UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Ramon Gonçalves Gonze

A Quantitative Information Flow Model for Attribute-Inference Attacks and Utility in Data Releases by Sampling

Belo Horizonte
2023

Ramon Gonçalves Gonze

A Quantitative Information Flow Model for Attribute-Inference Attacks and Utility in Data Releases by Sampling

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Mário Sérgio Alvim

Belo Horizonte
2023

Gonze, Ramon Gonçalves

G643q A quantitative information flow model for attribute-inference attacks and utility in data releases by sampling [manuscrito] / Ramon Gonçalves Gonze — 2023.
107 f. il.; 29 cm.

Orientador: Mario Sérgio Ferreira Alvim Júnior.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Departamento de Ciência da Computação, Instituto de Ciências Exatas
Referências: f.80-82.

1. Computação – Teses. 2. Dados estatísticos – Divulgação – Teses. 3. Amostragem (Estatística) – Teses. 4. Privacidade – Informática – Teses. 5. Gestão da informação – Teses. I. Alvim Júnior, Mario Sérgio Ferreira. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*74 (043)

Ficha Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A Quantitative Information Flow Model for Attribute-Inference Attacks and Utility in Data Releases by Sampling

RAMON GONÇALVES GONZE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARIO SÉRGIO FERREIRA ALVIM JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. JEROEN ANTONIUS MARIA VAN DE GRAAF
Departamento de Ciência da Computação - UFMG

PROFA. CATUSCIA PALAMIDESSI
- École Polytechnique - INRIA-LIX

Belo Horizonte, 11 de janeiro de 2023.

Dedico este trabalho aos meus pais, Helbert e Viviane, e ao meu irmão Breno.

Acknowledgments

I would like to thank my parents, brother, friends, relatives and everyone who encouraged me to keep doing the academical work I have been done.

I also would like to thank my advisor Professor Mário Sérgio Alvim for all the support given and the opportunities offered to me.

I am very grateful for all the help I received from many people, particularly for the technical discussions with my friends from UFMG, colleagues from my laboratory T-REX, Mireya Jurado and Catuscia Palamidessi.

Finally, I am grateful for UFMG, CAPES and the Brazilian people for financing this work.

"*Life doesn't stop because someone turned off the lights.*"
(Wayne Jolin)

Resumo

Divulgação de dados estatísticos é um processo presente na sociedade há bastante tempo, entretanto, a preocupação com privacidade é relativamente recente. O interesse em proteger dados individuais aumentou consideravelmente depois da elaboração de regulações sobre proteção de dados ao redor do mundo, como a *General Data Protection Regulation* (GDPR) na União Européia e a Lei Geral de Proteção de Dados (LGPD) no Brasil.

O esforço na comunidade científica para criar métodos de mitigação de risco à privacidade e para entender o compromisso entre privacidade e utilidade compõe uma grande área de pesquisa. Contudo, modelos matemáticos que buscam explicar formalmente este compromisso são, em alguma situações, incompreendidos pelos curadores de dados, i.e., entidades que coletam dados de uma população e adotam uma certa política para publicá-los podem não compreender quais os riscos e benefícios de tal política. Neste sentido, modelos e soluções que garantem que todas as partes envolvidas tenham ciência dos riscos e benefícios de cada política adotada se mostram importantes para que tomadas de decisões sejam realizadas de modo bem informado.

Como primeira contribuição deste trabalho, nós propomos um modelo que captura a vulnerabilidade de publicar-se uma amostra de uma população, em particular, a vulnerabilidade sob um ataque de inferência de atributo. Além disso descrevemos a utilidade de se publicar uma amostra para analistas de dados que têm como objetivo inferir a distribuição dos valores de um atributo em uma população.

O modelo foi desenvolvido utilizando o arcabouço *Quantitative Information Flow* (QIF) que fornece um aparato matemático para modelar formalmente sistemas como canais de informação. Nós desenvolvemos o modelo com o objetivo de ser facilmente explicável para não especialistas e para ser utilizado por curadores de dados quando estiverem tomando decisões sobre como publicar os seus dados. Como segunda contribuição, nós provemos fórmulas fechadas para vulnerabilidades à priori e à posteriori para ataques de inferência de atributo e para perda de utilidade à priori. As fórmulas fechadas são úteis para quantificar vulnerabilidades e perdas de utilidade em grandes amostras e populações.

Palavras-chave: Divulgação estatística. Amostragem. Privacidade. Fluxo de Informação Quantitativo.

Abstract

Statistical disclosure is a process that has been present in society for a long time, however the concern about privacy is relatively recent. The interest in protecting individual data increased considerably especially after the elaboration of regulations about data protection around the world, such as the General Data Protection Regulation (GDPR) in the European Union and the *Lei Geral de Proteção de Dados* (LGPD) in Brazil.

The effort in the scientific community to develop methods for the mitigation of privacy risks and to understand the trade-off between privacy and utility compose a large research area. However, mathematical models that explain formally this trade-off are, in some situations, misunderstood by data curators, i.e., entities that collect data from a population and adopt a certain policy to publish them can not understand what are the risks and benefits of that policy. In this sense, models and solutions that ensure that all parties involved are aware of the risks and benefits of each policy adopted are important for well informed decision-making.

As a first contribution of this work we propose a model that captures the vulnerability of publishing a sample from a population, in particular, the vulnerability of an attribute inference attack. We also describe the utility of the sample for data analysts who aim to infer the distribution of the values of an attribute in a population.

The model was developed using the framework of Quantitative Information Flow (QIF) that provides a mathematical apparatus to formally model systems as informational channels. We developed the model with the goal of being easily understandable by non experts and to be used by data curators when making decisions about how to publish their data. As a second contribution we provide closed formulas for prior and posterior vulnerabilities of attribute inference attack and for prior utility loss. The closed formulas are useful when quantifying vulnerabilities and utility losses in large datasets/samples.

Keywords: Statistical disclosure. Sampling. Privacy. Quantitative Information Flow.

List of Figures

List of Tables

List of Definitions, Theorems, etc.

Contents

Chapter 1

Introduction

Statistical disclosure is a procedure that has been present in society for a long time. Practices such as publication of demographic census by governments and the release of datasets by institutions and private companies become more common since the increment of computational power in the last decades.

The concern about privacy in public data releases increased considerably especially after the elaboration of Regulation 2016/679 of the European Union [16], known as GDPR - General Data Protection Regulation. These regulations have as object of analysis data of any type, from information that individuals put on social networks to censuses publications by governmental institutions. Since the creation of these regulations, works on statistical disclosure control gained more space in the scientific community.

There are several reasons, from both public and private entities, for publicly disclosing data about a group of individuals, i.e., there is a utility associated to each data release. For example, consider a hospital that has a database with information about its patients. In order to provide adequate treatment, doctors need sensitive information about the patient such as illnesses they have, personal life habits and others. These data are private patient information, and there is a relationship of trust between the doctor and the patient. On the other hand, many scientific researches in the health field depend on this data. Hospitals can contribute to the development of this kind of research by making their databases publicly available, but doing so without compromising the patients' privacy is a non-trivial challenge.

From the public part, consider as an example a government collecting data from its population in order to publish a census. Government policy makers can use the population's data to guide the creation of public policies. The knowledge about the population's living conditions is very important for the government when distributing resources. An illustration of this scenario is an institution in Brazil called *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* (Inep) [1] that is responsible for publishing Brazil's educational census. The institution collects various information from brazilian students from basic to higher education in order to support the formulation of educational policies at different levels of government, thus contributing to the economic and

[1] https://www.gov.br/inep/pt-br

social development of the country [20].

Although the census is considered very important by brazilian policy makers and researches at universities, a recent work published in 2021 [24] showed that Inep used to release microdata until 2019 leaks a lot of information about brazilian citizens. The technique used by the institution was de-identification, that consists in deleting all direct identifiers of participants (such as the Social Security Number and names) and release all other individual information. The authors exposed serious privacy breaches that were unacceptable according to LGPD.

When providing data to institutions, individuals expect their privacy to be guaranteed. In particular, there is the concern about attribute inference attack, i.e., when an individual is afraid of someone (an adversary) inferring information about him (in this case inferring an attribute value). Although there is a considerable effort in society to create efficient and robust mechanisms that can offer security to data, there are still vulnerabilities that are not efficiently contained. Traditionally, privacy is protected by encryption, access control and anonymization techniques. However, encryption and access control do not guarantee the privacy of sensitive information that can be inferred from public information, for instance, the inference of confidential information from public databases.

