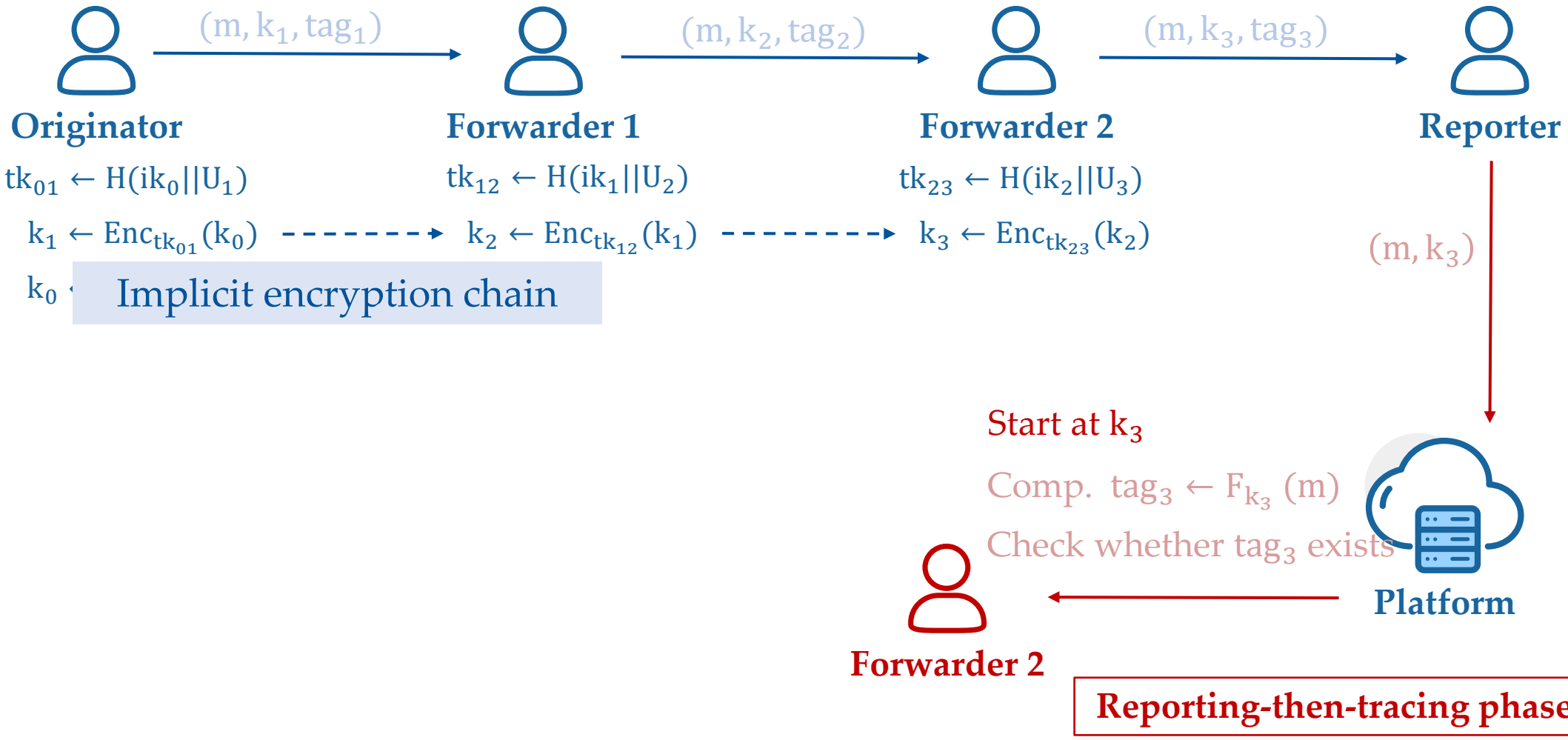
ik: Users' secret identity key; **U:** Users' public identity; **Enc:** Block cipher



The Design: Reporting then Tracing

Messaging phase

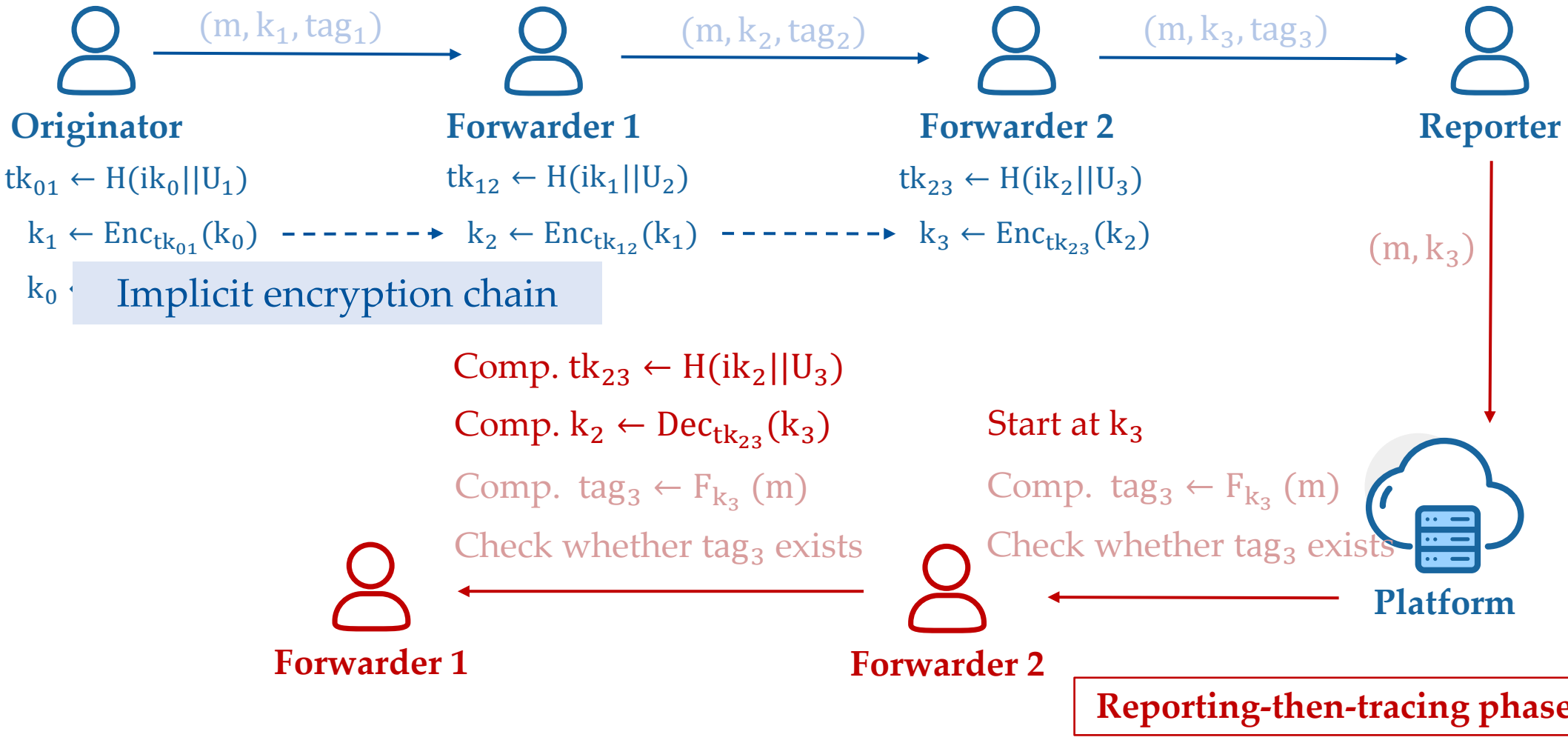
ik: Users' secret identity key; **U:** Users' public identity; **Enc:** Block cipher



