A Practical Complexity-Theoretic Analysis of Mix Systems

Dang Vinh Pham[1], Joss Wright[2], and Dogan Kesdogan[1]

[1] Siegen University, Siegen, Germany
[2] Oxford Internet Institute, University of Oxford, Oxford, United Kingdom

Abstract. The *Minimal-Hitting-Set attack* (HS-attack) [10] is a well-known passive intersection attack against Mix-based anonymity systems, applicable in cases where communication behaviour is non-uniform and unknown. The attack allows an observer to identify uniquely the fixed set of communication partners of a particular user by observing the messages of all senders and receivers using a Mix. Whilst the attack makes use of a provably minimal number of observations, it also requires solving an NP-complete problem. No prior research, to our knowledge, analyses the *average* complexity of this attack as opposed to its worst case.

We choose to explore the HS-attack, as opposed to statistical attacks, to provide a baseline metric and a practical attack for unambiguously identifying anonymous users. We show that the average complexity of the HS-attack can vary between a worst-case exponential complexity and a linear-time complexity according to the Mix parameters. We provide a closed formula for this relationship, giving a precise measure of the resistance of Mixes against the HS-attack in practice, and allowing adjustment of their parameters to reach a desired level of strength.

1 Introduction

Modern research into network-level anonymity is widely regarded to have begun in 1981 with the introduction of the Mix by Chaum [3]. The Mix hides the linkage between senders and recipients of messages by ensuring that all senders and recipients are a member of some *anonymity set*.

The concepts underlying the Mix remain the basis for a wide variety of practical and theoretical anonymity systems. Chaum's model of the Mix, in its pure theoretical form, provides an upper limit to what is achievable by these approaches and thus remains an important subject for analysis. The work presented here provides insight into the limits of the Mix model in practical use, and thus aims to guide choices involved in building real-world implementations of Mix variants.

Although a Mix provides unlinkability between input and output messages with respect to a *global passive attacker*, it cannot protect the links between senders and recipients against *long term traffic analysis* attacks when the sender group is *open* [2, 9].

The anonymity property can be modelled with standard security techniques: a global passive attacker model, which provides a strong but realistic adversary, and with the creation of *anonymity sets* as a basis for the anonymity property that we seek to enforce. We consider in this paper an abstract model called the *Pure Mix* [9] that can be used

V. Atluri and C. Diaz (Eds.): ESORICS 2011, LNCS 6879, pp. 508–527, 2011.
© Springer-Verlag Berlin Heidelberg 2011

to model more complex practical Mixes [14, 15]. Analysis of this model is believed to be applicable to other Mix models, with appropriate modifications, but these are not addressed here.

Berthold et al. [2] introduced a class of long term traffic analysis attacks on the Pure Mix, called *intersection attacks*, proving that Mixes with open sender groups cannot provide long term unlinkability if each sender repeatedly communicates with a fixed recipient. In practical usage, however, a sender may have several communication partners and a less restrictive model is therefore needed. Kesdogan et al. introduced the Disclosure [1] and Minimal-Hitting-Set (*HS*-) attacks [9, 10] for repeated communication with an arbitrary fixed set of recipients.

These attacks exploit the fact that a global passive attacker can selectively observe only *recipient anonymity sets* in which a particular sender, referred to as *Alice*, contributes a message. Given sufficiently many observations, Alice's recipient set is the smallest unique set intersecting each of the observations: the *unique minimal hitting set*. Unfortunately, computing this minimal set is known to be an NP-complete problem [8]. Many popular current attacks against Mixes analyse statistical properties [4, 5, 6, 11, 16] to deduce the most likely senders of given messages. These attacks, whilst allowing a level of innaccuracy in results, have the advantage of being much more efficiently computable.

We choose to focus on the Minimal-Hitting-Set attack for a number of reasons: firstly, as the HS-attack uses a provably minimal number of observations [9], it provides an important theoretical baseline for exact identification of a sender's recipients, resulting a metric for Mix anonymisation. Secondly, as we show, it remains a genuinely practical attack in many cases. As the attack is applicable to any distribution of user communications, even when this distribution is unknown to the attacker [10], it applies in many situations with unknown communication behaviour.

Finally, the underlying algorithmic structures related to the HS-attack, based on an NP-complete problem, are themselves an important topic. The research presented here sheds light on analysing the average case complexity of NP-complete problems, and thus those cases in which such problems are computationally tractable.

Contribution. This paper contributes, to our knowledge, the first robust and detailed security analysis of the Mix system based on the average computation required to unambiguously identify users with a provably minimal number of observations. It derives, for a given set of Mix parameters, a direct relationship between the number of observations required for identification and the average-case runtime complexity of the attack.

Our analysis is applicable to non-uniform user communication, and allows us to identify Mix parameters for which unambiguous identification of recipients is intractable in the average case. We also identify instances in which recipients of a sender can be efficiently identified by the HS-attack, providing a provably correct alternative to the more popular statistical approaches.

We show that the NP-completeness of the algorithm deployed by the HS-attack represents only a *worst-case* attack complexity, which provides a poor characterisation when considering systems where the *average* time to failure is of greater relevance.

In this work, therefore, we explore the complexity *structure* of an exact attack in order to determine the average-case complexity, and provide closed formulas that show the relation between the parameters of the Mix and the average-case complexity required to compromise its anonymity.

The Mix model we employ has been used to model practical Mix implementations such as Mixminion and Mixmaster [7, 15], as well as in other analyses [2, 1, 4, 9, 5, 16, 13].

Related Work. A security metric makes a quantitative statement concerning the resistance of a system to an attacker. Our attack model consists of a passively observing attacker against a given anonymity system. The attack that we consider relies solely on observations of this anonymity system. The success of the attack is therefore dependent on the attacker's *knowledge*, and knowledge gain, and on its *computational capabilities*. This form of attack is well known as *passive traffic analysis attack* in the literature. These attacks are hard to thwart, as they exploit the information leakage inherent in all anonymity systems.

Statistical long-term traffic analysis attacks are closest to our approach. These address cases in which Alice's communication behaviour reveals statistical patterns that allow identification of her likely recipients. By relaxing the requirement for absolute correctness, these attacks gain significant computational efficiency.

Greedy variants of the HS-attack, the SHS- and HS*-attacks, were suggested in [10]. These compute hitting sets guided by the frequency with which a peer was contacted while under observation. The result of the SHS-attack is a hitting set that is consistent with the observations made by the attacker, but which can miss Alice's real recipients (*exclusion-error*) or contain recipients not contacted by Alice (*inclusion-error*). In contrast to the SHS-attack, the HS*-attack accepts the greedily computed hitting set only if it is a unique smallest minimal hitting set. This attack can identify Alice's recipient set, but risks producing no result.

The Statistical-Disclosure attack (SDA) [4, 11, 6, 5] introduces the class of statistical attacks that focus on the likelihood that a single recipient is in Alice's recipient set. These attacks typically assume some knowledge of the distribution of communications amongst untargeted users, which must remain static during the attack. This approach introduces the possibility of both inclusion and exclusion errors, but results in much more efficient attacks. While both approaches provide advantages, a comparison of their relative effectiveness is beyond the scope of this work, and we will not discuss these attacks further.

One attack of note is the Perfect-Matching-Disclosure-attack (PMDA) [16], which applies statistical attacks to successively weight links between all senders and receivers in a Mix network. This more sophisticated statistical attack builds user communication-pattern profiles to inform its inferences, and allows for tradeoffs between accuracy and speed in disclosing communication links. Again, however, the nature and effectiveness of this attack is largely out of the scope of the current work, which focuses exclusively on provably correct attacks in order to provide a baseline *metric* for anonymity in Mixes.

Structure. Section 2 presents the Mix and attacker model used in this paper. The attack that we present is based on the *ExactHS algorithm* [12, 13] that computes all *minimal hitting sets*, in this case for sets of a user's possible communication partners. The results in this paper enable us to determine the average-case complexity of this algorithm, and thus the average complexity of unambiguously identifying a user's set of communication partners.

Proving the identity of a user's communication partner set is equivalent to proving that all other possible sets of recipients *cannot* be the user's partner set: a *disproof* of these sets. Section 3 presents our theoretical model that describes the number of peers in a possible set of recipients that must be considered in order to disprove it. The average worst case of this number of peers is derived in Sect. 4.