In the literature of statistical disclosure control, it is well known that there exists a trade-off between privacy and utility [11, 13, 17, 19]. In other words, providing some utility to data analysts when publishing data in general produces an inevitable loss of privacy. Within the challenges of researches in this area there are the trade-off characterization (e.g., studies that quantify the loss of privacy per unit of utility) and the development of methods to control data disclosure that guarantee acceptable levels of privacy and utility.

The cutting edge method used nowadays to guarantee privacy in statistical disclosures is differential privacy, first introduced by Dwork in 2006 [14]. Recent works [7, 8, 6, 5] that have improved privacy level in mechanisms that satisfies the properties of differential privacy are very important for the research area, however, a result that uses the technical language as, e.g., *"We propose a differentially private mechanism M_1 that uses a higher ε (which implies a lower perturbation in data thus keeping a higher utility) than mechanism M_2 keeping the same privacy level."* usually is not understandable by non experts in society and in institutions that publicly publish data. Explaining what some mathematical definitions and theorems represent in practical terms in the real world is a difficult task.

In this work we propose a model that is easily understandable by non experts and provide quantitative answers for some questions about privacy and utility in statistical disclosure. In particular, the model covers the process of publishing a sample from a population (i.e., a subset of records from the original dataset) that contains a single binary sensitive attribute. With the model we will be able to answer questions such as

"What is the expected probability of an adversary guessing correctly the attribute value of a single target when she observes a sample from the population?" regarding privacy concerns and *"When a data analyst observes a sample and tries to guess the distribution of a binary attribute in the population, how far her guess will be from the real distribution in expected values?"* regarding data utility.

The framework of Quantitative Information Flow (QIF) [3], a set of information– and decision-theoretic principles to reason about the flow of sensitive information through a system, provides a vast set of tools, definitions and properties that allows us to model scenarios of statistical disclosures that can be explainable to the society in general. Some recent works [4, 24, 21] have already used QIF to model some scenarios of statistical disclosures to quantify and explain, in easy terms, the vulnerabilities involved in those data publications.

1.1 Contributions

The contributions of this thesis are summarized as follows:

- We formally model attribute inference attack in sample publications of a single binary attribute using the framework of QIF. The attack covers an adversary that has a single target and wants to infer his attribute value. We consider 3 adversaries with different prior knowledge (i.e., what is known before an attack is performed) about the presence of the target in the sample release:

  (i) The adversary knows the target is in the sample,

  (ii) The adversary knows the target is outside the sample,

  (iii) The target's presence in the sample is not known by the adversary.

- We formally model utility in sample publications of a single binary attribute using the framework of QIF. The model describes a data analyst that tries to infer the attribute values distribution in the population.

- In order to capture a wider set of possible scenarios, we provide the analyses for both privacy and utility for two distinct group of adversaries who have different knowledge about how the data is generated, which we call $\mathcal{G}^f$ and $\mathcal{G}^d$. In $\mathcal{G}^f$ we model the adversary's prior knowledge as a uniform distribution on all possible attribute values distribution in the population In $\mathcal{G}^d$ we model the adversary's prior knowledge as a uniform distribution on all possible population datasets. These two

groups provide a wider view about the final information leakage when the prior knowledge about the population is different.

- We provide closed formulas for prior and posterior vulnerabilities of attribute inference attack and for prior utility loss. The closed formulas are useful when quantifying vulnerabilities and utility losses in large datasets/samples. Although we do not present a closed formula for posterior utility loss of adversaries in $\mathcal{G}^f$ and for prior and posterior utility loss for adversaries in $\mathcal{G}^d$, we provide an equations that can be computed in at most $O(n^3)$ and also an analysis about the behavior of those losses as the sample or population size increase.

1.2 Thesis outline

This thesis is organized as following. In Chapter 2 we present some related work in the research area on privacy and data disclosure and we present the basic definitions and theorems of the framework of QIF.

In Chapter 3 we present our model that uses the framework of QIF as basis to formally describe the scenario of a sample publication of a single binary attribute and the adversaries involved with privacy and utility.

In Chapter 4 we present lemmas and theorems related to the models described in Chapter 3, some of them providing closed formulas for privacy vulnerabilities and utility losses. In the end of this chapter we discuss the results about prior and posterior vulnerabilities/utility losses and compare their trade-off.

The last chapter concludes this thesis by making a review about the main contributions, the limitations and also discuss future work.

Chapter 2

Background

In this chapter we review the most important aspects about the research field on data disclosure control and the framework of QIF. We start by discussing important works about privacy and data disclosure control in Section 2.1. In the last section we detail the framework of QIF and present the main definitions that allowed us to build the models in Chapter 3 and derive the results about vulnerabilities and utility showed in Chapter 4.

2.1 Review on data disclosure control

The research area on data disclosure control started to be relevant in society decades ago. The concerns involving privacy of participants in a data release lead Tore Dalenius, in a work published in the 70's [9], to described a privacy goal for statistical databases:

> "Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access to the database."

As pointed out by Dwork in [13], different from cryptography, where we try to hide information from an adversary, statistical databases are published to be accessed by adversaries (some of them legitimate such as researches and data analysts), and indeed their purpose is to change beliefs about individuals. Because of that, she concluded that Dalenius's goal is not achievable in this scenario.

There are numerous cases in the history of breach of privacy caused by publicly data releases. A famous one is a work published by Sweeney in 2000 about the United States population [29]. She conducted experiments in the 1990 United States Census summary data and she found out that, using only basic information about individuals – more specifically, 5-digit ZIP, gender and date of birth – it was possible to uniquely identify 87% of the U.S. population. In other words, an adversary who knew the values

of these 3 attributes of a person that was a participant in that census, would have 87% of chance of finding exactly which record in the dataset corresponded to that person.

Another famous case is the work published by Narayanan and Shmatikov in 2008 [23]. The company Netflix promoted a competition, the Netflix Prize, that would award $1-million for the best improvement of their recommendation system for movies. They publicly released a dataset containing about 100 million of movie ratings from about 480,000 Netflix subscribers. The authors, crossing this dataset with another dataset from the Internet Movie Database (IMDB), successfully identified the Netflix records of known users, uncovering some sensitive information including their apparent political preferences.

These two cases illustrate that, only removing direct identifiable attributes (e.g., complete name and Social Security Number) about participants in a dataset and releasing all other information without any other treatment, can cause serious privacy breaches. In order to clarify the understanding of privacy risks, some researches published works with the goal of formalizing and categorizing them. Matthews et. al. in [22] present a literature review for privacy assessment where they highlight three privacy risks that are addressed by many works in the field:

(i) *re-identification risk* (also called *record linkage*), where an adversary tries to link a record from the database to its owner;

(ii) *attribute inference risk* (also called *attribute linkage* or *attribute disclosure*), where an adversary tries to guess the attribute value of a target;

(iii) *membership risk* (also called *population disclosure*), where an adversary tries to guess whether a person is present in the statistical publication or not.

In this work we focus our attention on attribute inference risk. We believe that inferring attribute values about individuals is the most important privacy concern present in data releases. Although more studies are needed to better understand the relationship between re-identification, membership and attribute inference risks, we give in the next paragraphs some motives that lead us to choose attribute inference as the object of study of this work.

Let us first take a look at membership inference risk. In general, when there is a concern about membership in a dataset, there is an implicit information associated to the presence or absence of a person in the dataset. For example, a hospital can publish a dataset containing information about patients that have a certain type of cancer. If an adversary found out that someone is in the dataset, she will automatically knows that the person has that type of cancer. The information *"Bob is in the dataset"*, by **itself**, does not cause harm to Bob if someone learns it about him. What Bob is concerned about is what the adversary can infer about him after learning he is in the dataset. Thus the real concern is about the person having cancer or not, and not the presence or absence of that person in the database.

In the case of re-identification risk, we can make a similar reasoning. We can model precisely the risk of *"An adversary linking a record from the dataset to an individual"*. Again, the information *"Bob is the record number 42"*, by **itself**, is not dangerous. The concern about this risk is that, after linking the record number 42 to Bob, the adversary may be able to look at all columns (each one representing an attribute) and infer exactly these attribute values about Bob.