The Design: Reporting then Tracing

Messaging phase

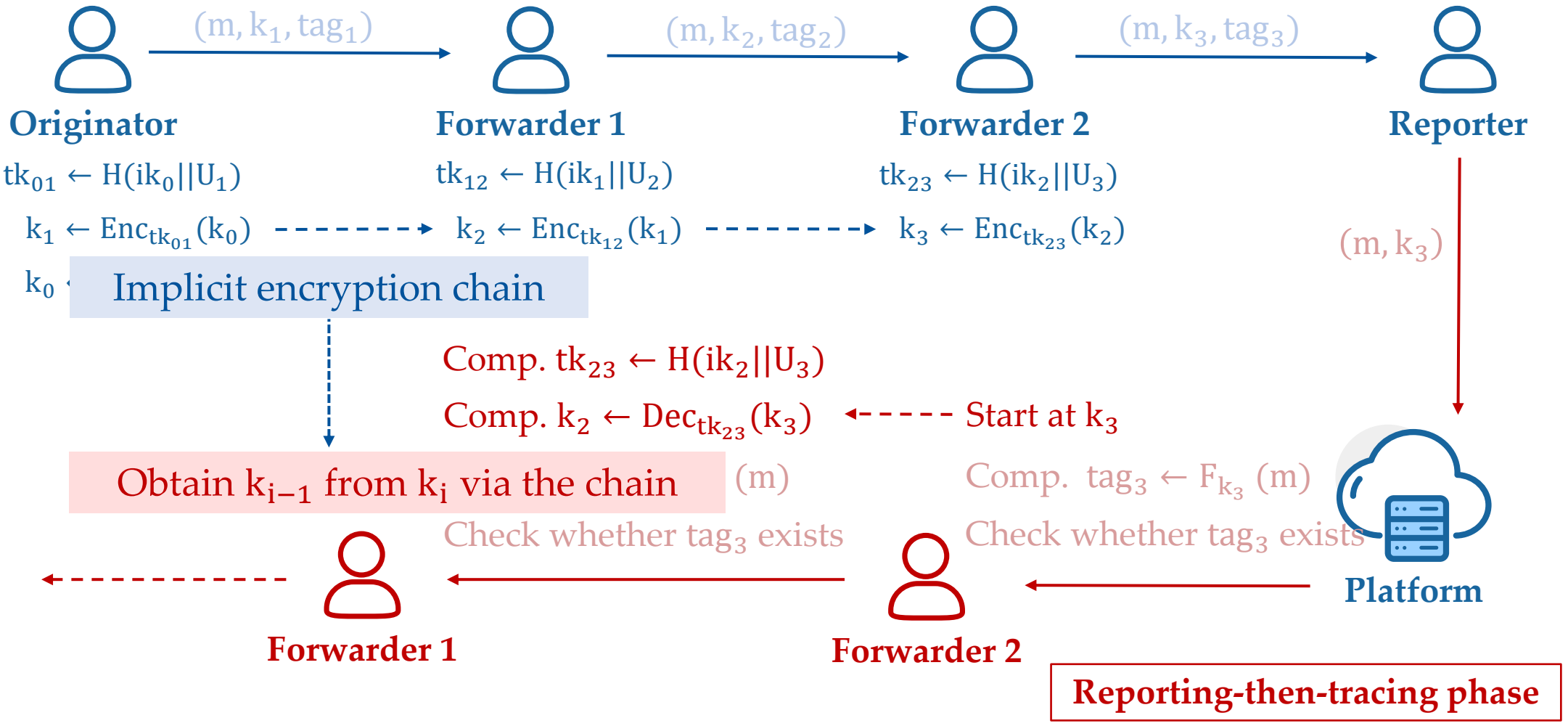
ik: Users' secret identity key; **U:** Users' public identity; **Enc:** Block cipher



The Design: Reporting then Tracing

Messaging phase

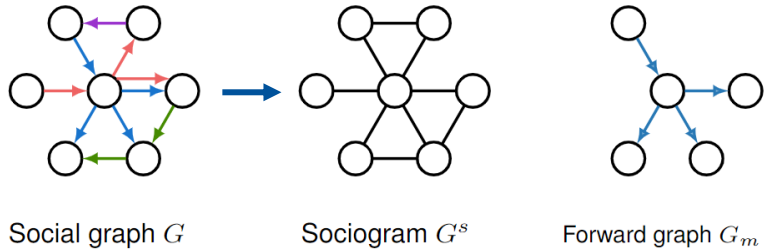
ik: Users' secret identity key; **U:** Users' public identity; **Enc:** Block cipher



The Design: Reporting then Tracing

Messaging phase:

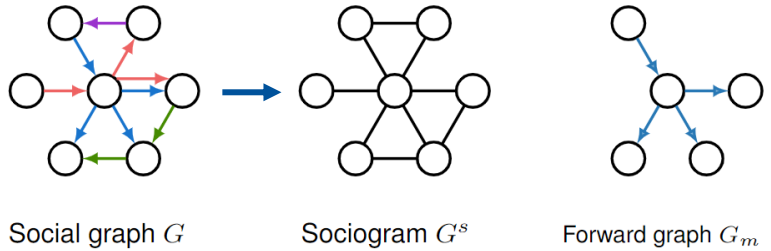
- Users forward messages.
- Platform collects sociogram & stores tags.



