The Effect of False Positives:
Why Fuzzy Message Detection Leads to
Fuzzy Privacy Guarantees?

István András Seres[1], Balázs Pejó[2], and Péter Burcsi[1]

[1] Eötvös Loránd University, Budapest, Hungary
[2] CrySyS Lab, HIT/VIK/BME, Hungary

Abstract. Fuzzy Message Detection (FMD) is a recent cryptographic primitive invented by Beck et al. (CCS'21) where an untrusted server performs coarse message filtering for its clients in a recipient-anonymous way. In FMD — besides the true positive messages — the clients download from the server their cover messages determined by their false-positive detection rates. What is more, within FMD, the server cannot distinguish between genuine and cover traffic. In this paper, we formally analyze the privacy guarantees of FMD from three different angles.
First, we analyze three privacy provisions offered by FMD: recipient unlinkability, relationship anonymity, and temporal detection ambiguity. Second, we perform a differential privacy analysis and coin a relaxed definition to capture the privacy guarantees FMD yields. Finally, we simulate FMD on real-world communication data. Our theoretical and empirical results assist FMD users in adequately selecting their false-positive detection rates for various applications with given privacy requirements.

Keywords: Fuzzy Message Detection · unlinkability · anonymity · differential privacy · game theory

1 Introduction

Fuzzy Message Detection (FMD) [3] is a promising, very recent privacy-enhancing cryptographic primitive that aims to provide several desired privacy properties such as recipients' anonymity. In recipient-anonymous communication systems, not even the intended recipients can tell which messages have been sent to them without decrypting all messages. The main practical drawback for the users in a recipient-anonymous scheme such as messaging and payment systems is to *efficiently and privately* detect the incoming messages or transactions. Decrypting all traffic in the system leads to a private but inevitably inefficient and bandwidth-wasting scan. This challenge is tackled by FMD, which allows the users to outsource the detection of their incoming traffic to an untrusted server in an efficient and privacy-enhanced way. It is assumed that messages/transactions are posted continuously to a potentially public board, e.g., to a permissionless public blockchain. It is expected that users are intermittently connected and

resource-constrained. In the FMD scheme, whenever users are online, they download their genuine transactions as well as false-positive transactions according to their custom-set false-positive detection rate. The cryptographic technique behind FMD guarantees that true and false-positive messages are indistinguishable from the server's point of view. Thus, the false-positive messages act as cover traffic for genuine messages.

The FMD protocol caught the attention of many practitioners and privacy advocates due to the protocol's applicability in numerous scenarios. In general, it supports privacy-preserving retrieval of incoming traffic from store-and-forward delivery systems. We highlight two applications currently being implemented by multiple teams and waiting to be deployed in several projects [4,8,23,32].

- **Anonymous messaging.** In a recipient-anonymous messaging application, the senders post their recipient-anonymous messages to a store-and-forward server. If the server employs FMD, recipients can detect their incoming (and false-positive) messages in an efficient and privacy-enhanced way. Recently, the Niwl messaging application was deployed utilizing FMD [23].
- **Privacy-preserving cryptocurrencies & stealth payments.** In privacy-preserving cryptocurrencies, e.g., Monero [29], Zcash [33], or in a privacy-enhancing overlay, payment recipients wish to detect their incoming payments without scanning the whole ledger. At the time of writing, several privacy-enhancing overlays for Ethereum (e.g., Zeth [32], Umbra [4]) as well as for standalone cryptocurrencies (e.g., Penumbra [8]) are actively exploring the possibility of applying FMD in their protocols.

Contributions. Despite the rapid adoption and interest in the FMD protocol, as far as we know, there is no study analyzing the provided privacy guarantees. Consequently, it is essential to understand the privacy implications of FMD. Furthermore, it is an open question how users need to choose their false-positive detection rates to achieve an efficiency-privacy trade-off suitable for their scenario. In this work, we make the following contributions.

- **Information-Theoretical Analysis.** We assess and quantify the privacy guarantees of FMD and the enhanced k-anonymity it provides in the context of anonymous communication systems. We focus on three notions of privacy and anonymity: relationship anonymity, recipient unlinkability, and temporal detection ambiguity. We demonstrate that FMD does not provide relationship anonymity when the server knows the senders' identity. What is more, we also study relationship anonymity from a game-theoretic point of view, and show that in our simplified model at the Nash Equilibrium the users do not employ any cover traffic due to their selfishness. Concerning recipient unlinkability and temporal detection ambiguity, we show that they are only provided in a meaningful way when the system has numerous users and users apply considerable false-positive detection rates.
- **Differential Privacy Analysis.** We adopt differential privacy (DP) [11] for the FMD scenario and coin a new definition, called Personalized Existing

Edge Differential Privacy (PEEDP). Moreover, we analyze the number of incoming messages of a user with (ε, δ)-differential privacy. The uncovered trade-off between the FMD's false-positive rates and DP's parameters could help the users to determine the appropriate regimes of false-positive rates, which corresponds to the level of tolerated privacy leakage.

– **Simulation of FMD on Real-World Data.** We quantitatively evaluate the privacy guarantees of FMD through open-source simulations on real-world communication systems. We show that the untrusted server can effortlessly recover a large portion of the social graph of the communicating users, i.e., the server can break relationship anonymity for numerous users.

Outline. In Section 2, we provide some background on FMD, while in Section 3, we introduce our system and threat model. In Section 4, we analyze the privacy guarantees of FMD while in Section 5 we study FMD using differential privacy. In Section 6, we conduct simulations on real-world communication networks, and finally, in Section 7, we conclude the paper.

2   Fuzzy Message Detection

The FMD protocol seeks to provide a reasonable privacy-efficiency trade-off in use cases where *recipient anonymity* needs to be protected. Users generate detection keys and send them along with their public keys to the untrusted server. Senders encrypt their messages with their recipient's public key and create flag ciphertexts using the intended recipient's public key. Detection keys allow the server to test whether a flag ciphertext gives a match for a user's public key. If yes, the server stores the message for that user identified by its public key. In particular, matched flag ciphertexts can be false-positive matches, i.e., the user cannot decrypt some matched ciphertexts. Users can decrease their false-positive rate by sending more detection keys to the server. Above all, the FMD protocol ensures *correctness*; whenever a user comes online, they can retrieve their genuine messages. The *fuzziness* property enforces that each other flag ciphertext is tested to be a match approximately with probability p set by the recipient.

Besides recipient anonymity, FMD also aims to satisfy *detection ambiguity*, which requires that the server cannot distinguish between true and false-positive matching flag ciphertexts provided that ciphertexts and detection keys are honestly generated. Hence, whenever a user downloads its matched messages, false-positive messages serve as cover traffic for the genuine messages. For formal security and privacy definitions of FMD and concrete instantiations, we refer the reader to Appendix A and ultimately to [3]. To improve readability, in Table 1, we present the variables utilized in the paper: we refer to the downloaded (genuine or cover) flag ciphertext as a fuzzy tag.

Privacy-efficiency trade-off. If user u's false-positive rate is $p(u)$, it received $in(u)$ messages and the total number of messages in the system is M, then the server will store $tag(u) \approx in(u) + p(u)(M - in(u))$ messages for u. Clearly, as

Variable	Description
U	Number of honest users (i.e., recipients and senders)
M	Number of all messages sent by honest users
$p(u)$	False-positive detection rate of recipient u
$tag(u)$	Number of fuzzy tags received by u (i.e., genuine and false positive)
$tag_v(u)$	Number of fuzzy tags received by u from v
$in(u)$	Number of genuine incoming messages of u
$out(u)$	Number of sent messages of u

Table 1: Notations used throughout the paper.

the number of messages stored by the server increases, so does the strength of the anonymity protection of a message. Note the trade-off between privacy and bandwidth efficiency: larger false-positive rate $p(u)$ corresponds to stronger privacy guarantees but also to higher bandwidth as more messages need to be downloaded from the server.[3] Substantial bandwidth can be prohibitive in certain use cases, e.g., for resource-constrained clients. Even though in the original work of Beck et al. [3] their FMD instantiations support a restricted subset of $[2^{-l}]_{l \in \mathbb{Z}}$ as false-positive rates, in our privacy analysis, we lift these restrictions and assume that FMD supports any false-positive rate $p \in [0, 1]$.