Section 5 applies our analyses to the ExactHS algorithm in order to obtain formulas for the average worst case complexity of identifying peers, and shows how this relates to the required number of observations. To support our theoretical results, we compare our analysis to simulations in Sect. 6. We provide conclusions and ideas for future work in Sect. 7.

2 Mix and Attacker Model

2.1 The Pure Mix Model

We consider the *Pure Mix* technique, as justified in [9], as a generalised and simplified model of practical real-world Mixes.

Our attacker model is that of a *global passive attacker* that observes all communication, but cannot inject, delay or alter messages. From this basis, we will use the following formal model of a pure Mix and information leakage for our analysis.

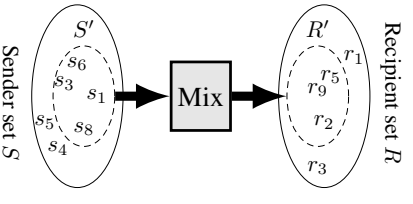

Fig. 1. Mix model

Formal Model of the Pure Mix Technique.

- A communication system consists of a set of senders, S, a set of recipients, R, and a Mix node[1] as shown in Fig. 1. If a sender $s \in S$ communicates with a recipient $r \in R$, then we say that r is a *peer partner* of s, or simply r is a *peer* of s.
- In each *communication round*[2] a subset $S' \subseteq S$ of all senders each send precisely one message to their peer partners. Let $R' \subseteq R$ be the set of intended recipients. The act of sending or receiving a message is not hidden to the attacker, therefore (S', R') represents the information leakage available to an attacker in each round.[3]

[1] S and R represent all users with the ability to send or receive messages in the system.

[2] A communication round consists of the Mix node collecting messages from a fixed number of distinct senders and, after applying the "Mix" protocol, forwarding the collected messages in random order to their intended recipients.

[3] Note that a sender can send to multiple recipients in distinct rounds, but cannot send multiple messages in a single round.

- We call S' the *sender anonymity set*, which is the set of all senders that may have sent a given message. The *recipient anonymity set* R' is the set of all recipients that may have received a message.
- We label the size of the *sender anonymity set*, $|S'|$, as b.
- The size of the *recipient anonymity set*, $|R'|$, is less than or equal to b, as each sender sends exactly one message per round but several senders may communicate with the same recipient. The size of the set of all recipients is $|R| = N$.

Attacker Model. The goal of the attacker is to compute, from a set of observations of traffic, all possible sets of peer partners of a target sender $Alice \in S$. These possibilities form *hypotheses* for the true set of Alice's peer partners, $\mathcal{H}_A$, which is assumed to be a fixed set of size $m = |\mathcal{H}_A|$. We call a peer $r \in \mathcal{H}_A$ an *Alice's peer*; a peer that does not communicate with Alice, $r \in R \setminus \mathcal{H}_A$, is called a *non-peer* and r is simply called *peer* if no distinction is required.

The attacker focuses on revealing Alice's peers by observing only those pairs (S', R'), where Alice participates as a sender. Under this condition we refer to the corresponding recipient set R' as an *observation*, $\mathcal{O}$. The set of all observations collected during t communication rounds is referred to as the *observation set* $\mathcal{OS} = \{\mathcal{O}_1, \ldots, \mathcal{O}_t\}$.

Alice's peer set can be revealed by the *Minimal-Hitting-Set attack* (HS-attack) [10], which computes all hypotheses from the set of observations. These hypotheses correspond to all sets of size m that are *hitting sets* in $\mathcal{OS}$. A hitting set is a set that intersects with all observations in $\mathcal{OS}$. A hitting set is *minimal* if no proper subset of it is a hitting set. The HS-attack succeeds if $\mathcal{OS}$ is consistent with only a single hypothesis. In this case Alice's peer set is unambiguously identified, and is thus the smallest *unique minimal hitting set* of size m. This attack has been proven to require a minimal number of observations to identify $\mathcal{H}_A$[9].

In applying the HS-attack, we assume that the size of Alice's peer set, m, is known, since learning m does not change the complexity class of the attack.

Learning m. The intuition behind our attack is that at least one of Alice's peers must appear in each observation[4], while this does not hold for any other set $\mathcal{H}$, where $\mathcal{H}_A \not\subseteq \mathcal{H}$. Therefore, after a large number of observations, t, Alice's peer set $\mathcal{H}_A$ remains the unique smallest minimal hitting set.

Assume the existence of a set in which $\mathcal{H} \neq \mathcal{H}_A$, where $|\mathcal{H}| < m$ happens to be a unique minimal hitting set. If p is the probability that any peer in $\mathcal{H}$ appears in a random observation, the probability that $\mathcal{H}$ remains a hitting set after t observations decreases according to an exponential function p^t. The probability of learning the wrong set of Alice's peers and the wrong value of m by the HS-attack is therefore negligible even for moderate t.

We can learn m in time $\sum_{m'=1}^{m} O(b^{m'} m' t b) = O(b^m m t b)$ by running the HS-attack for $m' = 1, \ldots, m$ with respect to the same t observations according to equation (2). If $m' < m$, then there will be no hitting sets of size m' and HS-attack thus detects incorrect m'.

Multiple Sending per Round. We assume that each sender sends only one message in each round to simplify the Mix model and our analysis. The HS-attack remains

[4] Recall that an "observation" refers to a round in which Alice participates.

applicable, however, if a sender can send multiple messages per round. This altered model does require slight modifications to the algorithm deployed by HS-attack, and thus a minor modification to the analysis. Due to space limitations, we omit this extended model here. Therefore, investigating the relation between the results from the simple model and from the extended model will be left for future work.

2.2 ExactHS Algorithm

ExactHS [12, 13], described in Alg. 1 determines the *hypotheses* in the Minimal-Hitting-Set attack. Unlike the original HS algorithm proposed in [10], which analyses all $\binom{N}{m}$ hitting sets, ExactHS considers *only* minimal hitting sets and thus drastically reduces computation [12, 13].

The method used by ExactHS to determine hitting sets corresponds to the theoretical model in Sect. 3, allowing us to apply the analyses of Sect. 4 to determine the average case complexity for unambiguously identifying Alice's peer set.

ExactHS recursively computes all minimal hitting sets with respect to the attacker's observation set $\mathcal{OS}$. We use the following notation:

$\mathcal{C}$: Set of at most m suspected[5] peers representing a subset of a possible hitting set. It is initially empty.

$\mathcal{OS}[r]$: Set of observations containing peer r, that is $\{\mathcal{O} \in \mathcal{OS} \mid r \in \mathcal{O}\}$. $|\mathcal{OS}[r]|$ is called the *frequency* of r. $|\mathcal{OS}[r]|$ is 0, if r is not in any observations of $\mathcal{OS}$.

$\mathcal{OS}[\{r_1, \ldots, r_k\}]$: Set of observations containing any $r_1, \ldots, r_k$, that is $\bigcup_{i=1}^{k} \mathcal{OS}[r_i]$.

We now describe in detail the steps taken by ExactHS on a line-by-line basis, as shown in Alg. 1.

Algorithm 1 ExactHS

```
1:  procedure ExactHS(OS', m', C)
2:      if OS' = {} then
3:          return C                                          ▷ C is a hitting set
4:      else if m' ≥ 1 then                                   ▷ add a peer to C, if C contains less than m peers
5:          choose O ∈ OS'
6:          while ({} ∉ OS') ∧ ( max_{r_1,...,r_m'} {Σ_{l=1}^{m'} |OS'[r_l]|} ≥ |OS'|) do
7:              choose r ∈ O                                  ▷ r will become element of C
8:              ExactHS(OS' \ OS'[r], m' − 1, C ∪ {r})        ▷ select remaining (m' − 1) peers of C
9:              OS' ← ⋃_{O_l ∈ OS'} {O_l \ {r}}               ▷ remove r in all observ. of OS'
10:             O ← O \ {r}                                   ▷ do not choose r in this recursion level again
```

The computation of the minimal hitting sets is initially invoked by calling the algorithm $ExactHS(\mathcal{OS}, m, \mathcal{C})$. For ease of reference we denote sets computed in the i-th level of recursion with the subscript i. Thus $\mathcal{C}_i, \mathcal{OS}'_i$ represents the sets calculated by ExactHS at the i-th recursive call of the algorithm. At each level of recursion in the algorithm, recursing to the next level extends the current set of peers $\mathcal{C}_i$ by exactly one peer, r, at Line 7 of Alg. 1. This peer is chosen from a designated observation $\mathcal{O} \in \mathcal{OS}'_i$ determined by the algorithm in Line 5. Thus: $\mathcal{C}_{i+1} = \mathcal{C}_i \cup \{r\}$.