The Design: Reporting then Tracing

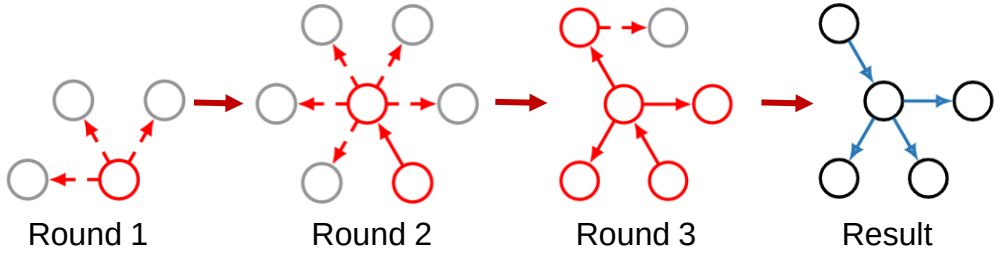
Messaging phase:

- Users forward messages.
- Platform collects sociogram & stores tags.



Reporting-then-Tracing phase:

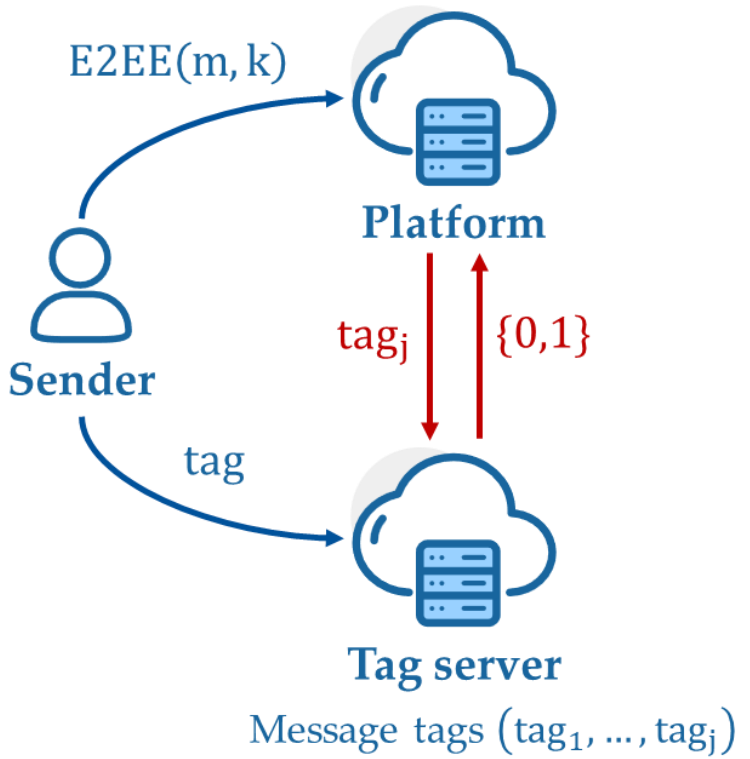
1) Compute tag key and tag.
2) Check the existence of the tag.
3) Start next round at existed vertices.



The Design: Noising the Traceback

Introduce random response to add noise.

- Two servers: Platform & Tag server

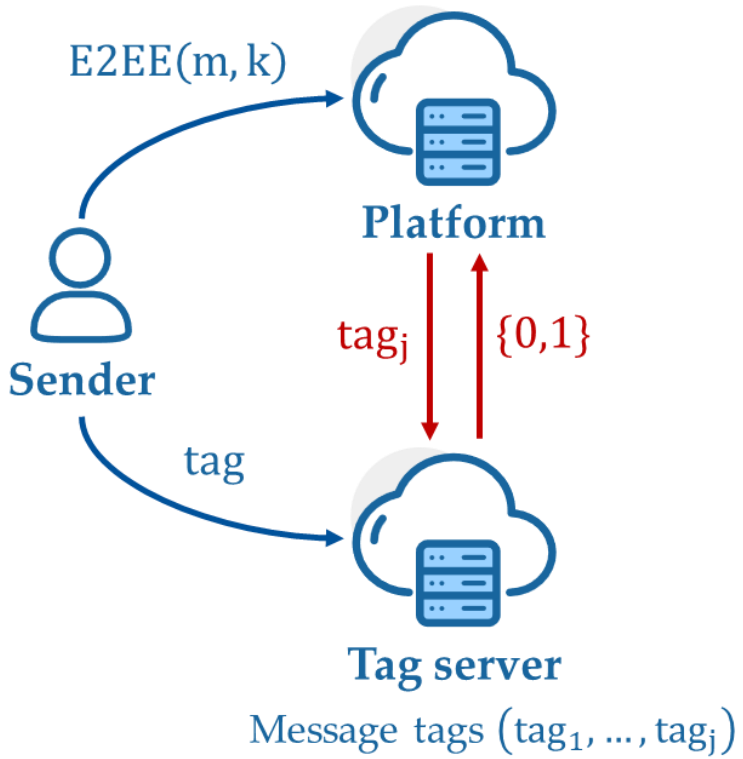


The Design: Noising the Traceback

Introduce random response to add noise.

- Two servers: Platform & Tag server
- Tag server responses a query as:

> if tag_j exists, return 1
>
> else, return 0 w.p. $1 - \psi$
>
> return 1 w.p. ψ