Provided privacy protection. The anonymity protection of FMD falls under the "hide-in-the-crowd" umbrella term as legitimate messages are concealed amongst cover ones. More precisely, each legitimate message enjoys an enhanced version of the well-known notion of k-anonymity [36].[4] In more detail, the anonymity guarantee of the FMD scheme is essentially a "dynamic", "personalized", and "probabilistic" extension of k-anonymity. It is dynamic because k could change over time as the overall number of messages could grow. It is personalized because k might differ from user to user as each user could set their own cover detection rates differently. Finally, it is probabilistic because achieved k may vary message-wise for a user due to the randomness of the amount of selected fuzzy messages.

To the best of our knowledge, as of today, there has not been a formal anonymity analysis of the "enhanced k-anonymity" achieved by the FMD protocol. Yet, there is already a great line of research demonstrating the weaknesses and the brittle privacy guarantees achieved by k-anonymity [10,25]. Intuitively, one might hope that enhanced k-anonymity could yield strong(er) privacy and anonymity guarantees. However, we show both theoretically and empirically and by using several tools that this enhanced k-anonymity fails to satisfy standard anonymity notions used in the anonymous communication literature.[5]

---

[3] Similar scenario was studied in [5] concerning Bloom filters.

[4] Note that Beck et al. coined this as *dynamic k-anonymity*, yet, we believe it does not capture all the aspects of their improvement. Hence, we renamed it with a more generic term.

[5] For an initial empirical anonymity analysis, we refer the reader to the simulator developed by Sarah Jamie Lewis [24].

3   System and Threat Model

System Model.   In a typical application where the FMD scheme is applied, we distinguish between the following four types of system components where the users can simultaneously be senders and recipients.

1. **Senders:** They send encrypted messages to a message board. Messages are key-private, i.e., no party other than the intended recipient can tell which public key was used to encrypt the message. Additionally, senders post flag ciphertexts associated with the messages to an untrusted server. The goal of the flag ciphertexts is to allow the server and the recipients to detect their messages in a privacy-enhanced manner.
2. **Message Board:** It is a database that contains the senders' messages. In many applications (e.g., stealth payments), we think of the message board as a public bulletin board; i.e., everyone can read and write the board. It might be implemented as a blockchain or as a centrally managed database, e.g., as would be the case in a messaging application. In either case, we assume that the message board is always available and that its integrity is guaranteed.
3. **Server:** It stores the detection keys of recipients. Additionally, it receives and stores flag ciphertexts from senders and tests the flag ciphertexts with the recipient's detection keys. It forwards matching flag ciphertexts and their associated data (messages, transactions, etc.) to recipients whenever they query it. Typically, flag ciphertexts match numerous recipients' public keys.[6]
4. **Recipients:** The recipient obtains matching flag ciphertexts from the server. An application-dependent message is attached as associated data to each flag ciphertext, e.g., e-mail, payment data, or instant message. The number of matching ciphertexts is proportional to the recipient's false-positive detection rate and all the messages stored by the untrusted server.

Threat model.   Our focus is on the privacy and anonymity guarantees provided by FMD. Hence, we assume that the FMD scheme is a secure cryptographic primitive, i.e., the cryptographic properties of FMD (correctness, fuzziness, and detection ambiguity) hold. Senders and recipients are assumed to be honest. Otherwise, they can be excluded from the messages' anonymity sets. We consider two types of computationally-bounded attackers that can compromise user's privacy in an FMD scheme. The adversaries' goal is to learn as much information as possible about the relationship between senders, recipients, and messages.

– **Server:** Naturally, the server providing the FMD service can endanger the user's privacy since it has continuous access to every relevant information related to message detection. Specifically, the server knows the users' false-positive rates. It can analyze each incoming message, flag ciphertext, and their corresponding anonymity sets.

---

[6] In this work, we stipulate that a single server filters the messages for all users, i.e., a single server knows all the recipients' detection keys.

- *Sender-oracle.* The server may know the sender of each message, i.e., a sender-oracle might be available in FMD. For instance, it is mandatory for the untrusted server if it only serves registered users. We assumed solely in Section 4.2 that such sender-oracle is available. If FMD is integrated into a system where senders apply anonymous communication (e.g., use Tor to send messages and flag ciphertexts to the server), then sender-oracle is not accessible to the FMD server.
  – **Eavesdropper:** A local passive adversary might observe the amount of data each user downloads from the server. Specifically, an eavesdropper could inspect the number of flag ciphertexts each user has received. Even though this attacker type does not have continual intrusion to the server's internal state, as we will show, it can still substantially decrease FMD user's privacy, e.g., if $p(u)$ is known, then the number of genuine incoming messages of users does not enjoy sufficiently high privacy protection (see Section 5).

4   Privacy Guarantees in FMD

In this section, we analyze and quantify various privacy and anonymity guarantees provided by the FMD scheme. Specifically, in Sections 4.1, 4.2, and 4.3, we measure recipient unlinkability, assess relationship anonymity, and estimate detection ambiguity, respectively. Note that for the latter two property we provide experimental evaluations in Section 6, and we formulate a game in Appendix D concerning relationship anonymity. We denote the security parameter with λ, and if an (probabilistic) algorithm A outputs x, then we write $A \rightarrow x$ ($A \xrightarrow{\$} x$). The Binomial distribution with success probability p and number of trials n is denoted as $\mathsf{Binom}(n, p)$, while a normal distribution with mean μ and variance σ^2 is denoted as $\mathcal{N}(\mu, \sigma^2)$.

4.1   Recipient Unlinkability

In anonymous communication systems, recipient unlinkability is the cornerstone of anonymity guarantees. It ensures that it is hard to distinguish whether two different messages were sent to the same recipient or different ones. Whenever recipient unlinkability is not attained, it facilitates possibly devastating passive attacks, i.e., if an adversary can infer which messages are sent to the same recipient, then the adversary can effortlessly launch intersection attacks, see Appendix E. In the absence of recipient unlinkability, it is also possible to efficiently map every message to its genuine recipient by 1) clustering the messages that are sent to the same recipient and 2) see the intersection of the users who downloaded the flag ciphertexts sent to the same recipient.

We consider a definition of recipient unlinkability similar to the one introduced in [2]. Informally, in the recipient unlinkablity game, we examine two recipients u_0, u_1 and a sender u_2. The challenger C generates uniformly at random $c \xleftarrow{\$} \{0, 1\}$ and instructs u_2 to send message m_α to u_c. Afterwards, C draws

a uniformly random bit $b \xleftarrow{\$} \{0,1\}$. If $b = 0$, then instructs u_2 to send a message m_β to u_c. Otherwise u_2 sends m_β to u_{1-c}. Adversary $\mathcal{A}$ examines the network, the flag ciphertexts and all communications and outputs b' indicating whether the two messages were sent to the same recipient.

Definition 1 (Recipient unlinkability (RU)). *An anonymous communication protocol Π satisfies recipient unlinkability if for all probabilistic polynomial-time adversaries $\mathcal{A}$ there is a negligible function $\mathsf{negl}(\cdot)$ such that*

$$\Pr[\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda) = 1] \leq \frac{1}{2} + \mathsf{negl}(\lambda), \tag{1}$$

where the privacy game $\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda)$ is defined in Figure 6 in Appendix B.

We denote the set of users who downloaded message m by $fuzzy(m)$, i.e., they form the anonymity set of the message m. We estimate the advantage of the following adversary $\mathcal{A}$ in the $\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda)$ game: $\mathcal{A}$ outputs 0 if $fuzzy(m_\alpha) \cap fuzzy(m_\beta) \neq \emptyset$ and outputs 1 otherwise. Note that $\mathcal{A}$ always wins if the same recipient was chosen by the challenger (i.e., $b = 0$) because it is guaranteed by the correctness of the FMD scheme that $u_c \in fuzzy(m_\alpha) \cap fuzzy(m_\beta)$. Therefore, we have that $\Pr[\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda) = 1 | b = 0] = 1$. If two different recipients were chosen by the challenger in the $\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda)$ game (i.e., $b = 1$), then $\mathcal{A}$ wins iff. $fuzzy(m_\alpha) \cap fuzzy(m_\beta) = \emptyset$. The advantage of the adversary can be computed as follows.

$$\Pr[\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda) = 1 | b = 1] = \Pr[\cap_{m \in \{m_\alpha, m_\beta\}} fuzzy(m) = \emptyset | b = 1] =$$

$$= \sum_{i=1}^{U} \sum_{\substack{V \subseteq U \\ |V|=i \\ \alpha \in V}} \Pr[fuzzy(m_\alpha) = V] \cdot \Pr[(V \cap fuzzy(m_\beta) = \emptyset] = \tag{2}$$

$$= \sum_{i=1}^{U} \sum_{\substack{V \subseteq U \\ |V|=i \\ \alpha \in V}} \left(\prod_{u_l \in V \setminus \{u_0\}} p(u_l) \cdot \prod_{u_l \in U \setminus V} (1 - p(u_l)) \right) \cdot \left(\prod_{u_l \in V} (1 - p(u_l)) \right).$$