Briefly, we are inclined to say that the real concern of individuals, when providing their data to a curator, is the risk of someone inferring their attribute values. As one goal of this work is to provide a model that is easily explainable to the society, we have decided to study attribute inference risk that is exactly the privacy concern of individuals in the real world.

Methods for preventing information leakage from third parties, e.g., cryptography, compose a large study area and there are many solutions available. The concerns about this scenario consider that a closed group of people (or entities) should have access to the data while external access must be blocked. However in public datasets the data must be available for everyone due to its purposes. We present below a list of different approaches that seems to be promising but actually have vulnerabilities:

- *Cryptography data does not work:* As discussed before, the data must be available for everyone (because its a public data) due to its purpose. Cryptography is effective against secret information that should not be revealed, but it is ineffective in statistical disclosures.

- *Data deidentification does not work:* The act of removing obvious identifiers such as the name and the Social Security Number does not protect the information leakage. Using auxiliar information (i.e., non-sensitive values such as date of birth and ZIP code) it is possible to infer sensitive information [11, 13].

- *Block queries to the dataset that return few records does not work:* Suppose queries in the form "Which records satisfy property P?" and a mechanism that blocks all queries that would return a set containing less than 10 records. The adversary can infer whether the target T satisfy property P doing the following: First she asks how many people in the dataset satisfy P, obtaining a result x. Second, she asks how many people in the dataset with a name different from T satisfy P, obtaining y as the result. Now the adversary calculates $x - y$ and can infer whether T satisfies P.

- *Allow only predefined types of queries does not work:* Imagine that a list of all allowed queries was created, intending to block queries related to sensitive information. However, it is possible to demonstrate that in query languages – such as SQL – does not exist an algorithm capable to check if two queries are the same or not.

- *Save the query history does not work:* Imagine that all the history is saved and used to check wheter every new query has a cross relation to the past ones. This approach fails for the same reasons from the last item.

Several works in the literature such as [30, 12, 17, 18, 19] put efforts in creating solutions to mitigate privacy risks. In the last decade and a half a method called differential privacy, first proposed by Dwork in [14], have been consolidated as the most refined method for guaranteeing privacy in statistical disclosures. Its popularity is largely due to its ability to significantly mitigate privacy breaches more effectively.

The research area in differential privacy is extremely active with new results and algorithms being proposed all the time. There are already companies applying differential privacy on their data publishing such as Google [15], Apple [25, 1] and Microsoft [10]. From governments, we have the United States Census Bureau (USCB) using differential privacy as the disclosure avoidance system in the US 2020 Decenal Census [2].

The usage of statistical disclosure control techniques and the interpretation of their guarantees, sometimes, are not totally clear for the final user, i.e., the entities (data owners) that will implement those methods in their data publications. Matthews et. al. pointed out in [22]:

> "While many methods of preserving privacy have been proposed, there are not, as of yet, any formal guidelines for many data releasing institutions to follow when releasing data to the public."

The lack of formal guidelines for institutions to publish their data is one of the main motivations for the work presented in this thesis. One of the contributions rely in models that are proposed to be explainable and that provide numerical values that are easily understandable and may allow data owners to take decisions about how to publish their data.

2.2 Quantitative Information Flow

Quantitative Information Flow (QIF) is a framework used to model how information flows in a system. Computer systems can be seen as black boxes that take some input and yield some output. One concern about this process is how much information is leaking from inputs to outputs, and QIF allows us to quantify this amount of leakage. In the following sections we describe how to model systems, inputs, outputs and information leakage.

2.2.1 Secrets

In QIF framework we call *secrets* the set of values we are interested to protect against an "adversary". Secrets can be a user's password, someone's age, political preferences, religion or any other relevant information. An *adversary* is an entity (e.g., a person or a company) which we are concerned about learning the value of a secret.

For instance, imagine that our system is a password checker C of a bank. Its task is to receive a password – a four digit number – as input and output a value in $\mathcal{Y} = \{\texttt{yes}, \texttt{no}\}$ depending on the password being correct or not for a given user. In this scenario the secret can be defined as the user's password. The set of all possible secrets is then $\mathcal{X} = \{\texttt{0000}, \ldots, \texttt{9999}\}$.

Suppose we are modeling the password checker behavior for Jennifer's password, which is, say, 3482. Let us denote by $C^{\texttt{3482}}$ the password checker behavior when the user's password is 3482. In order to check Jennifer's password, $C^{\texttt{3482}}$ will output yes if it receives $x = \texttt{3482}$ as input or output no if it receives any other number.

We are interested in measuring how much leakage the password checker yields when it outputs some $y \in \mathcal{Y}$ and how the system modifies the adversary's knowledge about Jennifer's password. But how can we describe the adversary knowledge about the secret? One way of doing that is using probability distributions. We call $\mathbb{D}\mathcal{X}$ the set of all possible probability distributions on $\mathcal{X}$. A probability distribution $\delta \in \mathbb{D}\mathcal{X}$ assigns, for all $x \in \mathcal{X}$, a probability δ_x the adversary attributes to x being the real secret. For example, if the adversary does not know anything about Jennifer's password, her knowledge is going to be a uniform distribution on $\mathcal{X}$, that is, $\delta_{\texttt{0000}} = \ldots = \delta_{\texttt{9999}} = 10^{-4}$.

Definition 2.2.1 (Secrets and prior)**.** *Given a finite set of secrets $\mathcal{X}$ we assume that the adversary's prior knowledge about the set is a probability distribution $\pi \in \mathbb{D}\mathcal{X}$ that specifies a probability π_x for all $x \in \mathcal{X}$.*

The closer to the uniform distribution the adversary knowledge is, the less is the "secrecy" about the secret. On the other hand, if the adversary's knowledge is a point distribution,[1] she is 100% sure about what is the secret's value. Table 2.1 shows examples of two different prior knowledges. We introduce next the concept of g-vulnerability.

[1] We say that a probability distribution is a *point distribution* when $Pr[x_i]$=1 for some i and $\forall j \neq i$: $Pr[x_j]$=0.

x	π_x
0000	10^{-4}
0001	10^{-4}
0002	10^{-4}
⋮	⋮
9999	10^{-4}

(a) The prior distribution π models an adversary that has no ideia about the password's value.

x	π'_x
0000	$1/2$
0001	$1/2$
0002	0
⋮	⋮
9999	0

(b) The prior distribution π' models an adversary that is sure the password is either 0000 or 0001, and equally likely.

Table 2.1: Different prior knowledges about the passwords that represent (a) an adversary with high uncertainty and (b) an adversary with low uncertainty.

2.2.2 g-vulnerability

We have already seen that the adversary has a prior knowledge about the secret – a probability distribution $\pi \in \mathbb{D}\mathcal{X}$. One possible view about what the adversary actually does with this knowledge is saying that she takes "actions" in order to get a "reward". In the password checker example (the system C^{3482} that checks whether the input corresponds to Jennifer's password), suppose the adversary is someone that wants to steal Jennifer's money. The adversary's action could be a guess for the password (i.e., a four digit number) and the adversary's reward could be (i) 100 000 dollars if the guess is correct (meaning that she can get all the money in the account) or; (ii) 0 dollars if the guess is incorrect (meaning that she gets nothing). In QIF this kind of reasoning can be modeled as a *gain function*.

Definition 2.2.2 (Gain function)**.** *Given a finite nonempty set of possible secrets $\mathcal{X}$ and a nonempty set of possible actions $\mathcal{W}$, a gain function is a function $g\colon \mathcal{W} \times \mathcal{X} \to \mathbb{R}$.*

Formally, we can assume the adversary is going to take an action w from a set of actions $\mathcal{W}$, and each action will give her a reward for each secret. The value $g(w, x)$ can be interpreted as "*what the adversary gains when she takes an action w and the real value of the secret is x*". Once the adversary's prior knowledge π about the set of secrets $\mathcal{X}$ and a gain function g are defined, we are ready to examine the secret secrecy, or its "vulnerability".

Considering the adversary is rational and will take the action that maximizes her gain, the *prior vulnerability* $V_g(\pi)$ will be the maximization of the expected gain of the adversary, and it is formalized next.