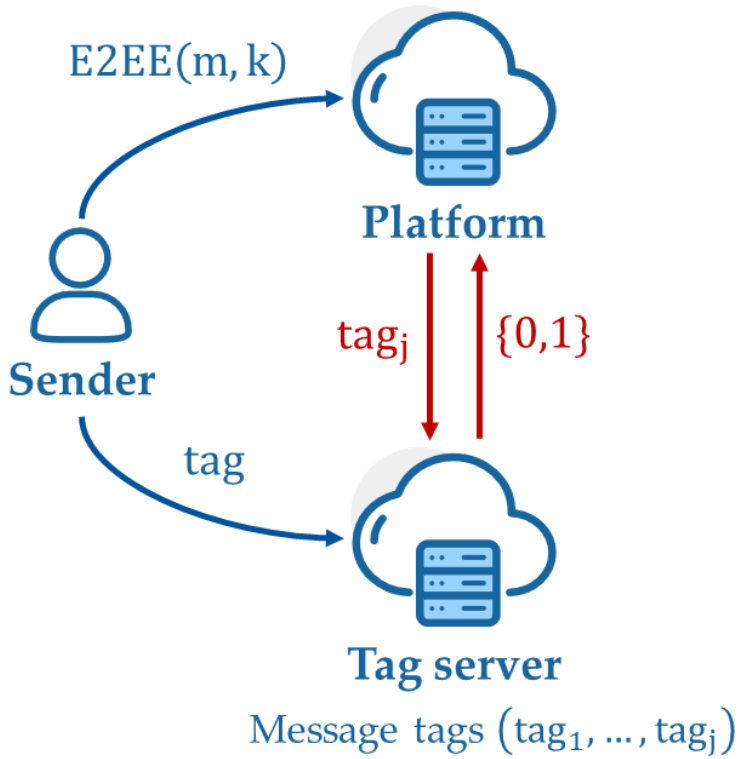


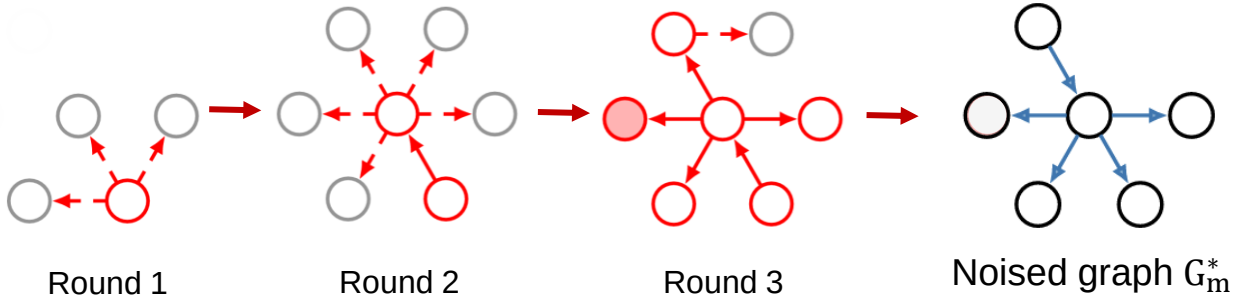
The Design: Noising the Traceback

Introduce random response to add noise.

- Two servers: Platform & Tag server
- Tag server responses a query as:

> if tag_j exists, return 1
>
> else, return 0 w.p. $1 - \psi$
>
> return 1 w.p. ψ



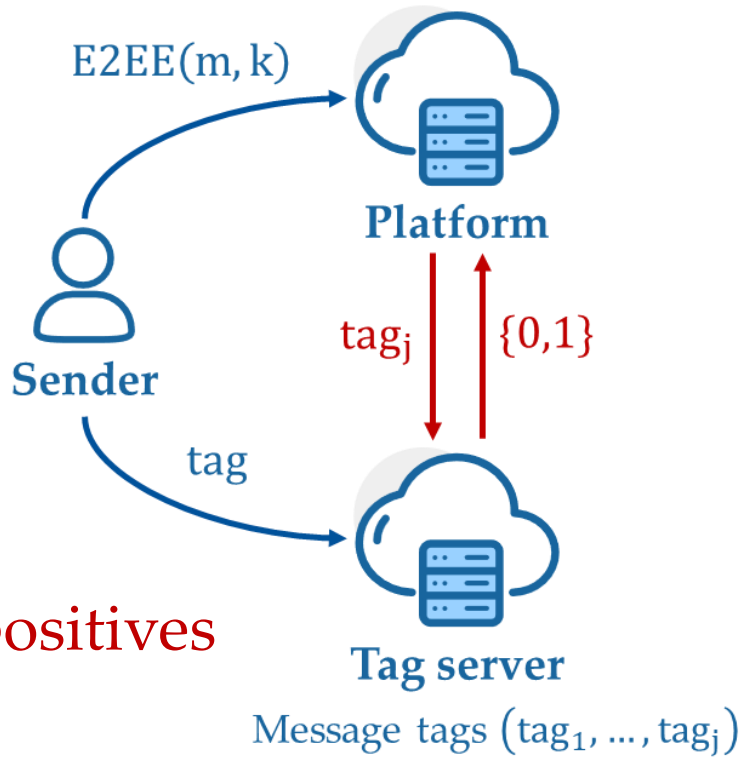


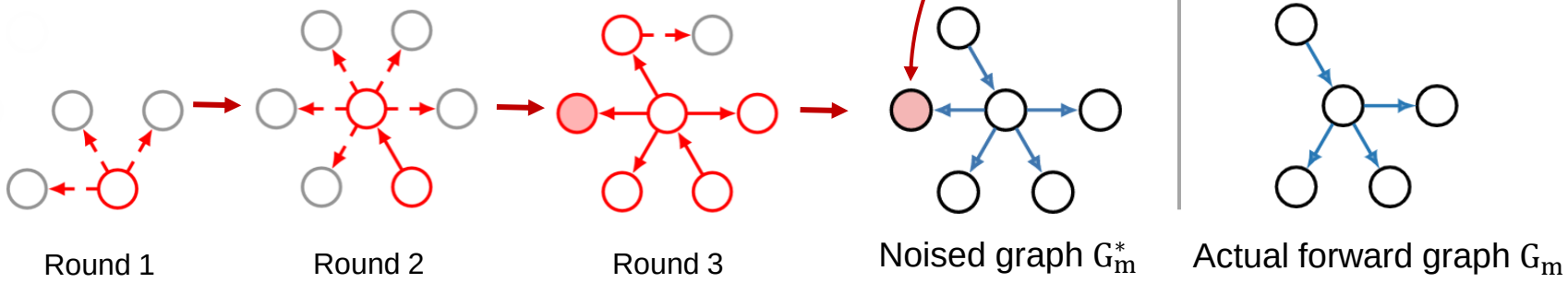
The Design: Noising the Traceback

Introduce random response to add noise.

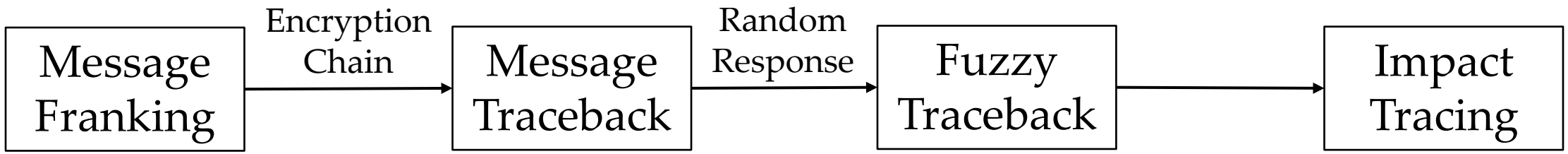
- Two servers: Platform & Tag server
- Tag server responses a query as:

> if tag_j exists, return 1
>
> else, return 0 w.p. $1 - \psi$
>
> return 1 w.p. ψ





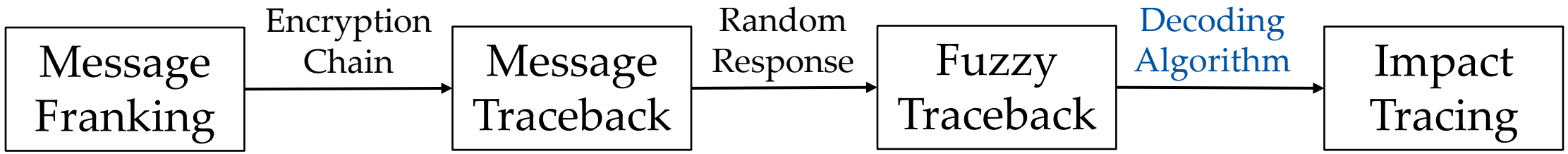
Remaining Concerns



Remaining Concerns

How to reveal influential spreaders from noised graph?