We simplify the adversarial advantage in Equation 2 by assuming that $\forall i : p(u_i) = p$ and that the sizes of the anonymity sets are fixed at $\lfloor pU \rfloor$. Moreover, computer-aided calculations show that the following birthday paradox-like quantity can be used as a sufficiently tight lower bound[7] for the recipient unlinkability adversarial advantage, whenever $p(u_i)$ are close to each other.

$$\prod_{j=1}^{\lfloor pU \rfloor} \frac{U - \lfloor pU \rfloor - j}{U} = \prod_{j=1}^{\lfloor pU \rfloor} \left(1 - \frac{\lfloor pU \rfloor + j}{U} \right) \approx \prod_{j=1}^{\lfloor pU \rfloor} e^{-\frac{\lfloor pU \rfloor + j}{U}} = \tag{3}$$

$$= e^{-\frac{\sum_j (\lfloor pU \rfloor + j)}{U}} = e^{-\frac{3\lfloor pU \rfloor^2 + \lfloor pU \rfloor}{2U}} \leq \Pr[\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda) = 1 | b = 1].$$

---

[7] This lower bound is practically tight since the probability distribution of the adversary's advantage is concentrated around the mean $\lfloor pU \rfloor$ anyway.

The approximation is obtained by applying the first-order Taylor-series approximation for $e^x \approx 1 + x$, whenever $|x| \ll 1$. We observe that the lower bound for the adversary's advantage in the recipient unlinkability game is a negligible function in U for a fixed false-positive detection rate p. Thus, in theory, the number of recipients U should be large in order to achieve recipient unlinkability asymptotically. In practice, the classical birthday-paradox argument shows us that the two anonymity sets intersect with constant probability if $p = \theta\left(\frac{1}{\sqrt{U}}\right)$. Our results suggest that a deployment of the FMD scheme should *concurrently* have a large number of users with high false-positive rates in order to provide recipient unlinkability, see Figure 1a for the concrete values of Equation 3.
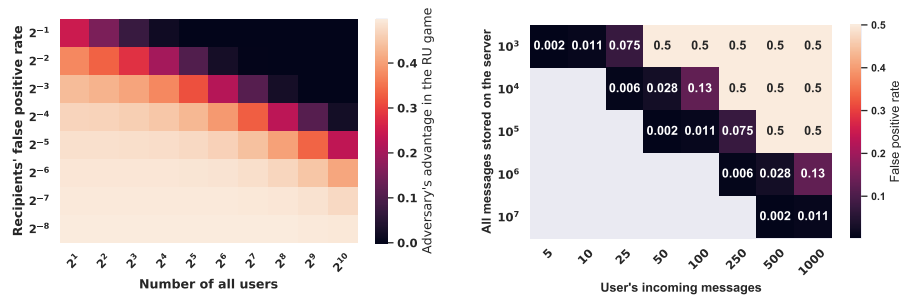
(a) Approximate values of the recipient unlinkability adversarial advantage in the $\mathcal{G}_{RU}^{\mathcal{A},\Pi}(\lambda)$ game according to Equation 3.

(b) Smallest false-positive rates to obtain temporal detection ambiguity against the utilized statistical tests.

Fig. 1: Recipient unlinkability and temporal detection ambiguity guarantees provided by the FMD scheme for various parameter settings.

4.2   Relationship anonymity

Relationship anonymity ensures that the adversary cannot determine the sender and the recipient of a communication at the same time. Intuitively, recipients applying low false-positive rates receive only a handful of fuzzy tags from peers they are not communicating with. Therefore, multiple fuzzy tags between a sender and a recipient can eradicate their relationship anonymity, given that the server knows the number of messages a sender issued. We assume the server knows the sender of each message, which holds whenever the untrusted server has access to a sender-oracle, see Section 3.

The number of fuzzy tags between a *non-communicating pair* of reciever u_1 and sender u_2 follows $\mathsf{Binom}(\mathsf{out}(u_2), p(u_1))$. If $tag_{u_2}(u_1)$ is saliently far from the expected mean $\mathsf{out}(u_2)p(u_1)$, then the untrusted server can deduce with high confidence that a relationship exists between the two users. We approximate the bi-

(a) Relationship anonymity for $100 \leq \mathsf{out}(s)$.

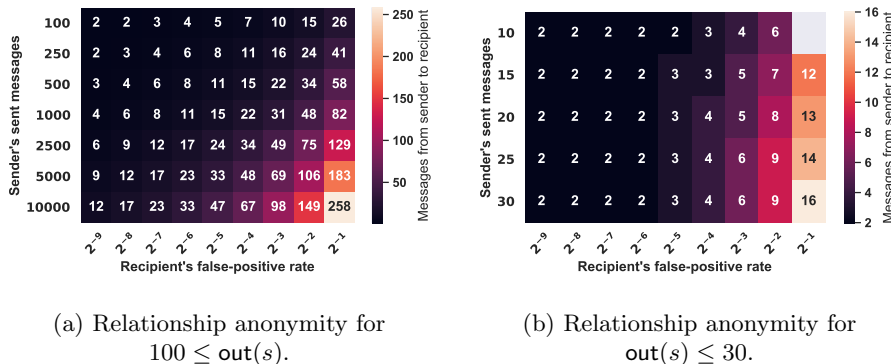(b) Relationship anonymity for $\mathsf{out}(s) \leq 30$.

Fig. 2: The minimum number of messages between a pair of users that statistically reveal the relationship of the communicating users (significance level 1%).

nomial distribution above with $\mathcal{N}^* := \mathcal{N}(\mathsf{out}(u_2)p(u_1), \mathsf{out}(u_2)p(u_1)(1-p(u_1)))$[8] so we can apply Z-tests to determine whether u_2 and u_1 had exchanged messages. Concretely, we apply two-tailed Z-test[9] for the hypothesis $H : tag_{u_2}(u_1) \sim \mathcal{N}(\mathsf{out}(u_2)p(u_1), \mathsf{out}(u_2)p(u_1)(1-p(u_1)))$. If the hypothesis is rejected, then users u_2 and u_1 are deemed to have exchanged messages.

Each recipient u_1 downloads on average $tag_{u_2}(u_1) \approx p(u_1)(out(u_2)-in_{u_2}(u_1))+ in_{u_2}(u_1)$ fuzzy messages from the messages sent by u_2, where $in_{u_2}(u_1)$ denotes the number of genuine messages sent by u_2 to u_1. We statistically test with Z-tests (when $100 \leq \mathsf{out}(u_2)$) and t-tests (when $\mathsf{out}(u_2) \leq 30$) whether $tag_{u_2}(u_1)$ could have been drawn from the $\mathcal{N}^*$ distribution, i.e., there are no exchanged messages between u_1 and u_2. The minimum number of genuine messages $in_{u_2}(u_1)$ that statistically expose the communication relationship between u_1 and u_2 is shown for various scenarios in Figure 2. We observe that the relationship anonymity of any pair of users could be broken by a handful of exchanged messages. This severely limits the applicability of the FMD scheme in use cases such as instant messaging. To have a meaningful level of relationship anonymity with their communicating peer, users should either apply substantial false-positive rates, or the server must not be able to learn the sender's identity of each message. The latter could be achieved, for instance, if senders apply an anonymous communication system to send messages or by using short-lived provisional pseudo IDs where no user would send more than one message.

Game Theoretic Analysis. Incentive compatibility has the utmost importance in decentralized and privacy-enhancing technologies. Therefore, we present a game-theoretic study of the FMD protocol concerning relationship anonymity.

---

[8] Note that this approximation is generally considered to be tight enough when $\mathsf{out}(u_2)p(u_1) \geq 5$ and $\mathsf{out}(u_2)(1-p(u_1)) \geq 5$.

[9] For senders with only a few sent messages ($\mathsf{out}(u_2) \leq 30$), one can apply t-tests instead of Z-tests.

We believe applying game theory to FMD by itself is a fruitful and over-arching direction. Our goal, besides conducting a preliminary analysis, is to raise interest and fuel future research towards this direction. The formalization of the game as well as the corresponding theorems and proofs can be found in Appendix D.