Definition 2.2.3 (g-vulnerability)**.** *Given a distribution $\pi \in \mathbb{D}\mathcal{X}$ and a gain function $g\colon \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the g-vulnerability of π is defined as*

$$V_g(\pi) := \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi_x \; g(w, x) \; .$$

Consider again the password checker example with Jennifer's bank password and the adversary that wants to steal her money. Let's define the gain function g as

$$g(w, x) = \begin{cases} \$100\,000 & \text{, if } w = x \\ \$0 & \text{, otherwise.} \end{cases}$$

Looking at the prior distributions π and π' from Table 2.1 and Definition 2.2.3 of prior vulnerability, we conclude that

- $V_g(\pi) = 10^{-4} \times \$100\,000 = \$10$.
- $V_g(\pi') = 1/2 \times \$100\,000 = \$50\,000$.

The adversary with high uncertainty (π) has a lower expected gain (only \$10), and the adversary with low uncertainty (π') has a higher expected gain (\$50 000).

Bayes vulnerability is another example of measure that is very useful to deal with a basic security concern: the probability of an adversary guessing correctly the real value of the secret in one try. This measure is described by the gain function g_{id}, defined by the identity matrix, i.e., $g_{id}(w, x){=}1$ if $w{=}x$ and $g_{id}(w, x){=}0$ if $w{\neq}x$.

Definition 2.2.4 (Bayes vulnerability)**.** *Given a finite set of secrets $\mathcal{X}$ and a prior distribution $\pi \in \mathbb{D}\mathcal{X}$, the Bayes vulnerability of π is defined as*

$$V_1(\pi) := \max_{x \in \mathcal{X}} \pi_x \; .$$

In Bayes vulnerability the adversary is guessing what's the secret value, so the set of actions $\mathcal{W} = \mathcal{X}$ and she gains 1 when her guess correct or 0 otherwise.

From the adversary's point of view, g-vulnerability is a maximization of gain, and it measures the threat to secrets. A complementary approach is to describe an adversary that take actions and, instead of measuring her gains, we measure her "losses". In this way we define next ℓ-uncertainty, that uses a loss function ℓ to measure the adversary's uncertainty about the secrets.

Definition 2.2.5 (Loss function)**.** *Given a prior distribution $\pi \in \mathbb{D}\mathcal{X}$ and a loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, the ℓ-uncertainty of π is*

$$U_\ell(\pi) := \min_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi_x \, \ell(w, x) \ .$$

In the next section we are going to introduce channels, objects that model systems that receive inputs and output observables. We will make the connection between secrets, adversaries and prior knowledge in order to reason about the adversary's posterior knowledge (i.e., after she observed a channel's output).

2.2.3 Channels

In QIF a system is modeled as an *information-theoretic* channel, i.e., a probabilistic function from inputs to outputs. We say that a channel $\mathsf{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ is a channel that maps secrets $x \in \mathcal{X}$ to probability distributions on outputs $y \in \mathcal{Y}$. It is possible to represent C as a matrix with $|\mathcal{X}|$ rows and $|\mathcal{Y}|$ such that $\mathsf{C}_{x,y} = Pr[y|x]$, i.e., the probability of the system outputs y given that the real value of the secret is x. Consequently, each row of C is going to be a probability distribution over $\mathcal{Y}$.

Definition 2.2.6 (Channel)**.** *Given a finite set of secrets $\mathcal{X}$ and a finite set of outputs $\mathcal{Y}$, a channel $\mathsf{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ is a mapping from secrets to probability distributions on $\mathcal{Y}$. The channel can be represented by a matrix such that each entry $\mathsf{C}_{x,y} = Pr[y \mid x]$ means "the probability of the output being y given that the secret is x".*

Recall the example of a password checker and also our character Jennifer whose password is 3482. We can define a channel C^{3482} that describes the behavior of our password checker C^{3482}. It would be defined in the following way:

$$\mathsf{C}^{3482}_{x,y} = \begin{cases} 1, & \text{if } (x{=}\texttt{3482} \text{ and } y{=}\texttt{yes}) \text{ or } (x{\neq}\texttt{3482} \text{ and } y{=}\texttt{no}) \\ 0, & \text{otherwise .} \end{cases}$$

Note that we can define other channels C^k for any four digit password $k \neq \texttt{3482}$ in a similar way.

Now we are ready to describe the effect of a channel on adversary's knowledge. We are interested in analyzing how information about the secret changes when our channel

outputs some value. We assume she knows how the system works, i.e., all entries of matrix C. After observing the output y, the adversary will update her knowledge (a probability distribution over $\mathcal{X}$) about the secret using Bayesian reasoning, i.e., the prior distribution on $\mathcal{X}$ becomes a posterior distribution $\delta^y = Pr[\mathcal{X} \mid \mathcal{Y} = y]$. After this update the adversary will have a *posterior knowledge* about the secret.

We also consider that our adversary is rational, so given her posterior knowledge about the secret and a gain function, she will choose the guess that maximizes her expected gain (or minimizes her expected loss if there is a loss function instead of a gain function).

To calculate the posterior distribution $Pr[\mathcal{X} \mid \mathcal{Y} = y]$, we first calculate the joint distribution $Pr[x, y]$ for all pairs (x, y). Fixing a prior distribution π and given that π_x is the probability *a priori* of x being the secret, the joint distribution can be defined as

$$Pr[x, y] = \pi_x \mathsf{C}_{x,y} \ . \tag{2.1}$$

We can organize the joint distribution in a matrix J with $|\mathcal{X}|$ rows and $|\mathcal{Y}|$ columns such that $\mathsf{J}_{x,y} = Pr[x, y]$. Now we need to calculate $Pr[y]$. If we look at J's columns, it is possible to calculate the marginal distribution on columns, and that is exactly a distribution on $\mathcal{Y}$. Formally, we have that

$$Pr[y] = \sum_{x \in \mathcal{X}} \mathsf{J}_{x,y} \ .$$

It is important to note that for each output $y \in \mathcal{Y}$ we get a new posterior probability distribution over the set of secrets $\mathcal{X}$. Each channel output is a possible "world", and each possible world is a probability distribution on $\mathcal{X}$. We call the possible worlds the *inner distributions*. Each possible world has a probability to occur, and we call the distribution on the possible worlds as the *outer distribution*. Therefore the adversary's posterior knowledge is going to be a distributions on distributions on $\mathcal{X}$ (i.e., $\mathbb{D}(\mathbb{D}\mathcal{X})$), also called a *hyper-distribution* $[\pi \triangleright \mathsf{C}]$.

Definition 2.2.7 (Hyper-distribution)**.** *Given a set of secrets $\mathcal{X}$, a prior distribution $\pi : \mathbb{D}\mathcal{X}$ and a channel $\mathsf{C} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$, a hyper-distribution $[\pi \triangleright \mathsf{C}]$ is a distribution on distributions on $\mathcal{X}$, i.e., $[\pi \triangleright \mathsf{C}] \in \mathbb{D}(\mathbb{D}\mathcal{X})$. We call the inner distributions of $[\pi \triangleright \mathsf{C}]$ the set of all possible "worlds" (posterior distributions $\delta \in \mathbb{D}\mathcal{X}$), and we call the outer distribution of $[\pi \triangleright \mathsf{C}]$ the distribution on inners themselves.*

Let us illustrate these definitions with an example. Let $\pi = (1/2, 1/3, 1/6)$ be the prior distribution on the set of secrets $\mathcal{X} = \{x_1, x_2, x_3\}$ and let the matrix in Table 2.2a be the representation of a channel C with the set of possible outputs $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$. Using Eq. (2.1) we can build the joint matrix J in Table 2.2b. As discussed above, the

C	y_1	y_2	y_3	y_4
x_1	$1/4$	$1/4$	$1/4$	$1/4$
x_2	$1/2$	$1/2$	0	0
x_3	0	$2/3$	0	$1/3$

(a) Channel matrix where $\mathsf{C}_{x,y} = Pr[y \mid x]$.

J	y_1	y_2	y_3	y_4
x_1	$1/8$	$1/8$	$1/8$	$1/8$
x_2	$1/6$	$1/6$	0	0
x_3	0	$1/9$	0	$1/18$

(b) Joint matrix where $\mathsf{J}_{x,y} = Pr[x, y]$.