- Design a decoding algorithm to identify influential spreaders.



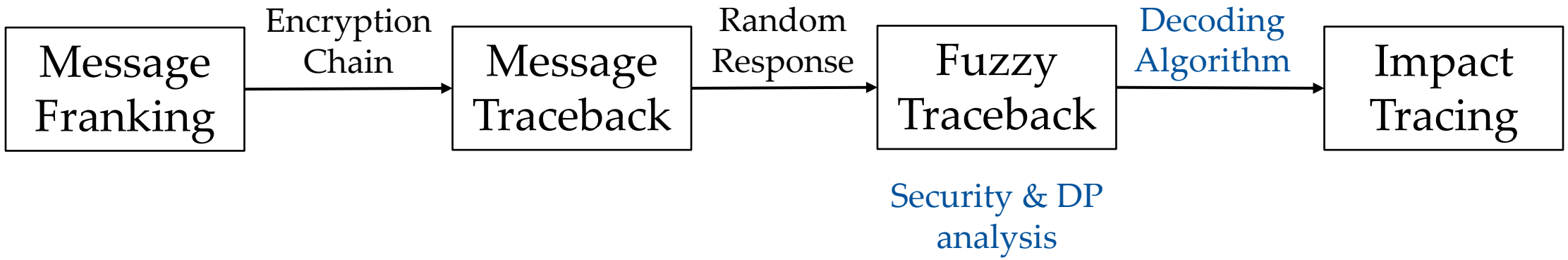
Remaining Concerns

How to reveal influential spreaders from noised graph?

- Design a decoding algorithm to identify influential spreaders.

Does the decoding algorithm violate privacy?

- Prove the noised graph satisfies differential privacy.



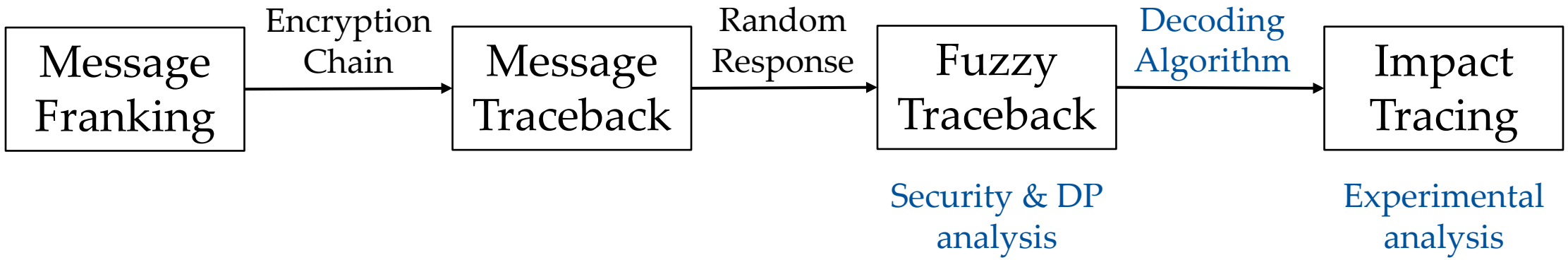
Remaining Concerns

How to reveal influential spreaders from noised graph?

- Design a decoding algorithm to identify influential spreaders.

Does the decoding algorithm violate privacy?

- Prove the noised graph satisfies differential privacy.



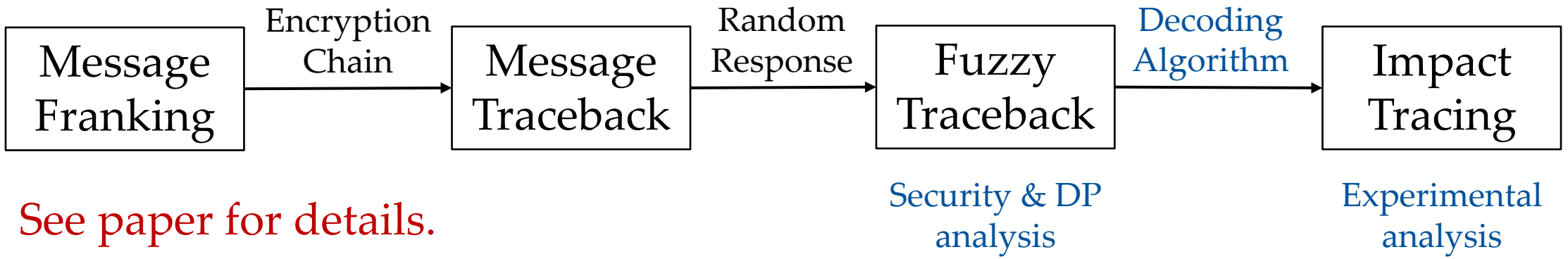
Remaining Concerns

How to reveal influential spreaders from noised graph?

- Design a decoding algorithm to identify influential spreaders.

Does the decoding algorithm violate privacy?

- Prove the noised graph satisfies differential privacy.



See paper for details.

Performance

Bandwidth & **Storage** (Byte)

Tracing Policy	Schemes†	Bandwidth		Storage	
		$\mathcal{S}$ - $\mathcal{P}$	$\mathcal{P}$ - $\mathcal{R}$	$\mathcal{C}$	$\mathcal{P}$
Source Tracing	PEB21 [39]	256	320	160	-
	LRTY22 [30]	243	243	243	-
	IAV22 [21]	380	484	380	-
Message Traceback	TMR19 [50]	96	80	34	104
	KTW22 [27]	203	203	16	136
Impact Tracing	Ours	96‡	72	16	6

$\mathcal{S}$, $\mathcal{R}$, $\mathcal{P}$, $\mathcal{C}$: Sender, recipient, platform, and client.