[5] During execution, $\mathcal{C}$ either becomes a minimal hitting set, or it will be proved not to be a subset of any minimal hitting sets.

$\mathcal{OS}'_{i+1}$, defined at Line 8, results from removing all observations intersecting with r in $\mathcal{OS}'_i$; we need only focus on those observations that have not already been evaluated by $\mathcal{C}_i \cup \{r\}$ in earlier recursive calls.

If, at Line 2, the algorithm detects that all remaining observations in $\mathcal{OS}'_{i+1}$ intersect with $\mathcal{C}_{i+1}$, $\mathcal{C}_{i+1}$ is proven to be a hitting set, and ExactHS will not compute any set containing this $\mathcal{C}_{i+1}$ in the future. Line 6 will also detect if $\mathcal{C}_{i+1}$ is not a subset of any hitting set; this also causes any set containing it to be ignored in future levels of recursion. We refer to sets excluded by the algorithm as *finalised* sets.

After a selection of r in recursion level i, ExactHS removes, at Line 9, r from all observations of $\mathcal{OS}'_i$ and, at Line 10, from the designated observation $\mathcal{O}$. The algorithm thus extends $\mathcal{C}_i$ with a new peer r'.

ExactHS stops choosing new peers if it detects, at Line 6, that the cumulative frequency of all remaining m' peers is lower than the number of remaining observations; that is, if $\max_{r_1,\ldots,r_{m'}}\{\sum_{l=1}^{m'} |\mathcal{OS}'[r_l]|\} \ngeq |\mathcal{OS}'|$. Further explanations are in Sect. 5.

Complexity. ExactHS creates a finalised set $\mathcal{C}$ by starting with an empty set $\mathcal{C} = \{\}$ and adding the i-th peer to $\mathcal{C}$ in the *choice phase* of the i-th level of recursion, starting at line 6 of the algorithm. The number of recursive invocations of the choice phase is bounded from above by m.

In each choice phase there are at most b possible choices of a peer r_i, as only peers $r_1, \ldots, r_b$ of a fixed observation $\mathcal{O}$ can be selected. Due to the bound m for the number of recursive invocations of the choice phase, and the bound b for the number of choices in each phase, the algorithm computes at most b^m minimal hitting sets. This bound is tight, and determines the *worst case runtime complexity* $O(b^m mtb)$ of ExactHS, as proved in [12, 13]. $t = |\mathcal{OS}|$ is the number of observations collected by the attacker and mtb is the effort required to construct one finalised set.

Let us consider a concrete example with the parameters $m = 2, b = 2$, the Alice's peer set $\mathcal{H}_A = \{1, 2\}$ and the observations $\{1, 3\}, \{2, 4\}$. Here, ExactHS would compute $b^m = 4$ minimal hitting sets, namely: $\{1, 2\}, \{1, 4\}, \{3, 2\}, \{3, 4\}$.

In general, however, if ExactHS were to prove at level $x \leq m$ that a set is, or is not, a hitting set, then the number of finalised sets computed by ExactHS is bounded from above by (1) and the runtime is bounded by (2). The space complexity of ExactHS, as proved in [12], is $O((x + 1)tb)$, which is linear.

$$\text{Maximal number of sets: } b^x \qquad (1) \qquad\qquad \text{Runtime: } O(b^x mtb) \qquad (2)$$

Hitting Set Structure. In order to make a more detailed analysis of the ExactHS algorithm, we partition the set of minimal hitting sets of size m. Let $\mathcal{H}$ be a minimal hitting set where $|\mathcal{H}| = m$. We therefore assign it to one of the $m + 1$ disjoint classes $\mathfrak{H}_0, \ldots \mathfrak{H}_m$ with the following structure:

$$\mathfrak{H}_0 = \{\mathcal{H}_A\} \qquad \text{and} \qquad \mathfrak{H}_j \subseteq (R \setminus \mathcal{H}_A)^j \times \mathcal{H}_A^{m-j} \text{ , for } j \leq m \ . \qquad (3)$$

A minimal hitting set $\mathcal{H}$ belongs to the class $\mathfrak{H}_j$ ($\mathcal{H} \in \mathfrak{H}_j$), if and only if it contains exactly $(m - j)$ distinct Alice's peers and j distinct non-peers. The class $\mathfrak{H}_0$ contains exactly one set, Alice's peer set $\mathcal{H}_A$, and $\mathfrak{H}_m$ represents minimal hitting sets consisting of only non-peers of Alice.

3 Estimation of the Number of Covered Observations

This section focuses on the complexity theoretic security of the Mix. We therefore assume that the observations in $\mathcal{OS}$ collected by the attacker provide sufficient information for the unambiguous identification of Alice's peer set $\mathcal{H}_A$. The main question we wish to answer is:

1. What is the average time complexity required to prove that $\mathcal{H}_A$ is a unique minimal hitting set?

Proving uniqueness of $\mathcal{H}_A$ in $\mathcal{OS}$ is hard as there are exponentially many possible hitting sets $\mathcal{H} = \{r_1, \ldots, r_m\} \neq \mathcal{H}_A$ that need to be disproved with respect to $\mathcal{OS}$. To mitigate this problem we avoid disproving all individual sets $\mathcal{H}$ answering the following question:

2. How many peers in $\mathcal{H}$ must be chosen to prove that $\mathcal{H}$ is not a hitting set?

We choose peers $r_1, \ldots, r_x \in \mathcal{H}$ by determining all observations including $\mathcal{C} = \{r_1, \ldots, r_x\}$, which we denote $\mathcal{OS}[\mathcal{C}]$. Given these chosen peers we know the observations $\mathcal{OS} \setminus \mathcal{OS}[\mathcal{C}]$ that have not yet been considered. We refer to the remaining peers in $\mathcal{H}$ as *non-chosen*. Whilst a peer is non-chosen, we do not known which observations contain that peer.

Assume, without loss of generality, that after choosing these x peers in $\mathcal{C} \subseteq \mathcal{H}$ we know that $\mathcal{H}$ cannot be a hitting set, because the cumulative frequency of the $(m - x)$ most frequent peers in $\mathcal{OS} \setminus \mathcal{OS}[\mathcal{C}]$ is less than $|\mathcal{OS} \setminus \mathcal{OS}[\mathcal{C}]|$. In this case we prove not only that $\mathcal{H}$ is not a hitting set, but also that any superset $\mathcal{H}'$ of $\mathcal{C}$ cannot be a hitting set, where $|\mathcal{H}'| = m$.

In general, if we know that every set can be disproved after choosing on average x peers, then using (2) the average runtime complexity of ExactHS is approximated by $O(b^x mtb)$, which answers our first question. A more detailed justification and discussion of this complexity is provided in Sect. 5.

The rest of this section provides the theoretical model for answering the question of how many peers in $\mathcal{H}$ must be chosen to prove that $\mathcal{H}$ is not a hitting set. The answer will be derived in Sect. 4.

3.1 Potential

In this section we introduce the definition of the *potential*: our estimation of the number of distinct observations covered by a set $\mathcal{H}$ in a given observation set $\mathcal{OS}$. This value allows us to estimate the number of peer choices required to disprove a set, and thus to understand the complexity of ExactHS. Note that this "estimation" is part of our analysis of the complexity, and does not affect the exactness of the attack itself.

We assume without loss of generality that all considered sets are of the structure $\mathcal{H} = \{r_1, \ldots, r_x, r_{x+1}, \ldots, r_m\}$. Each r_i represents a distinct peer, and the number of peers is $|\mathcal{H}| = m$. The first $0 \leq x \leq m$ peers $r_1, \ldots, r_x$ are always chosen, while the remaining $(m - x)$ peers are non-chosen. The potential of $\mathcal{H}$ is denoted by $Po(\mathcal{H})$.

$$Po(\mathcal{H}) = |\mathcal{OS}[\{r_1, \ldots, r_x\}]| + |\mathcal{OS}[r_{x+1}] \setminus \mathcal{OS}[\{r_1, \ldots, r_x\}]| + \ldots + |\mathcal{OS}[r_m] \setminus \mathcal{OS}[\{r_1, \ldots, r_x\}]| \quad (4)$$

There are two extreme cases. If all peers are chosen, then the potential is the number of observations covered by $\mathcal{H}$. If all peers are non-chosen, then the potential is the cumulative frequency of the peers of $\mathcal{H}$ in $\mathcal{OS}$. The more peers chosen in $\mathcal{H}$, the more accurately the potential represents the number of distinct observations intersecting with $\mathcal{H}$. $Po(\mathcal{H})$ thus never underestimates the number of observations intersecting with $\mathcal{H}$.