$[\pi \triangleright \mathsf{C}]$	$7/24$	$29/72$	$1/8$	$13/72$
x_1	$3/7$	$9/29$	1	$9/13$
x_2	$4/7$	$12/29$	0	0
x_3	0	$8/29$	0	$4/13$

(c) Hyper-distribution.

Table 2.2: For a set of secrets $\mathcal{X} = \{x_1, x_2, x_3\}$, a prior distribution $\pi = (1/2, 1/3, 1/6)$ and a channel matrix C, the three tables represent the process of building a hyper-distribution $[\pi \triangleright \mathsf{C}]$. Each column in the hyper-distribution represents an adversary's posterior knowledge about the secret (the inner distributions), and the first row (with values $(7/24, 29/72, 1/8, 13/72)$) is the outer distribution on all possible posterior knowledge.

adversary's posterior knowledge is a hyper-distribution $[\pi \triangleright \mathsf{C}]$, and we can write both outer and inner distributions in a single matrix as showed in Table 2.2c.

For instance, the column $(3/7, 4/7, 0)$ is a inner distribution and it will occur with a probability $7/24$. The fourth column of Table 2.2c says that there is an observable that happens with probability $1/8$ and gives the information to the adversary that x_1 is the secret's value.

Definition 2.2.8 (Posterior vulnerability)**.** *Given a set of secrets $\mathcal{X}$, a prior distribution $\pi \in \mathbb{D}\mathcal{X}$, a gain function g and a channel C, the posterior vulnerability $V_g[\pi \triangleright \mathsf{C}]$ that represents the expected gain of the adversary after observing the output of channel C is defined as*

$$\begin{aligned} V_g[\pi \triangleright \mathsf{C}] &:= \sum_{y \in \mathcal{Y}} \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \pi_x \, \mathsf{C}_{x,y} \; g(w, (p, t)) \\ &= \sum_{y \in \mathcal{Y}} Pr[y] V_g(\delta^y) \;, \end{aligned}$$

where $V_g(\delta^y)$ is the posterior vulnerability given that the adversary observed the output y. The posterior distribution $\delta^y \in \mathbb{D}\mathcal{X}$ (i.e., $Pr[\mathcal{X} \mid \mathcal{Y} = y]$) represents the adversary's posterior knowledge about the set of secrets when she observes the output y.

The matrix in Table 2.2c represents all the adversary's posterior knowledge, and the equation in Definiton 2.2.8 says that, given the adversary observed an output y, she will take the action w that maximizes her gain, and therefore the posterior vulnerability will be the average over all possible outputs.

Definition 2.2.9 (Posterior Bayes vulnerability)**.** *Given a finite set of secrets $\mathcal{X}$, a prior distribution $\pi \in \mathbb{D}\mathcal{X}$ and a channel $\mathsf{C} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y} = \mathcal{X}$, the posterior Bayes vulnerability is defined as*

$$V_1[\pi \rhd \mathsf{C}] := \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} \mathsf{J}_{x,y} \ ,$$

where J is the joint distribution from Eq. 2.1.

Recall that Bayes represents an adversary that is trying to guess the real value of the secret in one try. The equation presented in Definition 2.2.9 was derived from the equation in Definition 2.2.8, and its proof is available in Chapter 5.3 of [3].

In the next section we are going to discuss more about *additive* and *multiplicative* leakage, two comparisons between posterior and prior vulnerabilities that allows us to compare quantitatively systems that are leaking more or less information.

2.2.4 g-leakage

The posterior vulnerability in itself is already a good measure for system's secrecy. However, if the adversary's knowledge about the secret was very high before he observing the channel's output (i.e., the prior vulnerability $V_g(\pi)$), then his posterior vulnerability will also be high[2]. Looking only to posterior vulnerability, ignoring the prior, can lead to a misinterpretation of the vulnerability value.

In this way we introduce the definition of *additive* and *multiplicative* leakage, the absolute and relative difference between the posterior and prior vulnerabilities, respectively.

Definition 2.2.10 (Additive and Multiplicative g-leakage)**.** *Given a prior distribution $\pi \in \mathbb{D}\mathcal{X}$, a gain function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ and a channel $\mathsf{C} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the additive leakage $\mathcal{L}^+_g(\pi, \mathsf{C})$ is defined as*

$$\mathcal{L}^+_g(\pi, \mathsf{C}) := V_g[\pi \rhd \mathsf{C}] - V_g(\pi) \ ,$$

[2]It happens due to Monotonicity axiom which states that the adversary cannot have her knowledge about the secret (in expected value) decreased after observing the channel. In plain English, "information can't hurt". Formally this axiom states that $V_g[\pi \rhd \mathsf{C}] \geq V_g(\pi)$, for all π, C and gain function g. The proof can be found in Chapter 11 of [3].

and the multiplicative leakage $\mathcal{L}_g^{\times}(\pi, \mathsf{C})$ *is defined as*

$$\mathcal{L}_g^{\times}(\pi, \mathsf{C}) := \frac{V_g[\pi \rhd \mathsf{C}]}{V_g(\pi)} \ .$$

The usage of one measure or the other depends on the scenario, but in any case, looking at both brings us a more complete analysis about information leakage.

Chapter 3

Model for privacy and utility analyses

In this chapter we formalize adversary models for both analysis on privacy and utility. In the next section we present a detailed description of the general scenario captured by the models proposed in this work. In Section 3.2 we describe and define two groups of adversaries that we make two distinct assumptions about their prior knowledge about the set of secrets. Section 3.3 shows the adversary model that deals with privacy concerns on sample publications and Section 3.4 presents the adversary model for a data analyst who will be used to reason about the data publication's utility.

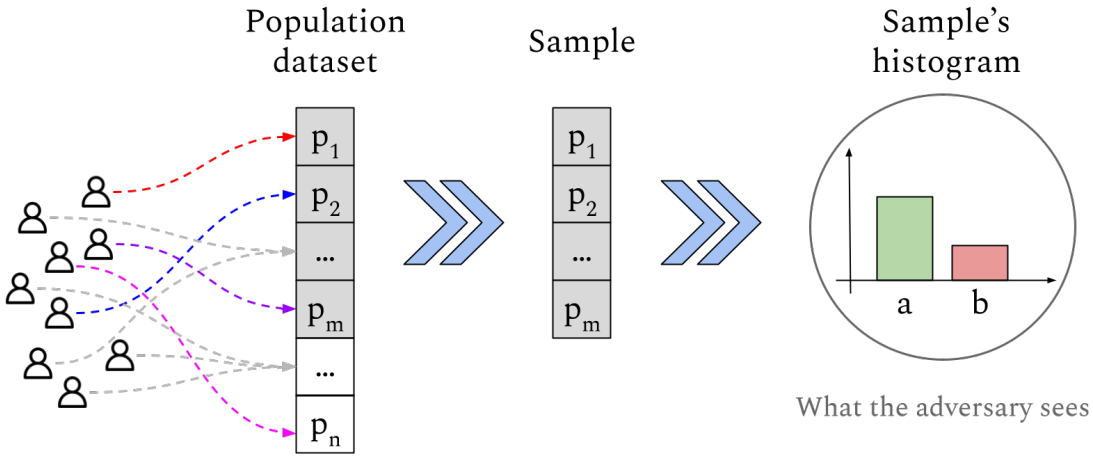
3.1 General scenario

Consider a scenario with a population of $n \geq 1$ individuals and a binary sensitive attribute with possible value in $\mathcal{V} = \{\mathtt{a}, \mathtt{b}\}$. Imagine there is a curator who collects and builds a dataset containing the attribute values of all n people from the population. Then, from the generated dataset, the curator selects a sample and publishes its histogram, i.e., the distribution on values `a` and `b`. Being more specific about this process, we are going to consider that a sample of size $1 \leq m \leq n$ will be randomly selected from the population (i.e., any set of m individuals is equally probable to be selected). A diagram of this process is shown in Figure 3.1.

As we are restricting to binary attributes, we can say that the publication is just an integer $0 \leq y \leq m$ that represents the number of people in the sample with value `a`, and consequently, $m - y$ will be the number of people in the sample with value `b`. We assume the adversary has access only to the sample's histogram and the number of people in the population n, i.e., she does not know anything about the values in the population dataset.

