Information and Computation 244 (2015) 76–112

Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Analyzing anonymity attacks through noisy channels

Sami Zhioua [1]

Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Al-Dhahran, 31261, Saudi Arabia

ARTICLE INFO

Article history:
Received 13 August 2013
Received in revised form 21 May 2015
Available online 20 August 2015

Keywords:
Anonymity protocols
Information theory
Information leakage
Security analysis

ABSTRACT

Anonymity protocols focus on protecting the identities of senders and/or receivers in a network communication. Most of these protocols rely on randomness to achieve their goal and therefore can very well be represented as noisy channels in the information theoretic sense. In this paper we examine the problem of measuring the anonymity degree of anonymity protocols. We investigate a new idea of measuring anonymity based on how much the rows of the channel probabilities matrix are different from each other. We propose a new and generic approximation algorithm for the open problem of finding where anonymity of a given protocol is minimized. We illustrate how the probabilities matrix is constructed for some known anonymity protocols and we use the information leakage measures to study known attacks on those protocols. The analysis shows counter intuitive results in particular for Timed dynamic pool (Cottrell) mixes. Finally, we discuss the applicability of the proposed measures on Tor Network.

1. Introduction

The most common tool for providing network security is cryptography, an old technique that has been revived and adapted to network security. Cryptography is the science and art of transforming messages to make them secure and immune to attacks. Although encrypting a message can help protect its content from being revealed to an undesired observer, the identities of the sender as well as the receiver remain generally known. Various are the situations where it is essential that the sender and/or the receiver remain anonymous. These situations can be roughly divided into four main categories: discussion of sensitive and personal issues, information searches, freedom of speech in intolerant environments, and polling/surveying. Hence, cryptography alone is not enough to guarantee anonymity. Some anonymity techniques should be used in order to confuse an observer and conceal the communication relationship between the sender and the receiver. Known anonymous communication systems include Mix-based systems [1,2], DC-Net [3], Crowds [4], Anonymizing proxies such as Ano-nymizer [5] and SafeWeb. Currently the most popular anonymity solution is Tor [6,7] which is a low-latency anonymizing network.

Anonymity attacks can be passive or active. Typically, a passive attacker tries to guess the identities of communicating users based on his observation of the network traffic. Most of the time he will end up with a probability distribution on the users. For example, after observing the execution of a protocol and given a particular message m in the network, the attacker suspects user x to be the sender of m with probability $p_1 = 0.4$, user y with probability $p_2 = 0.25$, etc. The degree of anonymity of the protocol is tightly related to the uncertainty of that probability distribution called the a posteriori

E-mail address: zhioua@kfupm.edu.sa.
[1] Fax: +966 3 8602174.

http://dx.doi.org/10.1016/j.ic.2015.08.003
0890-5401/

distribution. This explains why information-theoretic metrics have been widely adopted to quantify the anonymity degrees of protocols [8,9].

We adopt an information theoretic approach to represent a protocol where the concept of observation is central. A protocol is considered as a noisy channel [10] which represents the link between a set of anonymous events $\mathcal{A}$ and a set of observable events $\mathcal{O}$. Events in $\mathcal{A}$ represent the information to hide from a potential attacker while events in $\mathcal{O}$ are the ones that the attacker actually observes. It is well known that a noisy channel in information theory can be represented by a matrix of conditional probabilities $p(o|a)$ for $o \in \mathcal{O}$ and $a \in \mathcal{A}$. For example, if the set of anonymous events is the identities of the possible senders of a message, $p(o|a)$ is the probability to observe observation o given that the sender of the message is a. This representation allows a better understanding of anonymity protocols and attackers capabilities.

A good anonymity protocol should make it hard to the attacker to guess the anonymous event given the observable event. Let A and O be two probability distributions on $\mathcal{A}$ and $\mathcal{O}$ respectively. The extreme case is when the distributions A and O are completely independent. This is called *noninterference* and achieving it, unfortunately, is often not possible because in most of the cases the protocol needs to reveal information about A. For example, in an election protocol, the individual votes should be secret but ultimately, the result of the votes must be made public which reveals information about individual votes. Hence the degree of anonymity of a protocol is tightly related to the amount of information leaked about the anonymous event when an observation is observed. In particular, more information leakage means less anonymity to the users of the system and vice versa.

In information theory, the information leaked by a noisy channel is given by the notion of mutual information which is the difference between the a priori and the a posteriori distributions uncertainties (entropies): $H(A) - H(A|O)$. Mutual information (MI[2]) has been used to quantify the information leakage in mix-based systems [11]. Smith [12] showed through an interesting example that when an adversary tries to guess the value of the anonymous event in a single try, an information leakage measure based on Renyi min-entropy [13] is more suitable than MI. Both MI and Smith's measure depend on the knowledge of the a priori distribution (the initial probabilities that a user did some action) while in general this distribution is not known. Capacity [10] which is an abstraction of MI obtained by maximizing over the possible a priori distributions is proposed as an alternative [14,15]. After all, we are concerned by the minimum degree of anonymity the protocol provides.

The contributions of this paper are:

- We investigate a new and original idea to quantify the information leakage in a noisy channel based on how much the rows of the matrix are different from each others and we adopt a geometric approach to assess how much the corresponding points in the n-dimensional space are scattered. These measures can be seen as alternatives to the classical notion of MI and illustrate different flavors of behaviors which offer different views of the actual anonymity of protocols.
- We prove several relationships between the proposed measures and existing notions. In particular, we show that one variant of these measures coincides with MI which gives it an interesting geometric interpretation.
- We propose a generic reinforcement learning algorithm to compute the minimum anonymity of a protocol according to any arbitrary metric. The algorithm approximates the a priori distribution where the metric is optimal.
- We illustrate how a set of anonymity protocols can be represented using noisy channels and we use the information leakage measures to study known attacks on those protocols. The analysis shows counter intuitive results in particular for Timed dynamic pool mixes (Cottrell mixes).

The next section is a brief account of related work. Section 3 gives a background on anonymity degrees and illustrates how a protocol can be represented as a noisy channel. Section 4 presents existing information leak measures namely MI and Smith's measure. Section 5 exposes our ideas behind the new information leak measures and illustrates some interesting relationships with existing notions. The next section analyzes and compares the new measures. Section 7 details our approximation algorithm to compute the maximum information leakage based on reinforcement learning. Sections 8 and 9 show how the measures can be used to analyze known anonymity systems namely, crowds, onion-routing, DC-Net, and Cottrell mix. Section 10 considers the applicability of the proposed measures to assess the anonymity of Tor network. The last section discusses the contributions of the paper and concludes.

2. Related work

The first anonymity measure, called anonymity set, was introduced by Chaum [3] and is simply the set of users who are likely to be the sender or receiver of a particular message. Naturally, the anonymity of the users increases if the size of the anonymity set increases. Serjantov and Danezis [9] defined the effective set size based on the concept of entropy after they showed that the simple anonymity set is inadequate when not all the users are equally likely to have sent a particular message. Diaz et al. [8] proposed independently a similar measure and took the next step in attempting to normalize the entropy and thus define a degree of anonymity as a number between 0 and 1. These two simple entropy measures were

[2] Table D.2 in Appendix D lists all the notations used in this paper along with the associated formulas.