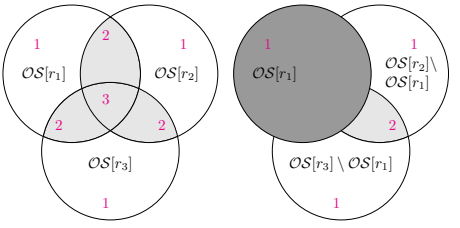

Fig. 2. *Left:* Overestimation by $Po(\{r_1, r_2, r_3\})$, where all peers r_1, r_2, r_3 are non-chosen. *Right:* Overestimation by $Po(\{r_1, r_2, r_3\})$, where r_1 is chosen.

Overestimations are observations that are covered by more than one non-chosen peers in $\mathcal{H}$ as illustrated by the leftmost diagram in Fig. 2. We will analyse the overestimation of the potential, since it enables us to conclude how many peers in $\mathcal{H} \neq \mathcal{H}_A$ need to be chosen to disprove it.

Potential: All Peers Non-Chosen. The set of observations covered by r_i is represented by a circle around $\mathcal{OS}[r_i]$ for $i = 1, 2, 3$ in the left-hand picture in Fig. 2. The grey area represents those observations that are covered by at least two peers r_i, r_j for $i \neq j$. The number in the area shows the number of times observations in that area are counted in the potential. In this example $\mathcal{H} = \{r_1, r_2, r_3\}$ and we can see on the left picture how $Po(\mathcal{H})$ overestimates $|\mathcal{OS}[\mathcal{H}]|$, which is the number of observations covered by $\mathcal{H}$. The overestimation is caused by those observations that are covered by more than one of the peers r_1, r_2, r_3. The exact number of observations covered by $\mathcal{H}$ in the left picture in Fig. 2 can be computed by the inclusion exclusion formula.

$$|\mathcal{OS}[\mathcal{H}]| = |\mathcal{OS}[r_1]| + |\mathcal{OS}[r_2]| + |\mathcal{OS}[r_3]| - |\mathcal{OS}[r_1] \cap \mathcal{OS}[r_2]| - \\ |\mathcal{OS}[r_1] \cap \mathcal{OS}[r_3]| - |\mathcal{OS}[r_2] \cap \mathcal{OS}[r_3]| + |\mathcal{OS}[r_1] \cap \mathcal{OS}[r_2] \cap \mathcal{OS}[r_3]|$$

As all peers in $\mathcal{H}$ are non-chosen, $Po(\mathcal{H}) = |\mathcal{OS}[r_1]| + |\mathcal{OS}[r_2]| + |\mathcal{OS}[r_3]|$. For the sake of simplicity we derive the following estimation from the equation above.

$$Po(\mathcal{H}) \leq |\mathcal{OS}[\mathcal{H}]| + |\mathcal{OS}[r_1] \cap \mathcal{OS}[r_2]| + |\mathcal{OS}[r_1] \cap \mathcal{OS}[r_3]| + |\mathcal{OS}[r_2] \cap \mathcal{OS}[r_3]|$$

Potential: General Case. The case when one peer r_1 is chosen while the other peers in $\mathcal{H}$ are non-chosen is illustrated by the right-hand picture of Fig. 2. By the definition of $Po(\{r_1, r_2, r_3\})$ in (4), choosing r_1 causes all observations containing it, represented by the dark circle, to be removed in the frequency consideration of the non-chosen peers. In this case $Po(\mathcal{H})$ overestimates $|\mathcal{OS}[\mathcal{H}]|$ by double-counting the grey area that represents observations that are covered by r_2 and r_3 but not by r_1. For simplicity we use the following estimation of $Po(\mathcal{H})$:

$$Po(\mathcal{H}) \leq |\mathcal{OS}[\mathcal{H}]| + |\mathcal{OS}[r_2] \cap \mathcal{OS}[r_3]| \ .$$

In general, if $0 \leq x \leq m$ peers $\{r_1, \ldots, r_x\}$ of $\mathcal{H} = \{r_1, \ldots, r_m\}$ are chosen, then the overestimation of the number of covered observations result from the non-chosen peers r_k, r_l for $x < k, l \leq m$. The overestimation is bounded by the size of the $\binom{m-x}{2}$ pairwise intersections $\mathcal{OS}[r_k] \cap \mathcal{OS}[r_l]$. This results in the following simplified estimation of the potential for the general case:

$$Po(\mathcal{H}) \leq |\mathcal{OS}[\mathcal{H}]| + \sum_{x<k,l\leq m;\, k\neq l} |\mathcal{OS}[r_k] \cap \mathcal{OS}[r_l]| \ . \tag{5}$$

Overestimation by Potential. In order to distinguish the effect of Alice's peers and non-peers to $Po(\mathcal{H})$, each peer $r \in \mathcal{H}$ is relabelled n for non-peers, and a for Alice's peer. Without loss of generality, every $\mathcal{H} \in \mathfrak{H}_j$, where $|\mathcal{H}| = m$ from now on has the following structure:

$$\mathcal{H} = \{\underbrace{n_1, \ldots, n_{x_1}, a_1, \ldots, a_{x_2}}_{x \text{ chosen peers}} \quad , \quad \underbrace{n_{x_1+1}, \ldots, n_j, a_{x_2+1}, \ldots, a_{m-j}}_{(m-x) \text{ non-chosen peers}}\} \ .$$

The number of chosen peers is $x = x_1 + x_2$, where $x_1 \leq j$ and $x_2 \leq m - j$. The variable j denotes the number of non-peers in hitting sets of the structure $\mathfrak{H}_j$. We still use the notation r_i to address the i-th peer in $\mathcal{H}$ if distinction is not important. As before, the first x peers $r_1, \ldots, r_x \in \mathcal{H}$ are chosen, while the remaining $(m-x)$ peers are non-chosen. We define $\mathcal{H}^{+A} = \mathcal{H} \cap \mathcal{H}_A$ as the subset containing only Alice's peers and $\mathcal{H}^{-A} = \mathcal{H} \setminus \mathcal{H}_A$ as the subset consisting of only non-peers.

The following estimations for $|\mathcal{OS}[\mathcal{H}]|$ and $|\mathcal{OS}|$ will be used next in inequality (9):

$$|\mathcal{OS}[\mathcal{H}]| \leq |\mathcal{OS}[\mathcal{H}^{+A}]| + \sum_{n\in\mathcal{H}^{-A}} |\mathcal{OS}[n] \setminus \mathcal{OS}[\mathcal{H}^{+A}]| \tag{6}$$

$$|\mathcal{OS}| \geq |\mathcal{OS}[\mathcal{H}^{+A}]| + \sum_{a\in(\mathcal{H}_A\setminus\mathcal{H}^{+A})} |\underbrace{\mathcal{OS}[a] \setminus \mathcal{OS}[\mathcal{H}_A \setminus \{a\}]}_{\text{observ. containing } a \text{ exclusively}}| \ . \tag{7}$$

An observation contains Alice's peer $a \in \mathcal{H}_A$ *exclusively* [9], if it does not contain any other peers of Alice.