Publishing a sample's histogram may allow data analysts to make inferences about

Figure 3.1: Pipeline of sample publication. The curator collects data from a population of n people and arrange them in random positions in an array p. Then the first m values of p are selected to be the sample, and finally the curator publishes the sample's histogram.



Font: Elaborated by the author.

the population, but at the same time, each individual does not desire that some adversary, using the released data, become able to infer his exact (or approximate) attribute value. Therefore, statistical disclosures has the challenge of providing a high level of utility to data analysts and protecting participants privacy. As discussed in Section 2.1 there are several evidences in the literature that publishing data with very high levels of both privacy and utility is not achievable. Our results showed in Chapter 4 corroborates this claim.

After stated the general scenario studied in this thesis, in the next section we start reasoning about the adversary's assumptions about the population and then we formalize the models based in QIF framework that permit our analyses about privacy and utility in Chapter 4.

3.2 Assumptions about adversaries' prior knowledge

In this section we present two ways of modeling the adversary's prior knowledge about the scenario described in the previous section. A system designer, when modeling its security, usually wants to protect the system against the highest number of possible adversaries, or, in the perfect scenario, against all of them. However, such claim is very hard to be done or sometimes not even doable. Differential privacy (the considered state-of-art in privacy protection, as mentioned in Section 2.1) for example, states a protection against a called "strong adversary", someone that knows the value of everyone in the

dataset but the target's value. In this work we focus in two group of weaker adversaries who have no information about the participants' values.

Consider an arbitrary situation where there is a set of elements R where only one element $r \in R$ can be the true value in the real world. Say there is an adversary trying to find what is the true r. In the field of statistics, the adversary's thinking can be modeled as a probability distribution on R, where each $Pr[r]$ corresponds to the probability the adversary gives to r being the true value. One way of representing a lack of knowledge about R, i.e., an adversary that does not know anything about any r, is saying that the probability distribution on R is the uniform one [1].

One relevant purpose of publishing a sample's histogram is to allow the consumers of this data to infer the distribution of attribute values in the population. Considering the case of a binary attribute studied in this work and following the reasoning in the last paragraph, we assume that the group of adversaries $\mathcal{G}^f$ is composed by the adversaries that assume that every possible distribution on $\{\mathtt{a}, \mathtt{b}\}$ in the population is equally probable, i.e., all possible frequencies of value $\mathtt{a}$ in the population has the same probability to occur.

Formally, as the population size is n, each distribution on $\{\mathtt{a}, \mathtt{b}\}$ can be represented by an integer $0 \leq k \leq n$ where k is the number of individuals with value $\mathtt{a}$ and $n-k$ is the number of individuals with value $\mathtt{b}$. Therefore the uniform distribution implies in $Pr[k] = 1/(n+1)$ for all k. From this uniform distribution we derive the distribution on population datasets. In Table 3.1a we show an example for a population with 3 individuals. The models for privacy and utility analyses related to $\mathcal{G}^f$ are formally defined in Sections 3.3.1 and 3.4.1, respectively.

Another possible assumption that can be done about the adversary's prior knowledge is about the distribution on population datasets. We then introduce the group of adversaries $\mathcal{G}^d$ that is composed by the adversaries that does not know anything about how the dataset was generated. The lack of knowledge in this situation is represented by a uniform distribution on all possible population datasets [2]. An example is shown in Table 3.1b. The formal definitions for privacy and utility analyses for $\mathcal{G}^d$ are presented in Sections 3.3.2 and 3.4.2, respectively.

We introduce now, using elements from QIF framework, formal mathematical definitions for privacy and utility analyses for the general scenario described in Section 3.1 and the groups of adversaries described in this section.

[1] One motivation for that comes from information theory and the maximum entropy associated to the uniform distribution.

[2] Note that this adversary can also be viewed as someone who assumes that each individual in the dataset is a random variable X with possible outcomes in $\{\mathtt{a}, \mathtt{b}\}$ and with probability distribution $Pr[X{=}\mathtt{a}] = Pr[X{=}\mathtt{b}] = 1/2$. In this case a dataset would be a set of independent and identically distributed random variables X.

Population dataset p	Frequency of a	Probability of frequency	$Pr[p]$
aaa	3	$1/4$	$1/4$
aab	2	$1/4$	$1/12$
aba			$1/12$
baa			$1/12$
abb	1	$1/4$	$1/12$
bab			$1/12$
bba			$1/12$
bbb	0	$1/4$	$1/4$

(a) Group of adversaries $\mathcal{G}^f$ that assumes a uniform distribution on all possible frequencies of value a in the population.

Population dataset p	$Pr[p]$
aaa	$1/8$
aab	$1/8$
aba	$1/8$
baa	$1/8$
abb	$1/8$
bab	$1/8$
bba	$1/8$
bbb	$1/8$

(b) Group of adversaries $\mathcal{G}^d$ that assumes a uniform distribution on all possible datasets.

Table 3.1: Example of prior knowledge of adversaries in $\mathcal{G}^f$ and $\mathcal{G}^d$ for a scenario with n=3 individuals in the population. Each value p_i of a population dataset array p represents the attribute value of a participant. Table 3.1a shows the group of adversaries that assumes a uniform distribution on possible frequencies and Table 3.1b shows the group of adversaries that assumes a uniform distribution on datasets.

3.3 Adversary model for privacy analysis

In the scenario described in Figure 3.1 suppose there is an adversary who is interested in inferring the attribute value of a single person from the population, that we will call the *target*. A population dataset can be represented as a binary string of size n where each position is an individual's attribute value. The set of all possible datasets is then the set of all possible binary strings of size n. The adversary's knowledge about the scenario includes the population dataset and the position of his target in this dataset (because she wants to infer the target's attribute value). Using g-vulnerability framework detailed in Section 2.2.2, we now define elements from QIF theory to model the entire scenario, from possible secrets, datasets and adversary behavior.

Definition 3.3.1 (Set of secrets $\mathcal{X}$). *Let $n \geq 1$ be the population size and consider a binary attribute with domain $\mathcal{V} = \{\mathtt{a}, \mathtt{b}\}$. The set of secrets $\mathcal{X}$ of all possible populations and target's index is defined as*

$$\mathcal{X} = \{(p, t) \mid p \in \mathcal{V}^n \wedge 1 \leq t \leq n\}, \tag{3.1}$$

where $\mathcal{V}^n$ is the set of possible binary arrays of size n and a secret (p, t) is a pair

where p is the population array and t is the target's index in p. We denote by p_i the attribute value of the i-th person.

Example 3.1 (Running example – set of secrets $\mathcal{X}$)**.** *Consider a scenario in which there is a population with $n=2$ individuals. We have that $\mathcal{V}^2 = \{\texttt{aa}, \texttt{ab}, \texttt{ba}, \texttt{bb}\}$ is the set of all possible populations. The adversary's target can be either the first or the second individual in the population array, thus the set of secrets is*

$$\mathcal{X} = \{(\texttt{aa}, 1), (\texttt{aa}, 2), (\texttt{ab}, 1), (\texttt{ab}, 2), (\texttt{ba}, 1), (\texttt{ba}, 2), (\texttt{bb}, 1), (\texttt{bb}, 2)\} \ .$$

The secret $(\texttt{ab}, 2)$, for example, represents the scenario where the population dataset is $p = \texttt{ab}$ and the adversary's target is the second person in $\texttt{ab}$.

In Section 3.2 we described two group of adversaries that make different assumptions about probability distribution on the set of all possible populations. For privacy analysis, as each secret is composed by both the population dataset and the target's index, we then need to define what is the adversary's knowledge about the target's index. Following the scheme in Figure 3.1, the target can be included or not in the published sample depending on its index value being greater or smaller or equal to m.

In order to capture all possible scenarios, we define three different adversaries with all the three possible knowledge about the target's presence in the published sample:

- *Adversary $\mathcal{A}^{in}$:* She knows the target is **in** the sample;
- *Adversary $\mathcal{A}^{out}$:* She knows the target is **outside** the sample;
- *Adversary $\mathcal{A}^{nk}$:* The target's presence in the sample is **not known**.