4.3   Temporal detection ambiguity

The FMD scheme is required to satisfy the security notion of *detection ambiguity* devised by Beck et al. [3]. Namely, for any message that yields a match for a detection key, the server should not be able to decide whether it is a true or a false-positive match. This definition is formalized for a single incoming message in isolation. Yet, the detector server can *continuously* observe the stream of incoming messages.[10] Consequently, the server might be able to assess whether the user has received a message in *a certain time interval*. To capture this time-dependent aspect, we relax detection ambiguity and coin the term *temporal detection ambiguity*. Informally, no adversary should be able to tell in a given time interval having access to all incoming flag ciphertexts whether a user received an incoming true-positive match. We provide the formal definition in Appendix B, and we empirically study temporal detection ambiguity on real communication data in Section 6. In Section 5, we measure the level of privacy protection the number of incoming messages enjoys from a differential privacy angle.

Any message that enters the communication system yields a match to a detection key according to its set false-positive rate. Specifically, the number of false-positive matches acquired by user u's detection key follows a $\mathsf{Binom}(M - \mathsf{in}(u), p(u))$ distribution. Similarly to Section 4.2, if M is large, then we can approximate the number of false-positive matches with a $\mathcal{N}(p(u)M, p(u)(1 - p(u))M)$ distribution and use statistical tests to assess that the number of downloaded messages by a recipient is statistically far from the expected number of downloaded messages. More precisely, the adversary can statistically test whether $tag(u)$ could have been drawn from $\mathcal{N}(p(u)M, p(u)(1 - p(u))M)$ (the approximation of $\mathsf{Binom}(M, p(u))$). We observe that in an epoch, a user should have either large false-positive rates or a small number of incoming messages to provide temporal detection ambiguity, shown in Figure 1b.

5   Differential Privacy Analysis

Differential privacy (DP) [11] is a procedure for sharing information about a dataset by publishing statistics of it while withholding information about single data points. DP is formalized in Definition 2; the core idea is to ensure that an arbitrary change on any data point in the database has a negligible effect on the query result. Hence, it is infeasible to infer much about any data point.

Definition 2 (Differential Privacy [11]). *An algorithm A satisfies ε-differential privacy if for all $S \subseteq Range(A)$ and every input pair D and D' differing in a*

---

[10] As an illustrative example collected from a real communication system, see Figure 3b.

single element Equation 4 holds.

$$\Pr(A(D) \in S) \le e^{\varepsilon} \cdot \Pr(A(D') \in S). \tag{4}$$

Personalized Existing Edge DP. A widely used relaxation of the above definition is (ε, δ)-DP, where Equation 4 is extended with a small additive term δ at the right end. There are over 200 modifications of DP [9], we combined several to make it suitable for FMD. Concretely, we create a novel definition called *Personalized Existing Edge DP (PEEDP)* (formally defined in Definition 3)[11] by combining four existing notions. We utilize *edge-DP* [17] which applies DP to communication graphs: D and D' are the original communication graphs with and without a particular edge respectively, and S is a set of graphs with fuzzy edges included. Furthermore, we apply *personalized DP* [18], which allocates different level of protection to incoming messages, as in FMD the users' false positive rates could differ.

Hiding the presence or absence of a message is only possible by explicitly removing real messages and adding fuzzy ones to the communication graph, which is indistinguishable from real ones. This setting (i.e., protecting existence and not value) corresponds to *unbounded DP* [20]. Hence, as also noted in [3], without a false negative rate (which would directly contradict correctness), FMD cannot satisfy DP: fuzzy messages can only hide the presence of a message not the absence. To tackle this imbalance, we utilize *asymmetric DP* [37] which only protects some of the records determined by policy P. It only differs from Definition 2 in the relationship of D and D' as Equation 4 should only hold for *every input pair D and D' where later is created by removing in D a single sensitive record defined by P*. By combining all these DP notions, we can formulate our PEEDP definition.

Definition 3 ($\bar{\varepsilon}$- Personalized Existing Edge Differential Privacy). *An algorithm A satisfies $\bar{\varepsilon}$-PEEDP (where $\bar{\varepsilon}$ is an element-wise positive vector which length is equal with the amount of nodes in D) if Equation 5 holds for all $S \subseteq Range(A)$ and every input graphs D and D' where later is created by removing in D a single incoming edge of user u.*

$$\Pr(A(D) \in S) \le e^{\bar{\varepsilon}_u} \cdot \Pr(A(D') \in S). \tag{5}$$

Once we formalized a suitable DP definition for FMD, it is easy to calculate the trade-off between efficiency (approximated by $p(u)$) and privacy protection (measured by ε_u). This is captured in Theorem 1 (proof can be found in Appendix C).

Theorem 1. *If we assume the distribution of the messages are IID then FMD satisfy $\left[\log \frac{1}{p(u)} \right]_{u=1}^{U}$-PEEDP.*

---

[11] We elaborate more on various DP notions in Appendix C.

Therefore, detection rates $p(u) = \{0.5^0, 0.5^1, 0.5^2, 0.5^4, 0.5^8\}$ in FMD correspond to $\varepsilon_u = \{0.000, 0.693, 1.386, 2.773, 5.545\}$ in $\bar{\varepsilon}$-PEEDP. Clearly, perfect protection (i.e., $\varepsilon_u = 0$) is reached only when all messages are downloaded (i.e., $p(u) = 1$). On the other hand, the other ε values are much harder to grasp: generally speaking, privacy-parameter below one is considered strong with the classic DP definition. As PEEDP only provides a relaxed guarantee, we can postulate that the privacy protection what FMD offers is weak.

Protecting the Number of Incoming Messages. In most applications, e.g., anonymous messaging or stealth payments, we want to protect the number of incoming messages of the users, $in(u)$. Intuitively, the server observes $tag(u) \sim in(u) + \mathsf{Binom}(M - in(u), p(u))$ where (with sufficiently large M) the second term can be thought of as Gaussian-noise added to mask $in(u)$, a common technique to achieve (ε, δ)-DP. Consequently, FMD does provide $(\varepsilon_u, \delta_u)$-DP[12] for the number of incoming messages of user u, see Theorem 2 (proof in Appendix C).

Theorem 2. *If we assume the distribution of the messages is IID than the FMD protocol provides $(\varepsilon_u, \delta_u)$-DP for the number of incoming messages $in(u)$ of user u where $\delta_u = \max_u(p(u), 1 - p(u))^{M - in(u)}$ and*

$$\varepsilon_u = \log \left[\max_u \left(\frac{p(u) \cdot (M - 2 \cdot in(u))}{(1 - p(u)) \cdot (in(u) + 1)}, \frac{(1 - p(u)) \cdot (M - in(u))}{p(u)} \right) \right].$$

M	100	100	100	**200**	1 000 000	1 000 000	1 000 000	**2 000 000**
$in(u)$	10	10	**20**	10	100	100	**1000**	100
$p(u)$	0.5^4	$\mathbf{0.5^2}$	0.5^4	0.5^4	0.5^8	$\mathbf{0.5^4}$	0.5^8	0.5^8
ε_u	7.2	5.6	7.1	8.0	19.4	16.5	19.4	20
δ_u	3e-3	6e-12	6e-3	5e-6	1e-1700	1e-28027	1e-1699	1e-3400

Table 2: Exemplary settings to illustrate the trade-off between the false-positive rate $p(u)$ and the privacy parameters of (ε, δ)-differential privacy for protecting the number of incoming messages.

To illustrate our results, we provide some exemplary settings in Table 2 and show how the false positive rate $p(u)$ translates into ε_u and δ_u. It is visible that increasing the detection rate does increase the privacy protection (i.e., lower ε_u and δ_u), and increasing the overall and incoming messages result in weaker privacy parameter ε_u and δ_u respectively. These results suggest, that even the number of incoming messages does not enjoy sufficient (differential) privacy protection in FMD, as the obtained values for ε_u are generally considered weak.

---

[12] Note that this is also a personalized guarantee as in [18].

6 Evaluation

We evaluate the relationship anonymity and temporal detection ambiguity guarantees of FMD through simulations on data from real communication systems.[13] We chose two real-world communication networks that could benefit from implementing and deploying FMD on top of them.

- **College Instant Messaging (IM) [30].** This dataset contains the instant messaging network of college students from the University of California, Irvine. The graph consists of 1899 nodes and 59 835 edges that cover 193 days of communication.
- **EU Email [31].** This dataset is a collection of emails between members of a European research institution. The graph has 986 nodes and 332 334 edges. It contains the communication patterns of 803 days.