We now mathematically formulate our earlier question; that is: how many peers must be chosen in order to prove that $\mathcal{H} \neq \mathcal{H}_A$ is not a hitting set in $\mathcal{OS}$? This is simple using the potential, as it estimates the number of observations covered by $\mathcal{H}$ in $\mathcal{OS}$. If $Po(\mathcal{H}) < |\mathcal{OS}|$ then $\mathcal{H}$ is clearly not a hitting set. On the other hand, if $Po(\mathcal{H}) \geq |\mathcal{OS}|$ then we must choose more peers in $\mathcal{H}$ for the disproof. The latter is formulated below. Inequality (9) then results from applying (5) and (6) on $Po(\mathcal{H})$ and (7) on $|\mathcal{OS}|$.

$$0 \leq Po(\mathcal{H}) - |\mathcal{OS}| \tag{8}$$

$$\begin{aligned} &\leq \sum_{x_2<k,l\leq m-j;\, k\neq l} |\mathcal{OS}[a_k] \cap \mathcal{OS}[a_l]| + \sum_{x_2<k\leq m-j;\, x_1<l\leq j} |\mathcal{OS}[a_k] \cap \mathcal{OS}[n_l]| \\ &\quad + \sum_{x_1<k,l\leq j;\, k\neq l} |\mathcal{OS}[n_k] \cap \mathcal{OS}[n_l]| + \sum_{n\in\mathcal{H}^{-A}} |\mathcal{OS}[n] \setminus \mathcal{OS}[\mathcal{H}^{+A}]| \\ &\quad - \sum_{a\in(\mathcal{H}_A\setminus\mathcal{H}^{+A})} |\mathcal{OS}[a] \setminus \mathcal{OS}[\mathcal{H}_A \setminus \{a\}]| \end{aligned} \tag{9}$$

For simplicity we restrict our analysis to those cases where the probability that a particular peer $r \in \mathcal{H}$ is contacted by a sender other than Alice, within a given observation $\mathcal{O}$, is significantly lower than the probability that Alice's peer is contacted by Alice. This allows us to ignore the possibility that some pair of peers $r_k, r_l \in \mathcal{H}$ is contacted by senders other than Alice in the same $\mathcal{O}$. This allows us to ignore counting the observations described below in (9):

$$\{\mathcal{O} \in \mathcal{OS}[r_k] \cap \mathcal{OS}[r_l] \mid r_k, r_l \in \mathcal{H} \text{ chosen by non-Alice senders in } \mathcal{O}\} \ . \quad (10)$$

We call the resulting simplified estimation of (9) the *difference* function $D(x, x_1, x_2, j)$:

$$\sum_{x_2<k,l\leq m-j;\, k\neq l} |\mathcal{OS}[a_k] \cap \mathcal{OS}[a_l]| + \sum_{x_2<k\leq m-j;\, x_1<l\leq j} |\mathcal{OS}[a_k] \cap \mathcal{OS}[n_l]| \ + \\ \sum_{n\in\mathcal{H}^{-A}} |\mathcal{OS}[n] \setminus \mathcal{OS}[\mathcal{H}^{+A}]| - \sum_{a\in(\mathcal{H}_A\setminus\mathcal{H}^{+A})} |\mathcal{OS}[a] \setminus \mathcal{OS}[\mathcal{H}_A \setminus \{a\}]| \ . \quad (11)$$

4 Number of Peer Choices for a Disproof

4.1 Expectation of the Difference

In this section we compute the expectation of the difference function for a simplified communication model of Alice and the other senders, which we call *uniform communication*.

In this model the *cumulative communication* of all other senders leads to a uniform *background distribution* of communication with the peers such that, without Alice's communication, each peer $r \in R$ appears with the same *cumulative probability* of P_{nA} in an observation. Therefore each sender can select its peer according to an arbitrary distribution provided that $\forall r \in R : P(r \in \mathcal{OS}) = P_{nA}$, where $P(r \in \mathcal{OS})$ denotes the probability that r appears in the observations $\mathcal{OS}$ of the attacker without considering Alice's communication.

To simplify our analysis we assume that, in every round, each of the $(b-1)$ non-Alice senders choose their peers uniformly from the set R of N recipients with probability $\frac{1}{N}$. Thus, for every peer $r \in R$ its cumulative probability of appearing in an observation is $P_{nA} = 1 - (\frac{N-1}{N})^{b-1}$. We further assume that Alice contacts one of her m peers $a \in \mathcal{H}_A$ in each round, chosen according to the uniform distribution with the probability of $P_A = \frac{1}{m}$.

$$E_1(x, x_1, x_2, j) = t\binom{m-j-x_2}{2}\frac{2}{m}P_{nA}$$

$$E_2(x, x_1, x_2, j) = t(j-x_1)(m-j-x_2)\frac{1}{m}P_{nA}$$

$$E_3(x, x_1, x_2, j) = tj\frac{j}{m}P_{nA}$$

$$E_4(x, x_1, x_2, j) = tjm^{-1}\left(1-(m-1)N^{-1}\right)^{b-1}$$

The difference described by equation (11) is generic and can be analysed with respect to arbitrary communication models. It is sufficient, however, to consider uniform communications, and Sect. 5.1 will show a mapping from non-uniform to uniform communications that provide analytical bounds valid for both instances. For

the sake of simplicity, all remaining analysis in this paper will refer to uniform communication unless otherwise stated.

The equations above represent the expectation of the four terms of equation (11), where the number of observations collected by the attacker is $t = |\mathcal{OS}|$.

The terms following t in E_1, E_2, E_3, E_4 are significant, and we discuss these here.

E_1: For Alice's peers $a_k, a_l \in \mathcal{H}^{+A}$, where $a_k \neq a_l$, the probability that Alice contacts a_k and one of the other $(b-1)$ senders contact a_l in an observation is $\frac{1}{m}P_{nA}$. Due to symmetry, the probability that a_k and a_l appear in an observation is $\frac{2}{m}P_{nA}$. This is multiplied by the number of possible pairs of non-chosen Alice's peers $\binom{m-j-x_2}{2}$.

E_2: For peers $a_k \in \mathcal{H}^{+A}$ and $n_l \in \mathcal{H}^{-A}$, the probability that Alice contacts a_k and one of the other $(b-1)$ senders contacts n_l is $\frac{1}{m}P_{nA}$. The factor $(m-j-x_2)$ shows the number of non-chosen Alice's peers a_k while the factor $(j-x_1)$ represents the number of non-chosen non-peers n_l.

E_3: Let $a_1, \ldots, a_j \in (\mathcal{H}_A \setminus \mathcal{H})$ be the j Alice's peers that are not in $\mathcal{H}$. The probability that a given non-peer $n_k \in \mathcal{H}^{-A}$ appears in an observation where Alice contacts one of $a_1, \ldots, a_j$ is $\frac{j}{m}P_{nA}$. The final factor j accounts for the fact that there are j non-peer n_k in $\mathcal{H}^{-A}$.

E_4: Alice's peer $a \in (\mathcal{H}_A \setminus \mathcal{H})$ is exclusive in an observation if Alice contacts a and none of the other $(b-1)$ senders contact any of the peers $a' \in (\mathcal{H}_A \setminus \{a\})$. The probability that a is exclusive is therefore $\frac{1}{m}\left(1-\frac{m-1}{N}\right)^{b-1}$. The factor j accounts for this exclusivity probability for the j Alice's peers $a_1, \ldots, a_j \in (\mathcal{H}_A \setminus \mathcal{H})$ not appearing in $\mathcal{H}$.

Combining these expectations results in an expectation, $E_D(x, x_1, x_2, j)$, for the difference function $D(x, x_1, x_2, j)$ of:

$$\frac{t}{m}\Big[((m-x-1)(m-j-x_2)+j^2)P_{nA} - j\left(1-\frac{m-1}{N}\right)^{b-1}\Big] \quad . \tag{12}$$

4.2 Average Number of Peer Choices

We obtain the average number of peer choices to disprove a set $\mathcal{H}$ by determining the value of x such that the expectation of the difference is 0. By detailed analysis of the property of E_D (in Appendix A) , we gain simple descriptions of assertions about the limits of the number of peer choices. These limits are summarised here.

Upper Bound of Average Worst Case Number of Peer Choices. If $\frac{N}{b-1} \geq 3m-1$ and N, b, m is fixed, then the *upper bound of the average worst case number of peer choices* is x_{uw}. This value provides an estimate of the average maximal number of peer choices for a disproof, approaching the bound from above. This can be reformulated to determine the parameters N, b, m, such that a particular bound x_{uw} is obtained by (14).

$$x_{uw} = m - \frac{1}{2} - \sqrt{\frac{N}{b-1} - m + \frac{1}{4}} \quad \text{, where } x_{uw} \leq m \tag{13}$$

$$b = \frac{N}{m^2 - 2mx_{uw} + x_{uw}^2 + x_{uw}} + 1 \tag{14}$$

For full proofs of these results, see Appendix A.1.