Store 1 billion messages: 5.6 GB

Performance

Bandwidth & **Storage** (Byte)

Tracing Policy	Schemes†	Bandwidth		Storage	
		$\mathcal{S}$ - $\mathcal{P}$	$\mathcal{P}$ - $\mathcal{R}$	$\mathcal{C}$	$\mathcal{P}$
Source Tracing	PEB21 [39]	256	320	160	-
	LRTY22 [30]	243		243	-
	IAV22 [21]	380		380	-
Message Traceback	TMR19 [50]	96	80	34	104
	KTW22 [27]	203	203	16	136
Impact Tracing	Ours	96‡	72	16	6

71%

$\mathcal{S}$, $\mathcal{R}$, $\mathcal{P}$, $\mathcal{C}$: Sender, recipient, platform, and client.

Store 1 billion messages: 5.6 GB

Performance

Bandwidth & **Storage** (Byte)

Tracing Policy	Schemes†	Bandwidth		Storage	
		$\mathcal{S}$ - $\mathcal{P}$	$\mathcal{P}$ - $\mathcal{R}$	$\mathcal{C}$	$\mathcal{P}$
Source Tracing	PEB21 [39]	256	320	160	-
	LRTY22 [30]	243		243	-
	IAV22 [21]	380		380	-
Message Traceback	TMR19 [50]	96	80	34	104
	KTW22 [27]	203			
Impact Tracing	Ours	96‡	72	16	6

71%

5%

94%

$\mathcal{S}$, $\mathcal{R}$, $\mathcal{P}$, $\mathcal{C}$: Sender, recipient, platform, and client.

Store 1 billion messages: 5.6 GB

Performance

Bandwidth & Storage (Byte)

Tracing Policy	Schemes†	Bandwidth		Storage	
		$\mathcal{S}$ - $\mathcal{P}$	$\mathcal{P}$ - $\mathcal{R}$	$\mathcal{C}$	$\mathcal{P}$
Source Tracing	PEB21 [39]	256	320	160	-
	LRTY22 [30]	243		243	-
	IAV22 [21]	380		380	-
Message Traceback	TMR19 [50]	96	80	34	104
	KTW22 [27]	203			
Impact Tracing	Ours	96‡	72	16	6

71%

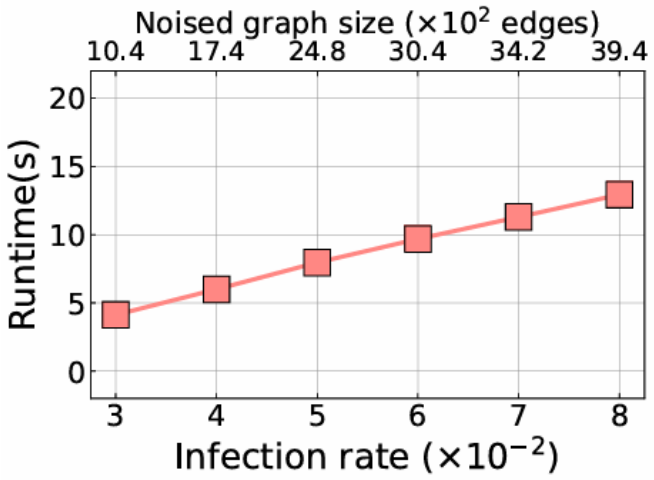
5%

94%

$\mathcal{S}$, $\mathcal{R}$, $\mathcal{P}$, $\mathcal{C}$: Sender, recipient, platform, and client.

Store 1 billion messages: 5.6 GB

Runtime of tracing (s)



Trace graph with 4,000 edges: 15 s

Transmit 1-KB message: 0.3 ms

Performance

Bandwidth & Storage (Byte)

Tracing Policy	Schemes†	Bandwidth		Storage	
		$\mathcal{S}$ - $\mathcal{P}$	$\mathcal{P}$ - $\mathcal{R}$	$\mathcal{C}$	$\mathcal{P}$
Source Tracing	PEB21 [39]	256	320	160	-
	LRTY22 [30]	243		243	-
	IAV22 [21]	380		380	-
Message Traceback	TMR19 [50]	96	80	34	104
	KTW22 [27]	203			
Impact Tracing	Ours	96‡	72	16	6

71%

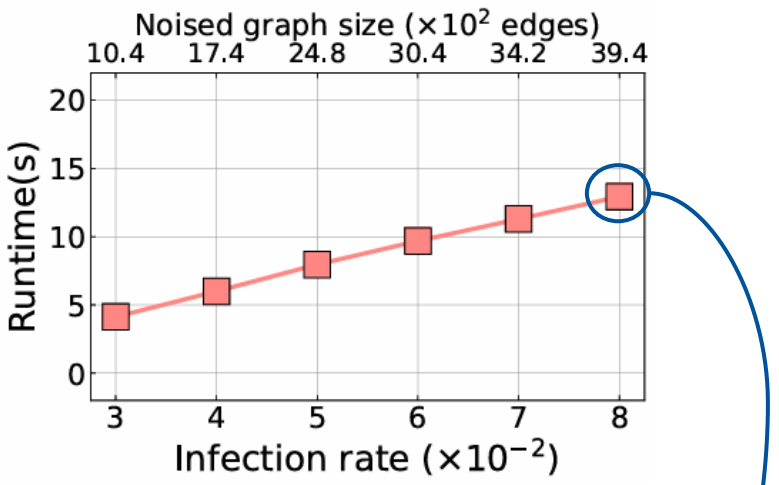
5%

94%

$\mathcal{S}$, $\mathcal{R}$, $\mathcal{P}$, $\mathcal{C}$: Sender, recipient, platform, and client.

Store 1 billion messages: 5.6 GB

Runtime of tracing (s)



Trace graph with 4,000 edges: 15 s

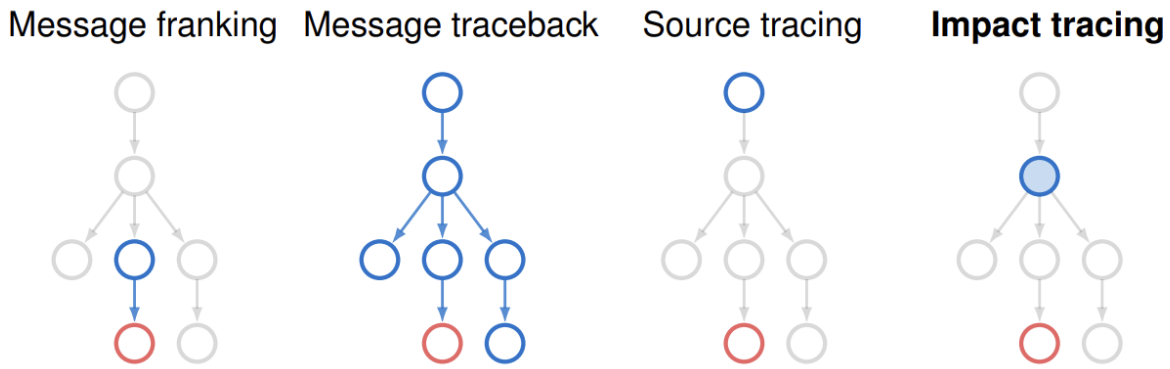
Transmit 1-KB message: 0.3 ms

Summary

Introduce impact tracing: balances traceability and privacy.