Users are roughly distributed equally among major Information Privacy Awareness categories [35], thus for each node in the datasets, we independently and uniformly at random chose a false-positive rate from the set $\{2^{-l}\}_{l=1}^{7}$. Note that the most efficient FMD scheme only supports false-positive rates of the form 2^{-l}. Moreover, for each message and user in the system, we added new "fuzzy" edges to the graph according to the false-positive rates of the messages' recipients. The server is solely capable of observing the message-user graph with the added "fuzzy" edges that serve as cover traffic to enhance the privacy and anonymity of the users. We run our experiments 10-fold where on average, there are around 16 and 48 million fuzzy edges for the two datasets, i.e., a randomly picked edge (the baseline) represents a genuine message with $\ll 1\%$.
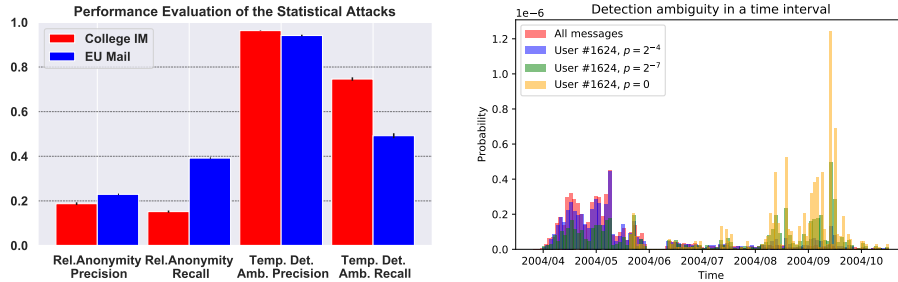
(a) The precision and recall of the statistical tests breaking relationship anonymity and temporal detection ambiguity, cf. Section 4.2 and 4.3.

(b) Temporal probability distribution of receiving a fuzzy tag for various false-positive detection rates. The exemplary user is taken from the College IM dataset.

Fig. 3: Privacy guarantees of FMD in simulations on real communication systems.

---

[13] The simulator can be found at https://github.com/seresistvanandras/FMD-analysis

6.1   Uncovering the relationship graph

The server's goal is to uncover the original social graph of its users, i.e., to expose the communicating partners. The relationship anonymity of a sender and a receiver can be easily uncovered by the statistical test introduced in Section 4.2 especially if a user is receiving multiple messages from the same sender while having a low false-positive rate. We found that statistical tests produce a 0.181 and 0.229 precision with 0.145 and 0.391 recall on average in predicting the communication links between all the pairs of nodes in the College IM and EU Email datasets, respectively, see Figure 3a. The results corresponding to the EU Email dataset are higher due to the increased density of the original graph. These results are substantial as they show the weak anonymization power of FMD in terms of relationship anonymity.
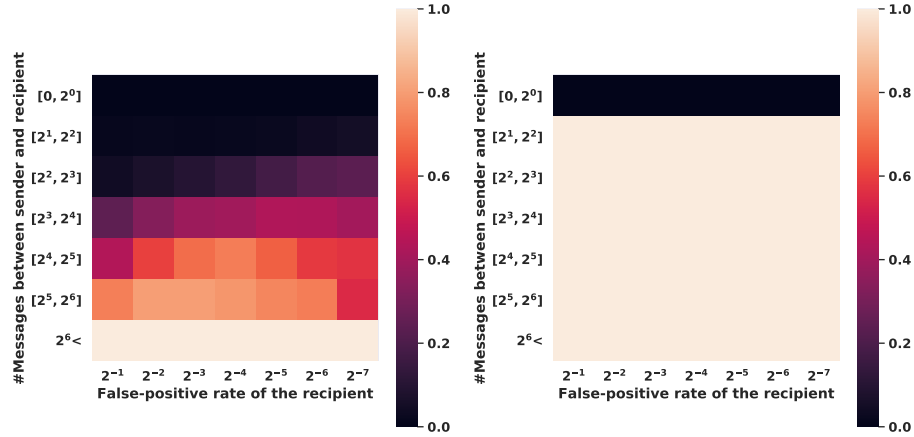

Fig. 4: Recall (left) and precision (right) of the statistical tests in breaking relationship anonymity (see Section 4.2) in simulations on the College IM dataset.

Specifically, communication relationships where merely a single message has been exchanged remain undetected by the applied statistical tests, cf. Figure 4. However, note that for every other pairs of users, neither of the analyzed datasets yields false positives by the used statistical tests. These simulation results demonstrate that relationship anonymity is effectively maintained against statistical attacks if each user sends only a single message from the server's point of view. This can be achieved by cryptographic tools or anonymous communication systems, e.g., Tor. On the other hand, recurrent communication relationships reveal the relationship of communicating peers. Thus, relationship anonymity is breached with perfect precision, cf. Figure 4. Simulation results confirm our intuition as well. Namely, statistical tests produce higher recall for nodes with lower false-

positive detection rates, while they are less effective for communicating pairs that exchanged very few messages.

6.2   Breaking Temporal Detection Ambiguity

We empirically quantify whether users can deny that they received an incoming (true positive) message. We consider 25 000 randomly selected messages with the corresponding fuzzy edges as one epoch. The server tried to assess using statistical tests (see Section 4.3) that a user has received an incoming message. The intuition is that users receive messages heterogeneously concerning time. Hence, surges in incoming traffic might not be adequately covered by fuzzy edges for users with low false-positive rates. Thus, these messages could be tight to the receiver with high probability, see Figure 3b for an illustrative example. Indeed, Figure 3a suggests that, in general, deniability can be broken effectively with high precision and recall. On the other hand, Figure 3b also shows that higher false-positive rates could provide enough cover traffic for messages within these conspicuous epochs, which is in line with the findings presented in Figure 1b.

7   Conclusion

In this paper, we present a privacy and anonymity analysis of the recently introduced Fuzzy Message Detection scheme. Our analysis is thorough as it covers over three directions. Foremost, an information-theoretical analysis was carried out concerning recipient unlinkability, relationship anonymity, and temporal detection ambiguity. It is followed by a differential privacy analysis which leads to a novel privacy definition. Finally, we gave an exhaustive simulation based on real-world data. Our findings facilitate proper parameter selection and the deployment of the FMD scheme into various applications. Yet, we also raise concerns about the guarantees what FMD provides and questions whether it is adequate/applicable for many real-world scenarios.

Limitations and Future Work. Although far-reaching, our analysis only scratches the surface of what can be analyzed concerning FMD, and substantial work and important questions remain as future work. Thus, a hidden goal of this paper is to fuel discussions about FMD so it can be deployed adequately for diverse scenarios. Concretely, we formulated a game only for one privacy property and did not study the Price of Stability/Anarchy. Concerning differential privacy, our assumption about the IID nature of the edges in a communication graph is non-realistic. At the same time, the time-dependent aspect of the messages is not incorporated in our analysis via Pan-Privacy.

References

1. Elliot Anshelevich, Anirban Dasgupta, Jon Kleinberg, Eva Tardos, Tom Wexler, and Tim Roughgarden. The price of stability for network design with fair cost allocation. *SIAM Journal on Computing*, 38(4):1602–1623, 2008.
2. Michael Backes, Aniket Kate, Praveen Manoharan, Sebastian Meiser, and Esfandiar Mohammadi. Anoa: A framework for analyzing anonymous communication protocols. In *2013 IEEE 26th Computer Security Foundations Symposium*, pages 163–178. IEEE, 2013.
3. Gabrielle Beck, Julia Len, Ian Miers, and Matthew Green. Fuzzy message detection. *IACR eprint*, 2021.
4. Matthew Solomon Ben DiFrancesco. Privacy preserving stealth payments on the ethereum blockchain, 2021.
5. Giuseppe Bianchi, Lorenzo Bracciale, and Pierpaolo Loreti. " better than nothing" privacy with bloom filters: To what extent? In *International Conference on Privacy in Statistical Databases*, pages 348–363. Springer, 2012.
6. Gergely Biczók and Pern Hui Chia. Interdependent privacy: Let me share your data. In *International conference on financial cryptography and data security*, pages 338–353. Springer, 2013.
7. Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer, 2013.
8. Henry de Valence. Determine whether penumbra could integrate fuzzy message detection, 2021.
9. Damien Desfontaines and Balázs Pejó. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2:288–313, 2020.
10. Josep Domingo-Ferrer and Vicenç Torra. A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pages 990–993. IEEE, 2008.
11. Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.
12. Cynthia Dwork. Differential privacy in new settings. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2010.
13. Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 2010.
14. Cynthia Dwork, Moni Naor, Toniann Pitassi, Guy N Rothblum, and Sergey Yekhanin. Pan-private streaming algorithms. In *ICS*, 2010.
15. Garrett Hardin. The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *science*, 162(3859):1243–1248, 1968.