Although the adversary $\mathcal{A}^{nk}$ probably represents the most common case that could be found in a real scenario, the others also represent realistic situations. For instance, imagine there is an institution doing an election poll in a country. They select a sample from the population and ask everyone whether they are going to vote in candidate A or B. Every employee from the institution that has access to the raw collected data knows who from the population is in the sample and who is not, and these employees have the prior knowledge of $\mathcal{A}^{in}$ and $\mathcal{A}^{out}$, respectively.

We are now ready to define the adversary's prior knowledge.

3.3.1 Group of adversaries $\mathcal{G}^f$

In this section we present the definitions related adversaries in $\mathcal{G}^f$, described in Section 3.2 and exemplified in Table 3.1a. Summarizing the adversary's prior knowledge, we have that:

- She knows that the population and sample sizes are n and m respectively,
- She assumes a uniform distribution on all possible frequencies of value $\mathtt{a}$ in the population,
- For $\mathcal{A}^{in}$, she knows the target is in the sample, for $\mathcal{A}^{out}$, she knows the target is outside the sample and for $\mathcal{A}^{nk}$, the target's presence in the sample is not known.

We then define the prior distributions $\pi^{in}, \pi^{out}, \pi^{nk} \in \mathbb{D}\mathcal{X}$ to formally model the prior knowledge of Adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$and $\mathcal{A}^{nk}$, respectively.

Definition 3.3.2 (Prior distributions for adversaries in $\mathcal{G}^f$)**.** *Let $\mathcal{X}$ be the set of secrets. The adversary knows that the population and sample's size are n and $1 \leq m \leq n$, respectively. She also assumes a uniform distribution on frequencies of value $\mathtt{a}$ in the population, as well as a uniform on datasets within a frequency (see Section 3.2 and Table 3.1a). We then define the prior distributions $\pi^{in}, \pi^{out}, \pi^{nk}$ for three different adversaries with distinct information about the target's presence in the sample:*

*(i) Adversary $\mathcal{A}^{in}$ that knows the target is **in** the sample:*

$$\pi^{in}_{(p,t)} = \begin{cases} \dfrac{1}{m(n+1)\binom{n}{n_{\mathtt{a}}(p)}} & , \text{ if } 1 \leq t \leq m \\ 0 & , \text{ otherwise.} \end{cases} \tag{3.2}$$

*(ii) Adversary $\mathcal{A}^{out}$ that knows the target is **outside** the sample:*

$$\pi^{out}_{(p,t)} = \begin{cases} \dfrac{1}{(n-m)(n+1)\binom{n}{n_{\mathtt{a}}(p)}} & , \text{ if } m < t \leq n \\ 0 & , \text{ otherwise.} \end{cases} \tag{3.3}$$

*(iii) Adversary $\mathcal{A}^{nk}$ whose information about the presence of the target in the sample is **not known**:*

$$\pi^{nk}_{(p,t)} = \frac{1}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}}, \tag{3.4}$$

where $n_{\mathtt{a}}(p)$ is the number of $\mathtt{a}$'s in array p.

Figure 3.2 shows a scenario of a population with 2 individuals and how priors π^{in}, π^{out} and π^{nk} would be defined.

Example 3.2 (Running example – prior distributions π^{in}, π^{out} and π^{nk})**.** *Consider the same scenario of Example 3.1 of a population with $n=2$ individuals, and also suppose the sample size is $m=1$. Figure 3.2 resumes this example.*

*For adversary $\mathcal{A}^{in}$, as she knows the target is **in** the sample, she gives probability zero to all secrets $(p,t) \in \{(\mathtt{aa},2),(\mathtt{ab},2),(\mathtt{ba},2),(\mathtt{bb},2)\}$. And for the other secrets, she assumes a uniform on frequencies and a uniform on datasets within a frequency. So $\pi^{in}_{(\mathtt{aa},1)} = 1/3$ because the probability $Pr[n_{\mathtt{a}}(p)=2] = 1/3$ (i.e., the probability of the number of $\mathtt{a}$'s in the population being 2) and there is only one dataset with 2 $\mathtt{a}$'s. On the other hand, $\pi^{in}_{(\mathtt{ab},1)} = 1/6$ because $Pr[n_{\mathtt{a}}(p)=1] = 1/3$ and there are 2 datasets with 1 $\mathtt{a}$ ($\mathtt{ab}$ and $\mathtt{ba}$), so the probability $1/3$ is distributed between these two datasets resulting in a probability $1/6$ for each one.*

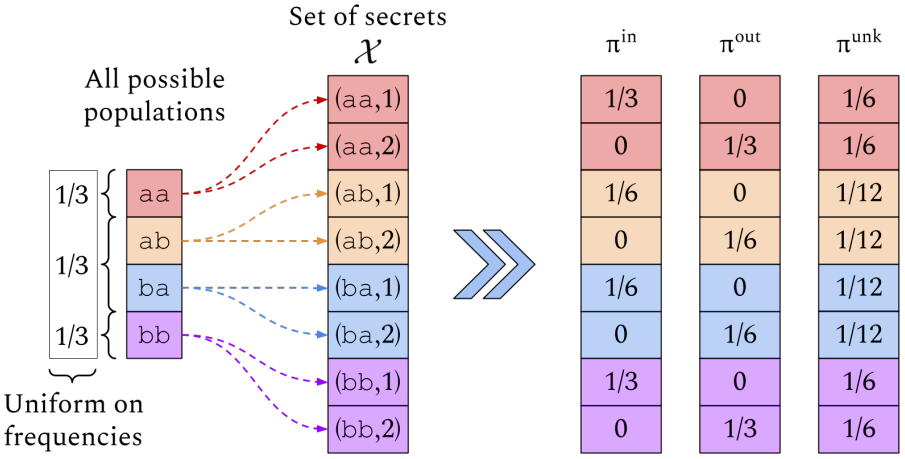
*In the case of adversary $\mathcal{A}^{out}$, as she knows the target is **outside** the sample, she gives probability zero to all secrets $(p,t) \in \{(\mathtt{aa},1),(\mathtt{ab},1),(\mathtt{ba},1),(\mathtt{bb},1)\}$. Thus, using the same reasoning we have used for $\mathcal{A}^{in}$, $\pi^{out}_{(\mathtt{aa},2)} = 1/3$ and $\pi^{out}_{(\mathtt{ab},2)} = 1/6$.*

Finally, adversary $\mathcal{A}^{nk}$ does not know anything about the target's presence in the sample, then the target's index can be both 1 or 2. Then, we have that $\pi^{nk}_{(\mathtt{aa},1)} = \pi^{nk}_{(\mathtt{aa},2)} = 1/6$ because $Pr[n_{\mathtt{a}}(p)=2] = 1/3$ and there are 2 options for the target's index. Also, $\pi^{nk}_{(\mathtt{ab},1)} = \pi^{nk}_{(\mathtt{ab},2)} = \pi^{nk}_{(\mathtt{ba},1)} = \pi^{nk}_{(\mathtt{ba},2)} = 1/12$ because $Pr[n_{\mathtt{a}}(p)=1] = 1/3$ and there are 2 datasets with 2 $\mathtt{a}$'s and, for each one, 2 options for the target's index.

3.3.2 Group of adversaries $\mathcal{G}^d$

In this section we present the prior knowledge formal definition for the group of adversaries $\mathcal{G}^d$ that assumes a uniform distribution on datasets, as detailed in Section 3.2

Figure 3.2: Example of prior distributions π^{in}, π^{out} and π^{nk} for a population of size $n = 2$ and a sample of size $m = 1$. The secret $(\texttt{ab}, 1)$, for instance, represents a population with 2 individuals where the first has value `a`, the second has value `b` and the adversary's target position in the population array is 1.



Font: Elaborated by the author.

and exemplified in Table 3.1b. The prior knowledge for this group of adversaries is composed by:

- The population and sample sizes n and m respectively,
- The assumption of a uniform distribution on all possible datasets,
- The information that the target is in the sample for adversary $\mathcal{A}^{in}$, that the target is outside the sample for adversary $\mathcal{A}^{out}$ and no information about the target's presence in the sample for adversary $\mathcal{A}^{nk}$.

We then define the prior distributions $\hat{\pi}^{in}, \hat{\pi}^{out}, \hat{\pi}^{nk} \in \mathbb{D}\mathcal{X}$ to formally model the prior knowledge of Adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$and $\mathcal{A}^{nk}$, respectively.