Design fuzzy message traceback and decoding algorithm.

Analyze security, privacy, and utility formally.

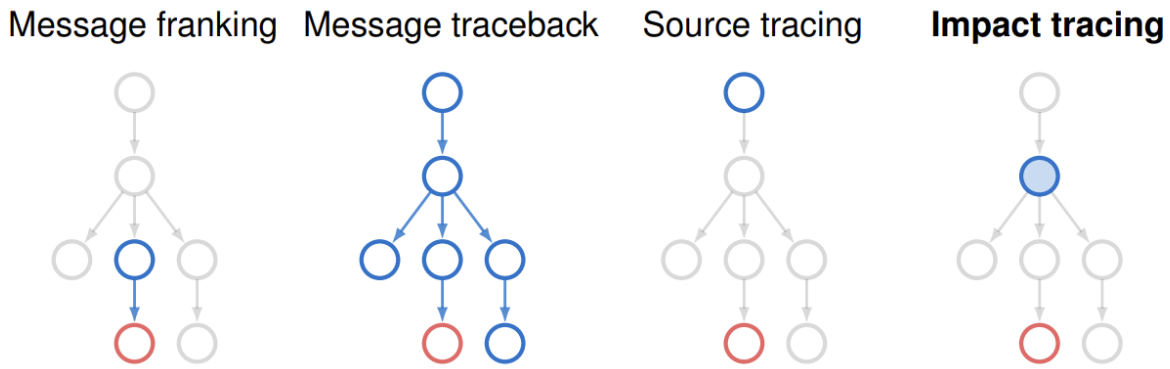


Summary

Introduce impact tracing: balances traceability and privacy.

Design fuzzy message traceback and decoding algorithm.

Analyze security, privacy, and utility formally.



Scan for code.

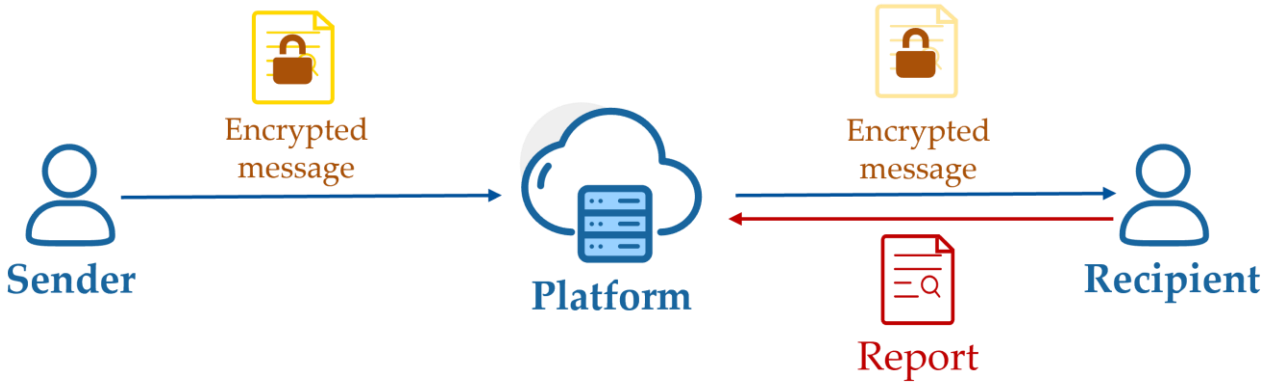
Thank for listening :) Any questions?

Bonus Slides

Message Franking

Message franking enables a recipient to report a message.

And, the platform can authenticate that the sender actually sent it.



Accountability: 1) Recipients cannot smear honest senders.

2) Senders cannot evade reporting.

Confidentiality: The platform learn nothing about unreported messages.

Simulation on Real-World Datasets

Influence evaluation

1. Simulate forwarding using SIR model.

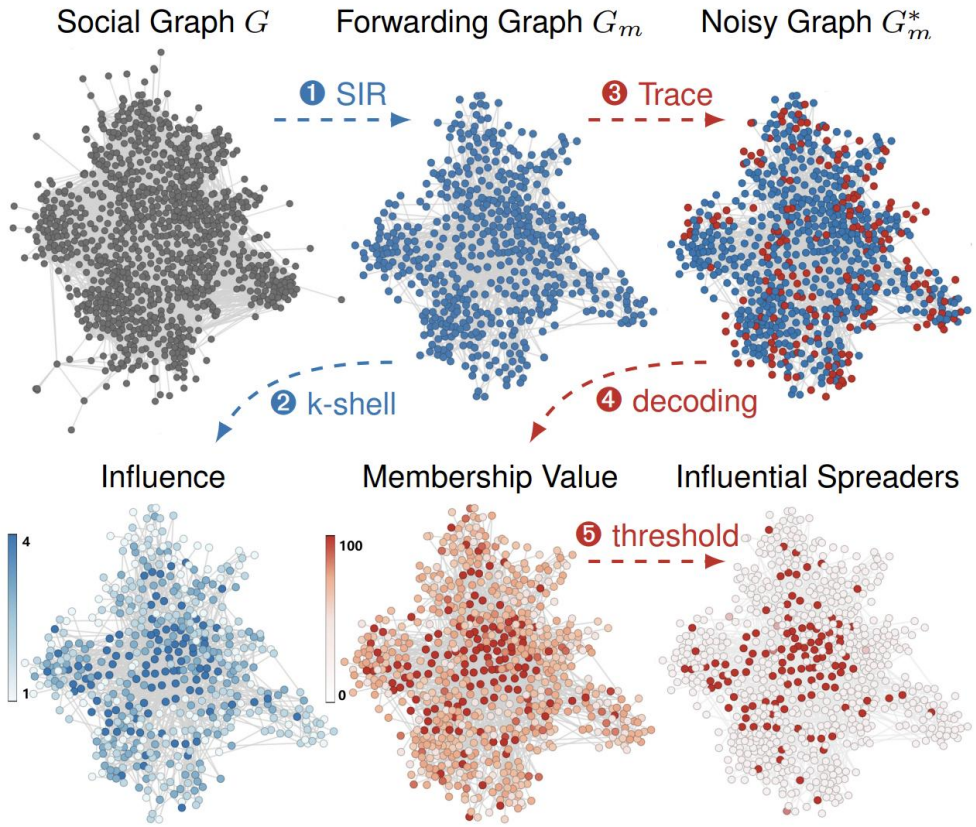
2. Evaluate vertices' impact using k-shell.