16. John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

17. Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178. IEEE, 2009.

18. Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31St international conference on data engineering*, pages 1023–1034. IEEE, 2015.

19. Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. Differentially private event sequences over infinite streams. *Proceedings of the VLDB Endowment*, 2014.

20. Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011.

21. Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pages 171–180, 2009.

22. Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *Stacs*. Springer, 1999.

23. Sarah Jamie Lewis. Niwl: a prototype system for open, decentralized, metadata resistant communication using fuzzytags and random ejection mixers, 2021.

24. Sarah Jamie Lewis. A playground simulator for fuzzy message detection, 2021.

25. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

26. Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996.

27. John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

28. Noam Nisan, Michael Schapira, Gregory Valiant, and Aviv Zohar. Best-response mechanisms. In *ICS*, pages 155–165. Citeseer, 2011.

29. Shen Noether. Ring signature confidential transactions for monero. *IACR Cryptol. ePrint Arch.*, 2015:1098, 2015.

30. Pietro Panzarasa, Tore Opsahl, and Kathleen M Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5):911–932, 2009.

31. Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610, 2017.

32. Antoine Rondelet. Fuzzy message detection in zeth, 2021.

33. Eli Ben Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE Symposium on Security and Privacy*, pages 459–474. IEEE, 2014.

34. Herbert A Simon. Altruism and economics. *The American Economic Review*, 83(2):156–161, 1993.

35. Aikaterini Soumelidou and Aggeliki Tsohou. Towards the creation of a profile of the information privacy aware user through a systematic literature review of information privacy awareness. *Telematics and Informatics*, page 101592, 2021.

36. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
37. Shun Takagi, Yang Cao, and Masatoshi Yoshikawa. Asymmetric differential privacy. *arXiv preprint arXiv:2103.00996*, 2021.
38. Tao Zhang, Tianqing Zhu, Renping Liu, and Wanlei Zhou. Correlated data in differential privacy: Definition and analysis. *Concurrency and Computation: Practice and Experience*, page e6015, 2020.
39. Bin Zhou and Jian Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and information systems*, 28(1):47–77, 2011.

A   FMD in more details

The fuzzy message detection scheme consists of the following five probabilistic polynomial-time algorithms (Setup, KeyGen, Flag, Extract, Test). In the following, let $\mathcal{P}$ denote the set of attainable false positive rates.

Setup$(1^\lambda) \xrightarrow{\$}$ pp. Global parameters pp of the FMD scheme are generated, i.e., the description of a shared cyclic group.

KeyGen$_{pp}(1^\lambda) \xrightarrow{\$} (pk, sk)$. This algorithm is given the global public parameters and the security parameter and outputs a public and secret key.

Flag$(pk) \xrightarrow{\$} C$. This randomized algorithm given a public key pk outputs a flag ciphertext C.

Extract$(sk, p) \rightarrow dsk$. Given a secret key sk and a false positive rate p the algorithm extracts a detection secret key dsk iff. $p \in \mathcal{P}$ or outputs $\perp$ otherwise.

Test$(dsk, C) \rightarrow \{0, 1\}$. The test algorithm given a detection secret key dsk and a flag ciphertext C outputs a detection result.

An FMD scheme needs to satisfy three main security and privacy notions: correctness, fuzziness and detection ambiguity. For the formal definitions of these, we refer to [3]. The toy example presented in Figure 5 is meant to illustrate the interdependent nature of the privacy guarantees achieved by the FMD scheme.
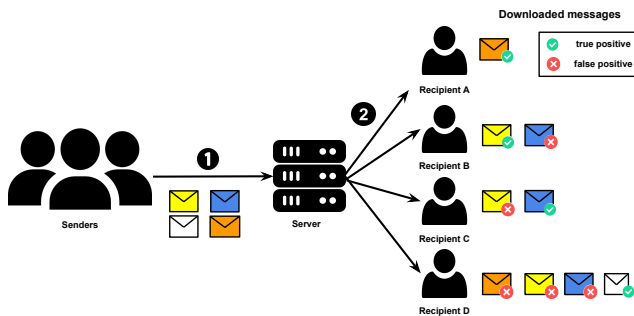
Fig. 5: A toy example of the FMD scheme. ❶ Several senders post anonymous messages to the untrusted server. ❷ Whenever recipients come online, they download messages that correspond to them (some false positve, some true positive). Recipient A,B,C and D have a false positive rate $0, \frac{1}{3}, \frac{1}{3}, 1$, respectively. Note that the server can map the messages that belong to A and D. However, the messages of Recipient B and C are 2-anonymous.

B   Formal definitions of security and privacy guarantees

The recipient unlinkability $\mathcal{G}_{\mathcal{A},\Pi}^{RU}(\lambda)$ game

1. Adversary $\mathcal{A}$ selects target recipients u_0, u_1 and a target sender u_2.
2. Challenger $\mathcal{C}$ instructs sender u_2 to send a message to u_c for $c \xleftarrow{\$} \{0,1\}$.
3. $\mathcal{C}$ uniformly at random generates a challenge bit $b \xleftarrow{\$} \{0,1\}$. If $b = 0$, $\mathcal{C}$ instructs u_2 to send a message to u_c. Otherwise, instructs u_2 to send a message to u_{1-c}.
4. $\mathcal{A}$ observes network traffic and flag ciphertexts and outputs b'.
5. Output 1, iff. $b = b'$, otherwise 0.

Fig. 6: The security game for the anonymity notion of recipient unlinkability.

Definition 4 (Temporal Detection Ambiguity). *An anonymous communication protocol Π satisfies temporal detection ambiguity if for all probabilistic polynomial-time adversaries $\mathcal{A}$ there is a negligible function $\mathsf{negl}(\cdot)$ such that*

$$\Pr[\mathcal{G}_{\mathcal{A},\Pi}^{TDA}(\lambda) = 1] \leq \frac{1}{2} + \mathsf{negl}(\lambda), \tag{6}$$

where the temporal detection ambiguity game $\mathcal{G}_{\mathcal{A},\Pi}^{TDA}(\cdot)$ is defined below.

The temporal detection ambiguity $\mathcal{G}_{\mathcal{A},\Pi}^{TDA}(\lambda)$ game

1. Adversary $\mathcal{A}$ selects a target recipient u_0.
2. Challenger $\mathcal{C}$ uniformly at random generates a challenge bit $b \xleftarrow{\$} \{0, 1\}$. If $b = 0$, $\mathcal{C}$ picks $k \xleftarrow{\$} [1, 2, \ldots, U]$ and instructs sender u_k to send a message to u_0. Otherwise, the challenger does nothing.
3. The anonymous communication protocol Π remains functional for a certain period of time, i.e., users keep sending messages using Π.
4. $\mathcal{A}$ observes network traffic and flag ciphertexts and outputs b'.
5. Output 1, iff. $b = b'$, otherwise 0.

Fig. 7: The security game for the privacy notion of temporal detection ambiguity

C   Differential Privacy Relaxations & Proofs

Our novel DP notion called PEEDP (short for Personalized Existing Edge DP) is an instance of d-privacy [7], which generalizes the neighbourhood of datasets (on which the DP inequality should hold) to an arbitrary metric d defined over the input space. Yet, instead of a top-down approach where we are presenting a complex metric to fit our FMD use-case, we follow a bottom-up approach and show the various building blocks of our definition. PEEDP is a straight forward combination of unbounded DP [20], edge-DP [17]), assymetric DP [37], and personalized DP [18]. Although Definition 3 is appropriate for FMD, it does not capture the FMD scenarios fully as neither time-dependent nature of the messages nor the dependencies and correlations between them are taken into account.

The first issue can be tackled by integrating other DP notions into PEEDP which provide guarantees under continuous observation (i.e., stream-data), such as pan-privacy [14]. Within this streaming context several definitions can be considered: user-level [13] (to protect the presence of users), event-level [12] (to protect the presence of messages), and w-event level [19] (to protect the presence of messages within time windows).

The second issue is also not considered in Theorem 1 as we assumed the messages are IID, while in a real-world applications this is not necessarily the case. Several DP notions consider distributions, without cherry-picking any we refer the readers to two corresponding surveys [9,38]. We leave it as a future work to tweak our definition further to fit into these contexts.