5 Runtime Complexity

We have now determined how many peers must be chosen in order to disprove a hypothesis set, and so can answer our original question: what is the average complexity to identify unambiguously Alice's peer set $\mathcal{H}_A$?

The ExactHS algorithm reduces the space of sets that must be disproved to identify $\mathcal{H}_A$ by two strategies. Firstly, ExactHS reduces the search space to consider only minimal hitting sets, which is sufficient to identify $\mathcal{H}_A$ in [12, 13]. Secondly, it deploys the estimation of the number of covered observations based on the potential and implements the difference function (Alg. 1 Lines 6, 8). In Alg. 1 the set $\mathcal{C}$ represents $(m-m')$ chosen peers and $\{r_1, \ldots, r_{m'}\}$ represents hypothetical non-chosen peers. The algorithm constructs $|\mathcal{OS}'| = |\mathcal{OS} \setminus \mathcal{OS}[\mathcal{C}]| = |\mathcal{OS}| - |\mathcal{OS}[\mathcal{C}]|$ and $\sum_{l=1}^{m'} |\mathcal{OS}'[r_l]| = Po(\mathcal{C} \cup \{r_1, \ldots, r_{m'}\}) - |\mathcal{OS}[\mathcal{C}]|$, where $\mathcal{OS}$ is the initial set of observations of the attacker, which is equivalent to Equation (8). This allows direct application of the bounds derived in the last section to ExactHS.

The worst case number of peer choices, x, to disprove a set in the last section therefore corresponds to the worst case number of recursion levels x invoked in ExactHS.

To avoid significantly overestimating the strength of the system, we assume that the variance of the average number of peer choices x is negligible. (1) therefore results in an average number of finalised sets computed by ExactHS to identify $\mathcal{H}_A$ of: b^x .

To obtain the corresponding runtime complexity, the last term must be multiplied by tbm, resulting in $O(b^x tbm)$, and reaches a worst case complexity of $O(b^m tbm)$ when $x = m$. The following analysis consequently refers only to the number of finalised sets computed by ExactHS.

5.1 Upper Bound of Average Worst Case

The upper bound of the average worst-case complexity results from the upper bound of the average worst-case number of peer choices x_{uw} determined by (13). Applying that to b^x we derive the upper bound for the average maximal number of finalised sets computed by ExactHS for the unambiguous identification of $\mathcal{H}_A$:

$$b^{m-\frac{1}{2}-\sqrt{\frac{N}{b-1}-m+\frac{1}{4}}} \approx b^{m-\frac{1}{2}-\sqrt{\frac{1}{P_{nA}}-m+\frac{1}{4}}} \quad . \tag{15}$$

From the relations $P_{nA} = 1 - (1 - \frac{1}{N})^{b-1} \approx \frac{b-1}{N}$ and $P_A = \frac{1}{m}$ we conclude that:

- If every peer not contacted by Alice is at least as likely to appear in an observation as peers contacted by Alice, the average worst case complexity roughly equals the worst case complexity $O(b^m tbm)$. That is if $P_{nA} = \frac{1}{m-\frac{1}{4}}$.
- The average worst case complexity becomes linear $O(tbm)$ if every peer not contacted by Alice appears in observations with a probability close to $\frac{1}{m^2}$.

Non-uniform Communication. The analyses above apply to non-uniform background distribution by setting $P_{nA} = \max_{r \in R'}\{P(r \in \mathcal{OS})\}$ in (15). This maps an instance

with non-uniform background communication and parameters $N' = |R'|, b, m$ to an instance of uniform communication with parameters $N = |R| = \frac{b-1}{P_{nA}}, b, m$, where R' and R is the recipient set of the first and second instances respectively. The average case complexity of the latter is at least as high as the former, as in uniform communication each of the $N = |R|$ peers appears with a probability of P_{nA} in an observation, while a smaller number of most likely peers of R' appears with that probability in non uniform communication.

Note that the cumulative background probability of the peers can be estimated in the global passive attacker model by considering observations in which Alice does not participate, enabling attackers, Mix providers and users to determine *a priori* the average worst case complexity of ExactHS for a distinct number of Alice's peer partners m.

We assume Alice's communication to be uniform when deriving the average case complexity not only for simplicity, but also because simulation reveals that it is the worst case for the average run time complexity. Informally, in a non-uniform communication some Alice's peers are even more statistic signification than the non-peers. Thus, making ExactHS focus on the most frequent peers reduces the hypothesis space and average time complexity. A formal proof of this is forthcoming.

Relation to Least Number of Observations by ExactHS. To determine efficiently the number of observations required by the ExactHS-attack, we can apply the algorithm to compute the lower bound of the HS-attack based on the *2x-exclusivity criteria* [9, 12] or use the mathematical analysis provided by [13].

We use here the formula for the least number of observations t to identify $\mathcal{H}_A$ by the minimal hitting set attack [13]. It provides, in contrast to the 2x-exclusivity formula in [9], a closed formula that directly represents the effect of Mix parameters.

$$t \approx m\big(\ln(b-1) - \ln(2^{1/m}-1)\big)\big(1 - mN^{-1}\big)^{1-b} \tag{16}$$

This formula shows that ExactHS can reveal Alice's peer set after a number of observations t that is sub-exponential with respect to N, b, m. The number of observations for the identification of Alice's peers is thus an insufficient metric for the strength of the Mix, and we need to consider the average case complexity of ExactHS. Section 6 compares the theoretical results of this paper with attacks on simulated data.

Countermeasure against Attack. To prevent the ExactHS attack in practice, Mix providers can adjust the average case complexity $O(b^x tbm)$ to be close to the worst case complexity, such that $x = m - \epsilon$ for fixed security parameters m and ϵ chosen by the provider. To obtain this, the batch size b can be determined with respect to N, m, x according to equation (14). By doing so, applying our attack against users who uniformly contact $m' \geq m$ peer partners requires a time complexity bounded by $O(b^{m'-\epsilon}tbm')$. Users with $m' < m$ peer partners, however, or non-uniform communication should be aware that revealing their peer partners will be faster than $O(b^{m'-\epsilon}tbm')$.

We have derived a formula for the lower bound of the average case complexity of ExactHS, which could be used to adjust the least average time required for an attack on a Mix, however we omit this due to space limitations.

Note that ExactHS and statistical attacks are based on very different principles. Therefore, Mix configurations that are susceptible to ExactHS are not necessarily susceptible to the statistical attacks and vice versa. While it is outside of the scope of this paper, a comparison of the effectiveness of both classes of attacks with respect to different Mix configurations and countermeasures would be an interesting topic for future research.

6 Simulation

To support our mathematical analysis, we now show the ExactHS algorithm applied to randomly generated observations. These observations are generated under the uniform communication model of Sect. 4.1, which is chosen to allow direct comparison between the simulation and our theoretical results.

An attack is *successful* if ExactHS can unambiguously identify Alice's peer set $\mathcal{H}_A$; the simulation generates new observations until this occurs. The average number of observations required by an attack is therefore the mean of the number of observations of all successful attacks. To ensure that our results are statistically significant, experiments were repeated until 95% of the results fall within 5% of the empirically observed mean.

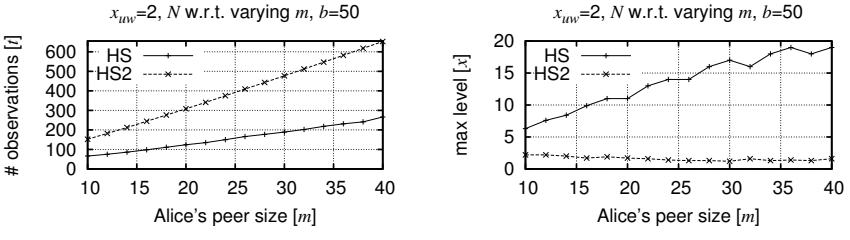

Fig. 3. Parameters N, b, m, where $x_{uw} = 2$. *Left:* Number of observations when ExactHS succeeds. *Right:* Empirical recursion level for disproof by ExactHS.

Average Worst Case. To demonstrate that our analysis closely predicts the empirical average worst case complexity of ExactHS, we apply attacks on observations of a Mix with parameters N, b, m that are chosen according to (14), where $x_{uw} = 2$. It is therefore expected that ExactHS succeeds on those configurations within a polynomial run time of $O(b^2tbm)$, while its average worst case recursion level is bounded by 2.