Impact tracing

3. Trace with fuzzy message traceback.

4. Decode the traced noisy graph.

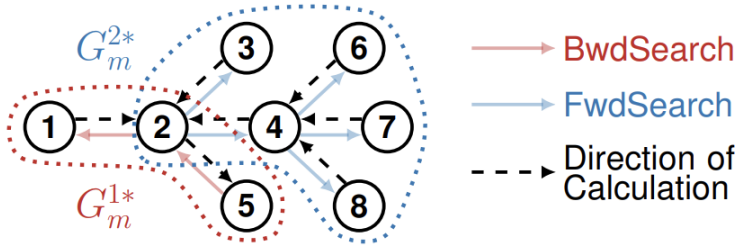
5. Output a set of influential spreaders.



The Design: Decoding the Result

Decoding algorithm: Computes a fuzzy value for each vertex.

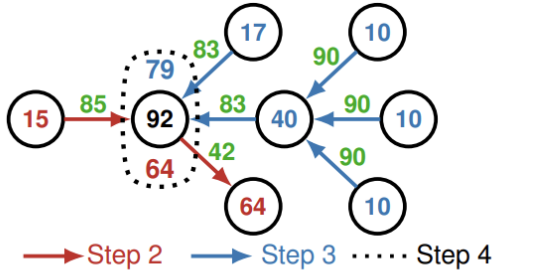
- The false positives of a vertex satisfy binomial distribution.
- A vertex has *only* one true precursor.
- A vertex is true positive if one of its descendants is true.



(a) Illustration of G_m^{1*} and G_m^{2*}

Graph	Vertex	$n(v)$	$n^*(v)$	$a_v(\%)$
G_m^{1*}	v(2)	50	1	85
G_m^{2*}	v(2)	50	2	83
	v(4)	100	3	90
	v(5)	10	1	42

(b) Step 1: caculate α_v



(c) Step 2-4: caculate β_v^1, β_v^2, and μ_v

IAS-DP: Defining Privacy

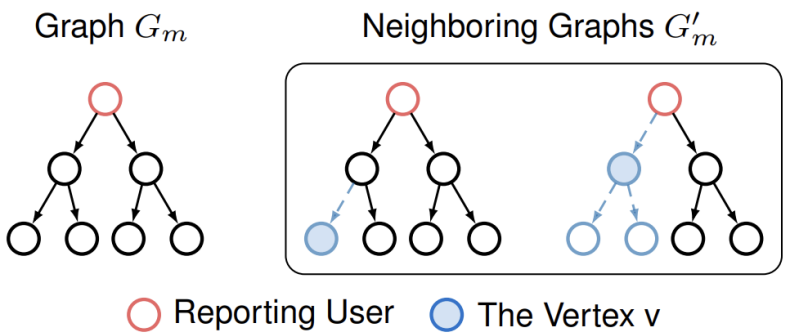
Individualized Asymmetric Subtree Differential Privacy

- **Individualized**: Privacy budgets vary per user.
- **Asymmetric**: Traceability adopts one-side noise.
- **Subtree**: All impacts caused by one user on one forwarding graph.

Definition 2 (ε_v-IAS-DP). A randomized algorithm $\mathcal{M}$ is ε_v-IAS-DP if given a graph G the following equation holds for any $S \in Range(\mathcal{M})$ and neighboring subgraph pair (G_m, G'_m), where G'_m is obtained by removing a subtree $tree(v)$ in G_m.

$$\Pr[\mathcal{M}(G_m) = S] \leq e^{\varepsilon_v} \cdot \Pr[\mathcal{M}(G'_m) = S],$$

where ε_v is the privacy budget for vertex v.



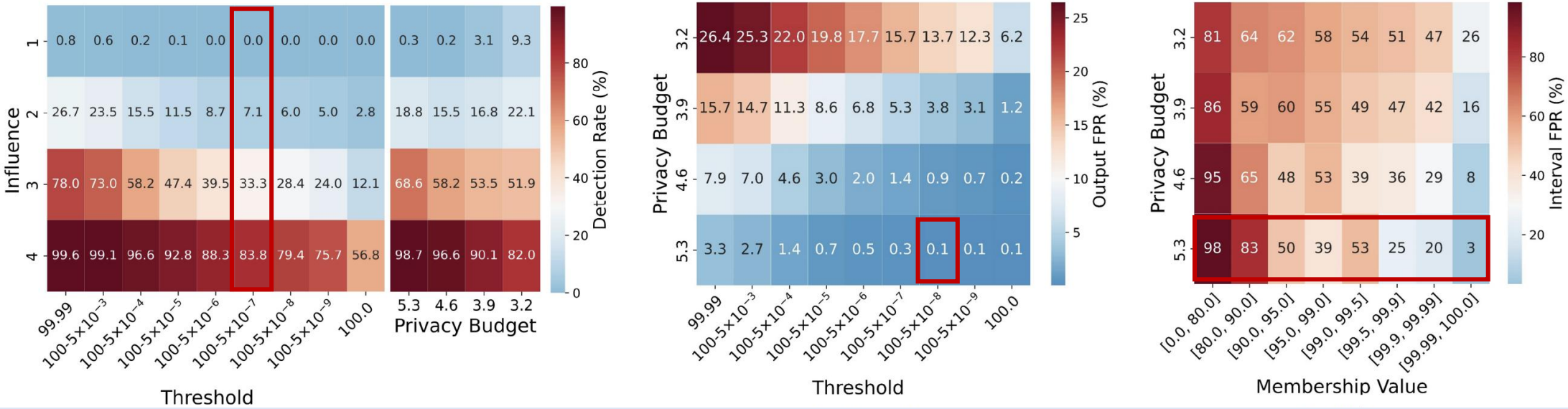
Theorem 1. The fuzzy message traceback scheme satisfies ε_v-IAS-DP, where $\varepsilon_v = \ln(1/\psi^n)$, n is the number of edges in $tree(v)$, and ψ is the FPR of random response.

Evaluating Utility & Privacy

Detection rate: The output contains as most influential users as possible.

Output FPR: The output should contain as few false positives as possible.

Interval FPR: Non-influential users should be hidden by sufficient noise.



Identifies 84% of the most influential spreaders and no the least influential users with 99.9% correctness (i.e., less than 0.25 false positives on average).