Proof (of Theorem 1). Due to the IID nature of the messages it is enough to show that Equation 5 holds for an arbitrary communication graph D with arbitrary message m of an arbitrary user u. The two possible world the adversary should not be able to differentiate between $D = D'/\{m\}$, i.e., whether the particular message exists or not. Due to the asymmetric nature of Definition 3 (i.e.,

it only protects the existence) Equation 7 does not need to be satisfied. On the other hand, if the message exists than Equation 8 and 9 must be satisfied where $S_1 =$ {message m is downloaded by user u} and $S_2 =$ {message m is not downloaded by user u}.

$$\Pr(A(D') \in S) \leq e^{\varepsilon_u} \cdot \Pr(A(D) \in S) \tag{7}$$

$$\Pr(A(D) \in S_1) \leq e^{\varepsilon_u} \cdot \Pr(A(D') \in S_1) \tag{8}$$

$$\Pr(A(D) \in S_2) \leq e^{\varepsilon_u} \cdot \Pr(A(D') \in S_2) \tag{9}$$

If we reformulate the last two equations with the corresponding probabilities we get $1 \leq e^{\varepsilon_u} \cdot p(u)$ and $0 \leq e^{\varepsilon_u} \cdot (1 - p(u))$ respectively. While the second holds trivially the first corresponds to the formula in Theorem 1. $\qquad\square$

Proof (of Theorem 2). The users' number of incoming messages are independent from each other hence we can focus on a single user u. The proof follows the idea from [21][14]: we satisfy Equation 4 (with $+\delta$ at the end) when $A(D) = tag(u) \sim D + \mathsf{Binom}(M - in(u), p(u))$ for $D = in(u)$ and $D' = in(u) \pm 1$, i.e., we show that the following Equation holds.

$$\Pr(A(D) = tag(u) \in S | D = in(u), M, p(u)) \leq$$
$$e^{\varepsilon} \cdot \Pr(A(D') = tag'(u) \in S | D' = in(u) \pm 1, M' = M \pm 1, p(u)) + \delta$$
$$\Rightarrow \qquad \Pr(in(u) + \mathsf{Binom}(M - in(u), p(u)) \in S) \leq$$
$$e^{\varepsilon} \cdot \Pr(in(u) \pm 1 + \mathsf{Binom}(M \pm 1 - (in(u) \pm 1), p(u)) \in S) + \delta$$

First, we focus on δ and provide a lower bound originating from the probability on the left when $\Pr(\cdot) \leq e^{\varepsilon} \cdot 0 + \delta$. This corresponds to two cases as seen in the Equation below: when $D' = in(u) + 1$ with $S = \{in(u)\}$ and when $D' = in(u) - 1$ with $S = \{M\}$. The corresponding lower bounds (i.e., probabilities) correspond to the event when user u does not download any fuzzy messages and when user u does downloads all messages respectively. Hence, the maximum of these are indeed a lower bound for δ.

$$\Pr(A(in(u)) = in(u)) \leq e^{\varepsilon} \cdot \Pr(A(in(u) + 1) = in(u)) + \delta \Rightarrow (1 - p(u))^{M - in(u)} \leq \delta$$
$$\Pr(A(in(u)) = M) \leq e^{\varepsilon} \cdot \Pr(A(in(u) - 1) = M) + \delta \quad \Rightarrow \quad p(u)^{M - in(u)} \leq \delta$$

Now we turn towards ε and show that $(\varepsilon, 0)$-DP holds for all subset besides the two above, i.e., when $S = \{in(u) + y\}$ with $y = [1, \ldots, M - in(u) - 1]$. First, we reformulate Equation 4 as seen below.

$$\frac{\Pr(in(u) + \mathsf{Binom}(M - in(u), p(u)) \in S)}{\Pr(in(u) \pm 1 + \mathsf{Binom}(M - in(u), p(u)) \in S)} \leq e^{\varepsilon}$$

---

[14] We present the proof for singleton sets, but it can be extended by using the following formula: $\frac{A+C}{B+D} < \max(\frac{A}{B}, \frac{C}{D})$.

Then, by replacing the binomial distributions with the corresponding probability formulas we get the following two equations for $D' = in(u) + 1$ and $D' = in(u) - 1$ respectively.

$$\frac{\binom{M-in(u)}{y} \cdot p(u)^y \cdot (1-p(u))^{M-in(u)-y}}{\binom{M-in(u)}{y-1} \cdot p(u)^{y-1} \cdot (1-p(u))^{M-in(u)-y+1}} = \frac{M - in(u) - y + 1}{y} \cdot \frac{p(u)}{1 - p(u)} \leq e^\varepsilon$$

$$\frac{\binom{M-in(u)}{y} \cdot p(u)^y \cdot (1-p(u))^{M-in(u)-y}}{\binom{M-in(u)}{y+1} \cdot p(u)^{y+1} \cdot (1-p(u))^{M-in(u)-y-1}} = \frac{y + 1}{M - in(u) - y} \cdot \frac{1 - p(u)}{p(u)} \leq e^\varepsilon$$

Consequently, the maximum of these is the lower bound for e^ε. The first formula's derivative is negative, so the function is monotone decreasing, meaning that its maximum is at $y = in(u) + 1$. On the other hand, the second formula's derivative is positive so the function is monotone increasing, hence the maximum is reached at $y = M - in(u) - 1$. By replacing y with these values respectively one can verify that the corresponding maximum values are indeed what is shown in Theorem 2. □

D   Game-Theoretical Analysis

Here — besides a short introduction of the utilized game theoretic concepts — we present a rudimentary game-theoretic study of the FMD protocol focusing on relationship anonymity introduced in Section 4. First, we formalize a game and highlight some corresponding problems such as the interdependence of the user's privacy. Then, we unify the user's actions and show the designed game's only Nash Equilibrium, which is to set the false-positive detection rates to zero, rendering FMD idle amongst selfish users. Following this, we show that a higher utility could been reached with altruistic users and/or by centrally adjusting the false-positive detection rates. Finally, we show that our game (even with non-unified actions) is a potential game, which have several nice properties, such as efficient Nash Equilibrium computation.

- **Tragedy of Commons** [15]: users act according to their own self-interest and, contrary to the common good of all users, cause depletion of the resource through their uncoordinated action.
- **Nash Equilibrium** [27]: every player makes the best/optimal decision for itself as long as the others' choices remain unchanged.
- **Altruism** [34]: users act to promote the others' welfare, even at a risk or cost to ourselves.
- **Social Optimum** [16]: the user's strategies which maximizes social welfare (i.e., the overall accumulated utilities).
- **Price of Stability/Anarchy** [1,22]: the ratio between utility values corresponding to the best/worst NE and the SO. It measures how the efficiency of a system degrades due to selfish behavior of its agents.

– **Best Response Mechanism** [28]: from a random initial strategy the players iteratively improve their strategies

Almost every multi-party interaction can be modeled as a game. In our case, these decision makers are the users using the FMD service. We assume the users bear some costs C_u for downloading any message from the server. For simplicity we define this uniformly: if f is the cost of retrieving any message for any user than $C_u = f \cdot tag(u)$. Moreover, we replace the random variable $tag(u) \sim in(u) + \mathsf{Binom}(M - in(u), p(u))$ with its expected value, i.e., $C_u = f \cdot (in(u) + p(u) \cdot (M - in(u)))$.

Besides, the user's payoff should depend on whether any of the privacy properties detailed in Section 4 are not satisfied. For instance, we assume the users suffer from a privacy breach if relationship anonymity is not ensured, i.e., they uniformly lose L when the recipient u can be linked to any sender via any message between them. In the rest of the section we slightly abuse the notation u as in contrast to the rest of the paper we refer to the users as $u \in \{1, \ldots, U\}$ instead of $\{u_0, u_1, \ldots\}$. The probability of a linkage via a particular message for user u is $\alpha_u = \prod_{v \in \{1,\ldots,U\}/u}(1 - p(v))$. The probability of a linkage from any incoming message of u is $1 - (1 - \alpha_u)^{in(u)}$.[15] Based on these we define the FMD-RA Game.