Figure 3 shows the result of our simulation for fixed $b = 50$. The value of N is determined by (14) given fixed $x_{uw} = 2$; m values are shown on the x-axis. The value of N ranges from 3200 for $m = 10$ to 70000 for $m = 40$.

As the attack requires very few observations to succeed, the empirical probability distribution of the peers of the non-Alice senders at the termination of the attack strongly diverge from the function $P_{nA} \approx \frac{b-1}{N}$ from which they are drawn[6].

[6] Assume for example that $P_{nA} = 1/400$, but the attack succeeds after $|\mathcal{OS}| = 100$ observations, then the probability of each peer included by an observation in $\mathcal{OS}$ exceeds P_{nA} by at least a factor of four.

Due to the law of large numbers, this side effect diminishes for large number of observations. We therefore consider the application of ExactHS where the number of observations is twice that required by (16). This is shown in the graphs by the line labelled (HS2). This doubling is simply to aid demonstration of our results by reducing the side effects due to the small number of observations.

The left plot shows on the y-axis the average number of observations to identify Alice's peer set unambiguously. The line (HS) represents the mean of the least possible number of observations required by ExactHS in an information theoretic sense. The line (HS2) shows the number of observations which corresponds to twice the value of (16).

The right plot shows on the y-axis the average worst case level required to disprove a set by ExactHS under the conditions represented by the lines (HS) and (HS2). The line (HS) shows that the level is significantly higher than x_{uw} if ExactHS identifies $\mathcal{H}_A$ with the information theoretic minimal number of observations. This is due to the probabilities of many non-peers exceeding P_{nA} due to a low number of observations. With more observations, as in (HS2), we can see that the average worst case number of required peer choices is about x_{uw} for all selected N, b, m as predicted by (14). Collecting even more additional observations when applying ExactHS does not noticeably change the worst case number of peer choices.

7 Conclusion

Previous non-statistical analyses of Mixes have been based almost exclusively on the least number of observations for an attack, and on the fact that the unambiguous identification of Alice's peer set requires the solution of an NP-complete problem.

This paper is the first presentation, to our knowledge, of a detailed complexity-theoretic analysis of the problem of identifying a user's peer set beyond the worst case complexity determined by the NP-completeness of the underlying problem. We achieve this by contributing closed formulas that determine the average case complexity with respect to the Mix parameters. These theoretical results are further supported by simulations.

It is clear from our results that the identification of Alice's peers in a Mix network, whilst being intractable in the worst case, contains a broad range of realistic Mix configurations that are polynomially solvable. These configurations are serious threats for anonymity that can now be identified by our results (13), (15). Our analyses enable further to identify those configurations that are solvable only in exponential time by ExactHS, allowing for an increase in the anonymity of these systems.

In order to gain the average case complexity of the system, we employ the most efficient known algorithm that provides an exact result. Whilst the possibility exists that a more efficient algorithm could be discovered[7], our results are the first to provide an analysis of this form.

In the future, we intend to extend the analysis in this work to more complex and real-world Mix models. It is hoped that this will allow us to understand the effect that different mixing strategies have on anonymity. In a wider context, our analyses are concerned with the identification of average polynomial-time-solvable instances of an

[7] As is possible with, for example, the prime factorisation algorithms employed in cryptanalysis.

NP-complete problem. The results presented here may therefore be of use in identifying average polynomial-time instances of other interesting NP-complete problems, which would have wider applications beyond the restricted scope of security and privacy.

References

[1] Agrawal, D., Kesdogan, D., Penz, S.: Probabilistic Treatment of MIXes to Hamper Traffic Analysis. In: IEEE Symposium on Security and Privacy, pp. 16–27 (2003)
[2] Berthold, O., Langos, H.: Dummy traffic against long term intersection attacks. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 110–128. Springer, Heidelberg (2003)
[3] Chaum, D.L.: Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. Communications of the ACM 24(2), 84–88 (1981)
[4] Danezis, G.: Statistical Disclosure Attacks: Traffic Confirmation in Open Environments. In: Proceedings of Security and Privacy in the Age of Uncertainty, pp. 421–426 (2003)
[5] Danezis, G., Diaz, C., Troncoso, C.: Two-sided statistical disclosure attack. In: Borisov, N., Golle, P. (eds.) PET 2007. LNCS, vol. 4776, pp. 30–44. Springer, Heidelberg (2007)
[6] Danezis, G., Serjantov, A.: Statistical Disclosure or Intersection Attacks on Anonymity Systems. In: Fridrich, J. (ed.) IH 2004. LNCS, vol. 3200, pp. 293–308. Springer, Heidelberg (2004)
[7] Danezis, G., Troncoso, C.: Vida: How to use bayesian inference to de-anonymize persistent communications. In: Goldberg, I., Atallah, M.J. (eds.) PETS 2009. LNCS, vol. 5672, pp. 56–72. Springer, Heidelberg (2009)
[8] Garey, M.R., Johnson, D.S.: Computers and Intractability; A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York (1990)
[9] Kesdogan, D., Agrawal, D., Pham, V., Rauterbach, D.: Fundamental Limits on the Anonymity Provided by the Mix Technique. In: IEEE Symposium on Security and Privacy (2006)
[10] Kesdogan, D., Pimenidis, L.: The Hitting Set Attack on Anonymity Protocols. In: Fridrich, J. (ed.) IH 2004. LNCS, vol. 3200, pp. 326–339. Springer, Heidelberg (2004)
[11] Mathewson, N., Dingledine, R.: Practical Traffic Analysis: Extending and Resisting Statistical Disclosure. In: Martin, D., Serjantov, A. (eds.) PET 2004. LNCS, vol. 3424, pp. 17–34. Springer, Heidelberg (2005)
[12] Pham, V.: Analysis of the Anonymity Set of Chaumian Mixes. In: 13th Nordic Workshop on Secure IT-Systems (2008)
[13] Pham, D.V., Kesdogan, D.: A Combinatorial Approach for an Anonymity Metric. In: Boyd, C., González Nieto, J. (eds.) ACISP 2009. LNCS, vol. 5594, pp. 26–43. Springer, Heidelberg (2009)
[14] Serjantov, A., Danezis, G.: Towards an Information Theoretic Metric for Anonymity. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 259–263. Springer, Heidelberg (2003)
[15] Troncoso, C., Danezis, G.: The bayesian traffic analysis of mix networks. In: ACM Conference on Computer and Communications Security, CCS 2009, pp. 369–379 (2009)
[16] Troncoso, C., Gierlichs, B., Preneel, B., Verbauwhede, I.: Perfect matching disclosure attacks. In: Borisov, N., Goldberg, I. (eds.) PETS 2008. LNCS, vol. 5134, pp. 2–23. Springer, Heidelberg (2008)

A Analysis of Expectation Function for Number of Peer Choices

Relation to Number of Chosen Peers

Claim 1. *The expectation $E_D(x, x_1, x_2, j)$ is a monotonically decreasing function with respect to the number of chosen peers x, where $1 \leq x \leq m - \frac{1}{2}$.*

The proof consists of two parts. We will show that $E_D(x, x_1, x_2, j)$ is monotonically decreasing given that x_1 is fixed and then for the case that x_2 is fixed.

Proof (Monotonicity of $E_D(x, x_1, x_2, j)$ given fixed x_1). This analysis refers to the case that the number of chosen non-peers x_1 is fixed in the chosen peers x. By definition $x_2 = (x - x_1)$, therefore we replace all x_2 in (12) by $(x - x_1)$. The following function determines the gradient of the resulting function by computing its partial derivative with respect to x: $\frac{\partial E_D(x,x_1,x-x_1,j)}{\partial x} = \frac{tP_{nA}}{m}(2x - 2m - x_1 + j + 1)$.

This equation is less-than or equal 0, if:

$$x \leq m + 0.5(x_1 - j) - 0.5 \ . \tag{17}$$

We consider the inequality (17) for different cases of $(x_1 - j)$. By definition $x_1 \leq j$, therefore only the following cases exist:

$x_1 = j$: In this case E_D is a decreasing function if $x \leq m - \frac{1}{2}$.