Definition 5. *The FMD-RA Game is a tuple $\langle \mathcal{N}, \Sigma, \mathcal{U} \rangle$, where the set of players is $\mathcal{N} = \{1, \ldots, U\}$, their actions are $\Sigma = \{p(1), \ldots, p(U)\}$ where $p(u) \in [0,1]$ while their utility functions are $\mathcal{U} = \{\varphi_u(p(1), \ldots, p(U))\}_{u=1}^{U}$ such that for $1 \leq u \leq U$:*

$$\varphi_u = -L \cdot \left(1 - (1 - \alpha_u)^{in(u)}\right) - f \cdot (in(u) + p(u) \cdot (M - in(u))). \qquad (10)$$

It is visible in the utility function that the bandwidth-related cost (second term) depends only on user u's action while the privacy-related costs (first term) depend only on the other user's actions. This reflects well that relationship anonymity is an interdependent privacy property [6] within FMD: by downloading fuzzy tags, the users provide privacy to others rather than to themselves. As a consequence of this tragedy-of-commons [15] situation, a trivial no-protection Nash Equilibrium (NE) emerges. Moreover, Theorem 3 also states this NE is unique, i.e., no other NE exists.

Theorem 3. *Applying no privacy protection in the FMD-RA Game is the only NE: $(p^*(1), \ldots, p^*(U)) = (0, \ldots, 0)$.*

Proof. First we prove that no-protection is a NE. If all user u set $p(u) = 0$ than a single user by deviates from this strategy would increased its cost. Hence no rational user would deviate from this point. In details, in Equation 10 the privacy related costs is constant $-L$ independently from user u's false-positive

---

[15] It is only an optimistic baseline as it merely captures the trivial event when no-one downloads the a message from any sender v besides the intended recipient u.

rate while the download related cost would trivially increase as the derivative of this function (shown in Equation 11) is negative.

$$\frac{\partial \varphi_u}{\partial p(u)} = -f \cdot (M - in(u)) < 0 \qquad (11)$$

Consequently, $p^* = (p^*(1), \ldots, p^*(U)) = (0, \ldots, 0)$ is indeed a NE. Now we give an indirect reasoning why there cannot be any other NEs. Lets assume $\hat{p} = (\hat{p}(1), \ldots, \hat{p}(U))$ is a NE. At this state any player could decrease its cost by reducing its false positive-rate which only lower the download related cost. Hence, $\hat{p}$ is not an equilibrium. $\qquad \square$

This negative result highlights that in our simplistic model, no rational (selfish) user would use FMD; it is only viable when altruism [34] is present. On the other hand, (if some condition holds) in the Social Optimum (SO) [16], the users do utilize privacy protection. This means a higher total payoff could be achieved (i.e., greater social welfare) if the users cooperate or when the false-positive rates are controlled by a central planner. Indeed, according to Theorem 4 the SO≠NE if, for all users, the cost of the fuzzy message downloads is smaller than the cost of the privacy loss. The exact improvement of the SO over the NE could be captured by the Price of Stability/Anarchy [1,22], but we leave this as future work.

Theorem 4. *The SO of the FMD-RA Game is not the trivial NE and corresponds to higher overall utilities if $f \cdot (M - \max_u(in(u))) < L$.*

Proof. We show that the condition in the theorem is sufficient to ensure that SO≠NE by showing that greater utility could be achieved with $0 < p'(u)$ than with $p(u) = 0$. To do this we simplify out scenario and set $p(u) = p$ for all users. The corresponding utility function is presented in Equation 12 while in Equation 13 we show the exact utilities when p is either 0 or 1.

$$\varphi_u(p) = -L \cdot (1 - (1 - (1 - p)^{U-1})^{in(u)}) - f \cdot (in(u) + p \cdot (M - in(u))) \quad (12)$$

$$\varphi_u(0) = -L - f \cdot in(u) \qquad \varphi_u(1) = -f \cdot M \qquad (13)$$

One can check with some basic level of mathematical analysis that the derivative of Equation 12 is negative at both edge of $[0,1]$ as $\frac{\partial \varphi_u(p)}{\partial p}(0) = \frac{\partial \varphi_u(p)}{\partial p}(1) = -f \cdot (M - in(u))$. This implies that the utility is decreasing at these points. Moreover, depending on the relation between the utilities in Equation 13 (when $p = 0$ and $p = 1$), two scenario is possible as we illustrate in Figure 8. From this figure it is clear that when $\varphi_u(0) < \varphi_u(1)$ (or $f \cdot (M - in(u)) < L$) for all users that the maximum of their utilities cannot be at $p = 0$. $\qquad \square$

Potential Game. We also show that FMD-RA is a potential game [26]. This is especially important, as it guaranteed that the Best Response Dynamics terminates in a NE.
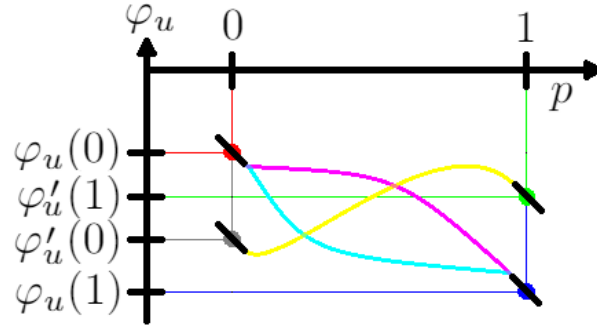
Fig. 8: Illustration of the utility functions: the yellow curve's maximum must be between zero and one since the gray dot is below the green where the derivative is negative.

Definition 6 (Potential Game). *A Game $\langle \mathcal{N}, \mathcal{A}, \mathcal{U} \rangle$ (with players $\{1, \ldots, U\}$, actions $\{a_1, \ldots, a_U\}$, and utilities $\{\varphi_1, \ldots, \varphi_U\}$) is a Potential Game if there exist a potential function Ψ such that Equation 14 holds for all players u independently of the other player's actions.*[16]

$$\varphi_u(a_u, a_{-u}) - \varphi_u(a'_u, a_{-u}) = \Psi(a_u, a_{-u}) - \Psi(a'_u, a_{-u}) \tag{14}$$

Theorem 5. *FMD-RA is a Potential Game with potential function shown in Equation 15.*

$$\Psi(p(1), \ldots, p(U)) = -f \cdot \sum_{u=1}^{U} p(u) \cdot (M - in(u)) \tag{15}$$

Proof. We prove Equation 14 by transforming both side to the same form. We start with the left side: the privacy related part of the utility does only depend on the other user's action, therefore this part falls out during subtraction. On the other hand the download related part accumulates as shown below.

$$\varphi_u(p(u), (p(-u)) - \varphi_u(p(u)', p(-u)) =$$
$$-f \cdot (in(u) + p(u) \cdot (M - in(u))) - (-f \cdot (in(u) + p(u)' \cdot (M - in(u)))) =$$
$$-f \cdot p(u) \cdot (M - in(u)) - (-f \cdot p(u)' \cdot (M - in(u)))$$

Coincidentally, we get the same result if we do the subtraction on the right side using the formula in Equation 15 as all element in the summation besides u falls out (as they are identical because they do not depend on user u's action). □

---

[16] a_{-u} is a common notation to represent all other players action except player u. Note that $p(-u)$ stands for the same in relation with FMD.

E   Attacks on Privacy

We show several possible attacks against the FMD scheme, that might be fruitful to be analyzed in more depth.

Intersection Attacks.   The untrusted server could possess some background knowledge that it allows to infer that some messages were meant to be received by *the same recipient*. In this case, the server only needs to consider the intersection of the anonymity sets of the "suspicious" messages. Suppose the server knows that l messages are sent to the same user. In that case, the probability that a user is in the intersection of all the l messages' anonymity sets is drawn from the $\mathsf{Binom}(U, p^l)$ distribution. Therefore, the expected size of the anonymity set after an intersection attack is reduced to $p^l U$ from pU.

Sybil attacks.   The collusion of multiple nodes would decrease the anonymity set of a message. For instance, when a message is downloaded by K nodes out of U, and N node is colluding, then the probability of pinpointing a particular message to a single recipient is $\frac{\binom{N+1}{K}}{\binom{U}{K}}$. This probability clearly increases as more node is being controlled by the adversary. On the other hand, controlling more nodes does trivially increase the controller's privacy (not message-privacy but user-privacy) as well. However, formal reasoning would require a proper definition for both of these privacy notions.

Neighborhood attacks.   Neighborhood attacks had been introduced by Zhou et al. in the context of deanonymizing individuals in social networks [39]. An adversary who knows the neighborhood of a victim node could deanonymize the victim even if the whole graph is released anonymously. FMD is susceptible to neighborhood attacks, given that relationship anonymity can be easily broken with statistical tests. More precisely, one can derive first the social graph of FMD users and then launch a neighborhood attack to recover the identity of some users.