$x_1 < j$: In this case E_D is always a decreasing function. The proof derives from the definition $x = (x_1 + x_2)$, where $x_2 \leq (m - j)$. Replacing x_2 in the first equation by the latter inequality, we obtain:

$$x \leq m + (x_1 - j) \quad \Rightarrow \quad x \leq m + 0.5(x_1 - j) - 0.5, \text{ since } x_1 - j \leq -1 \ .$$

Therefore (17) is always fulfilled in this case.

This proves that $E_D(x, x_1, x_2, j)$ is a monotonically decreasing function with respect to the number of chosen peers x, where $1 \leq x \leq m - \frac{1}{2}$, given that x_1 is fixed. □

Proof (Monotonicity of $E_D(x, x_1, x_2, j)$ given fixed x_2). We now consider the case that the number of Alice's peers is fixed in the number of chosen peers x. The gradient of $E_D(x, x_1, x_2, j)$ with respect to x is now: $\frac{\partial E_D(x,x-x_2,x_2,j)}{\partial x} = \frac{tP_{nA}}{m}(-m + x_2 + j)$.

The relation $(x_2 + j) \leq m$ is given by definition, therefore the gradient is always less-than or equal to 0. This proves that $E_D(x, x_1, x_2, j)$ is a monotonically decreasing function, given that x_2 is fixed. □

We conclude from these two proofs that $E_D(x, x_1, x_2, j)$ is a monotonically decreasing function with respect to the number of chosen peers x, where $1 \leq x \leq m - \frac{1}{2}$. This completes the proof of Claim 1. All analyses in the rest of the paper implicitly assume $x \in [1, \ldots, m - 1]$.

Relation to Order of Peer Choice. This section will show that, in general, if one prefers to chose non-peers in $\mathcal{H} \in \mathfrak{H}_j$ first and then the remaining peers of Alice, then the number of choices required to disprove $\mathcal{H}$ is maximised.

Claim 2. *Let x be a fixed number of chosen peers and x_1 be the number of chosen non-peers, where $x_1 \leq j \leq x$. The expectation $E_D(x, x_1, x_2, j)$ with respect to x_1 is a monotonically increasing function.*

Proof. To analyse how E_D is related to the number of non-peer choices x_1, we compute the partial derivative of $E_D(x, x_1, x - x_1, j)$ with respect to $x_1 \leq j \leq x$, where x is fixed. This is: $\frac{\partial E_D(x,x_1,x-x_1,j)}{\partial x_1} = \frac{tP_{nA}}{m}(m - x - 1)$.

This equation is clearly greater than 0 (since $x \leq m - 1$ is assumed), therefore $E_D(x, x_1, x - x_1, j)$ is a monotonically increasing function for x_1 in the complete interval $[0, \ldots, j]$. □

Note that $E_D(x, x_1, x - x_1, j)$ for $x_1 > j$ is, by definition of x_1, not defined. Given that $\mathcal{H}$ has $x \geq j$ chosen peers, $Po(\mathcal{H})$ is maximal if $x_1 = j$ of the chosen peers are non-peers. Disproving $\mathcal{H}$ therefore requires the maximal number of chosen peers if the non-peers are chosen first. To simplify the notation, and because of the importance of the number of non-peers, we will replace the notation $E_D(x, x_1, x - x_1, j)$ by the shorter notation $E_D(x, x_1, j)$ in the sequel.

A.1 Average Worst Case Number of Peer Choices

In this section we assume a worst case algorithm that chooses the peers of a set $\mathcal{H} \in \mathfrak{H}_j$ such that the number of peer choices x to disprove $\mathcal{H} \neq \mathcal{H}_A$ is maximal. According to the previous section this is the case if the non-peers are always chosen first in $\mathcal{H}$.

Claim 3. *Let $\frac{N}{b-1} \geq 2(m - 1)$. The maximal number of peer choices x, such that $E_D(x, x_1, j) = 0$ with respect to N, b, m, j, is:*

$$x_w = m - 0.5 - \sqrt{jN(b-1)^{-1} - j^2 + j - mj + 0.25} \ . \tag{18}$$

We call x_w the *average worst case number of peer choices*.

Proof. In order to ensure that all non-peers are chosen first, we set $x_1 = j$. Given this, the maximal number of peer choices is the value x, such that $E_D(x, x_1, j)$ in (12) is 0.

$$\begin{aligned} 0 &= E_D(x, j, j) \\ &\leq \frac{t}{m}\Big[((m - x - 1)(m - x) + j^2)(1 - (1 - \frac{b-1}{N})) - j(1 - (b-1)\frac{m-1}{N})\Big] \ . \end{aligned}$$

We obtain (18) by computing the positive root of the last right hand side function for the variable x. Equation (18) is valid if the term within the square root is at least 0. That is, if:

$$0 \leq jN(b-1)^{-1} - j^2 + j - mj + 0.25 \ .$$

Since $j \leq m$ the above equation holds if: $N(b-1)^{-1} \geq 2(m-1)$.

Note that it is sufficient to assume $x_1 = j$ and $x \geq j$ for the proof. There is no need to consider the case $x < j$ for the average worst case number of peer choices, where $x_1 < j$ separately.

For an intuitive explanation, we assume a set $\mathcal{H} \in \mathfrak{H}_j$ for a maximal value j, such that $x_1 = j = x$ is the maximal number of non-peer choices to disprove $\mathcal{H}$. Let $\mathcal{H}' \in \mathfrak{H}_{j'}$ be another set, where $j' > j$. Since we assume that each Alice's peer are more frequently observed by the attacker than any non-peer, the relation $Po(\mathcal{H}') < Po(\mathcal{H})$ holds in most of the cases. We can particularly follow that $E_D(x, x_1, j') < E_D(x, x_1, j)$ implying that the maximal number of peer choices to disprove $\mathcal{H}$, as well as $\mathcal{H}'$ is x. Analysing the case $x_1 = j$ and $x \geq j$ is thus sufficient. A formal proof of this follows from a generalised form of (19), but is omitted here for brevity. □

Gradient of Worst Case Function x_w. We now analyse the case where (18) is a monotonically decreasing function with respect to $\mathfrak{H}_j$ to simplify succeeding analyses. The next equation is the partial derivative of (18) with respect to j.

$$\frac{\partial x_w}{\partial j} = -\frac{1}{2} \frac{\left(N(b-1)^{-1} - 2j + 1 - m\right)}{\left(jN(b-1)^{-1} - j^2 + j - mj + 0.25\right)^{\frac{1}{2}}} \tag{19}$$

x_w is thus monotonically decreasing if the numerator in the above is at least 0.

$$0 \leq N(b-1)^{-1} - 2j + 1 - m \qquad \Rightarrow \qquad j \leq 0.5\left(N(b-1)^{-1} - m + 1\right)$$

Thus, if the maximal number of non-peer choices in a disproof is not larger than $\frac{1}{2}(\frac{N}{b-1} - m + 1)$, (18) is a monotonically decreasing function. If $\frac{N}{b-1} \geq 3m - 1$, then this case is necessarily fulfilled and we assume this condition for the remaining analyses.

Upper Bound of Average Number of Peer Choices. This section determines the upper bound of the average worst case number of peer choices x_w.

Claim 4. *Let $\frac{N}{b-1} \geq 3m-1$ and x_w be the average worst case number of peer choices. The maximal value of x_w for fixed N, b, m is:*

$$x_{uw} = m - \frac{1}{2} - \sqrt{\frac{N}{b-1} - m + \frac{1}{4}} \quad , \textit{ where } 0 \leq x_{uw} \leq m \ . \tag{13}$$

We call x_{uw} the *upper bound of the average worst case number of peer choices*.

Proof. Let $\frac{N}{b-1} \geq 3m - 1$, then x_w is monotonic decreasing with respect to j. It is therefore maximal if we set $j = 1$ in (18) and thus obtain (13). □

In case of $\frac{N}{b-1} < 3m - 1$, the right hand side of equation (13) might not provide a maximal value for x_{uw}. Therefore we can conclude in this case that if $\frac{N}{b-1} = m - \frac{1}{4}$, then $x_{uw} \geq m - \frac{1}{2}$ and that x_{uw} increases if the value of $\frac{N}{b-1}$ decreases. This justifies the conclusions of Sect. 5.1.

From this analysis we can obtain an approximation of the "lower bound of the average case complexity" of ExactHS. The derivation of these are omitted due to space limitation.