Definition 3.3.3 (Prior distributions for adversaries in $\mathcal{G}^d$)**.** *Let $\mathcal{X}$ be the set of secrets. The adversary knows that the population and sample's size are n and $1 \leq m \leq n$, respectively. She also assumes that every dataset is equally probable to be the real population. We then define the prior distributions $\hat{\pi}^{in}, \hat{\pi}^{out}, \hat{\pi}^{nk}$ for three different adversaries with distinct information about the target's presence in the sample:*

*(i) Adversary $\mathcal{A}^{in}$ that knows the target is **in** the sample:*

$$\hat{\pi}^{in}_{(p,t)} = \begin{cases} \frac{1}{m2^n} & , \text{ if } 1 \leq t \leq m \\ 0 & , \text{ otherwise.} \end{cases} \tag{3.5}$$

*(ii) Adversary $\mathcal{A}^{out}$ that knows the target is **outside** the sample:*

$$\hat{\pi}^{out}_{(p,t)} = \begin{cases} \frac{1}{(n-m)2^n} & , \text{ if } m < t \leq n \\ 0 & , \text{ otherwise.} \end{cases} \tag{3.6}$$

*(iii) Adversary $\mathcal{A}^{nk}$ whose information about the presence of the target in the sample is **not known**:*

$$\hat{\pi}^{nk}_{(p,t)} = \frac{1}{n2^n}. \tag{3.7}$$

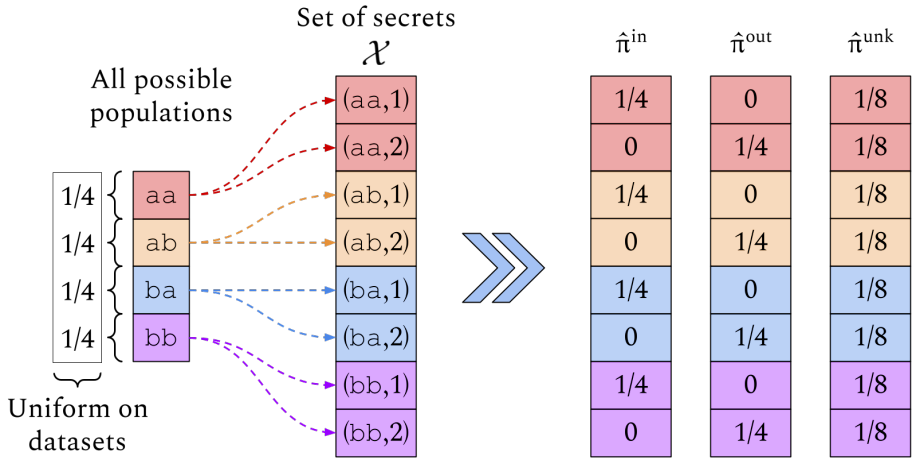
Example 3.3 (Running example – prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$)**.** *Consider the scenario of Example 3.1 where $n = 2$ and $m = 1$. Figure 3.3 resumes this example.*

For adversary $\mathcal{A}^{in}$, she knows the target is in the sample, then she gives probability zero to all $(p,t) \in \{(\texttt{aa}, 2), (\texttt{ab}, 2), (\texttt{ba}, 2), (\texttt{bb}, 2)\}$. And for all other secrets (p,t) such that $t = 1$, $Pr[(p,t)] = 1/4$.

Adversary $\mathcal{A}^{out}$ is the opposite of $\mathcal{A}^{in}$, i.e., $Pr[(p,t)] = 1/4$ for all (p,t) such that $t = 2$ and probability zero for all secrets (p,t) such that $t = 1$, because she knows the target is outside the sample.

The third adversary $\mathcal{A}^{nk}$ that does not know anything about the target's presence in the sample, we have just a uniform probability distribution on all secrets, i.e., $Pr[(p,t)] = \frac{1}{n2^n}$ for all (p,t).

Figure 3.3: Example of prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ for a population of size $n = 2$ and a sample of size $m = 1$. The secret $(\texttt{ab}, 1)$, for instance, represents a population with 2 individuals where the first has value `a`, the second has value `b` and the adversary's target position in the population array is 1.



Font: Elaborated by the author.

In the next section we define precisely the adversary actions as a gain function and provide a formal definition of a channel that represents the process of a sample release.

3.3.3 Adversary's actions and system channel

Once we have modeled the adversary's prior knowledge, we can now reason about the actions she is allowed to take (or the attack she is allowed to execute), and the concerns about information leakage present in the sample's publication. As exposed in the beginning of this chapter, the adversary studied in this work executes an attribute inference attack. She has a predefined single target from the population.

The adversary can guess that the target's value is either `a` or `b`, and we say that she wins 1 if she guesses correctly and 0 otherwise. Using g-vulnerability framework the gain function g formally defined next models this attack.

Definition 3.3.4 (Gain function g - Attribute Inference Attack)**.** *Let $\mathcal{X}$ be the set of secrets. The adversary wants to infer the target's attribute value, so the set of possible guesses is $\mathcal{W} = \{\mathtt{a}, \mathtt{b}\}$. The gain function $g : \mathcal{W} \times \mathcal{X} \to \{0, 1\}$ is defined as*

$$g(w, (p, t)) = \begin{cases} 1 & , \text{ if } p_t = w \\ 0 & , \text{ otherwise,} \end{cases} \tag{3.8}$$

where the condition $p_t = w$ means the target's attribute value is equal to the adversary's guess w.

Example 3.4 (Running example – gain function g)**.** *Backing to Example 3.1 and following Definition 3.3.4, the gain function g that represents the adversary's actions and their corresponding gains are represented in Table 3.2.*

Example 3.5 (Running example – Prior vulnerability)**.** *Considering the same scenario of Example 3.1 with $n = 2$ and supposing that $m = 1$ and adversaries from $\mathcal{G}^f$, we can calculate the adversary's expected probability of guessing correctly the target's attribute value as following.*

$g(w,(p,t))$	$(\texttt{aa},1)$	$(\texttt{aa},2)$	$(\texttt{ab},1)$	$(\texttt{ab},2)$	$(\texttt{ba},1)$	$(\texttt{ba},2)$	$(\texttt{bb},1)$	$(\texttt{bb},2)$
a	1	1	1	0	0	1	0	0
b	0	0	0	1	1	0	1	1

Table 3.2: Gain function g from Example 3.1 that represents the adversary's gains when she guesses a or b for his target's attribute value. Remember that the second element of a secret (p,t) corresponds to the target's index, so for instance, $g(\texttt{a},(\texttt{aa},1))=1$ because the adversary is guessing that his target has the value a and, as the target is in the first position of the dataset aa, her guess was correct.

- *For adversary $\mathcal{A}^{in}$:*

$$
\begin{aligned}
V_g(\pi^{in}) &= \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \pi^{in}_{(p,t)}\cdot g(w,(p,t)) \\
&= \max\left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_1=\texttt{a},\\ t=1}} \frac{1}{m(n+1)\binom{n}{n_{\texttt{a}}(p)}}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_1=\texttt{b},\\ t=1}} \frac{1}{m(n+1)\binom{n}{n_{\texttt{a}}(p)}} \right\} \\
&= \frac{1}{3}\cdot\max\left\{ \underbrace{\frac{1}{\binom{2}{2}}}_{(p,t)=(\texttt{aa},1)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\texttt{ab},1)}, \underbrace{\frac{1}{\binom{2}{0}}}_{(p,t)=(\texttt{bb},1)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\texttt{ba},1)} \right\} \\
&= \frac{1}{3}\cdot\frac{3}{2} \\
&= \frac{1}{2}\ .
\end{aligned}
$$

- *For adversary $\mathcal{A}^{out}$:*

$$
\begin{aligned}
V_g(\pi^{out}) &= \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \pi^{out}_{(p,t)}\cdot g(w,(p,t)) \\
&= \max\left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_1=\texttt{a},\\ t=2}} \frac{1}{(n-m)(n+1)\binom{n}{n_{\texttt{a}}(p)}}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_1=\texttt{b},\\ t=2}} \frac{1}{(n-m)(n+1)\binom{n}{n_{\texttt{a}}(p)}} \right\} \\
&= \frac{1}{3}\cdot\max\left\{ \underbrace{\frac{1}{\binom{2}{2}}}_{(p,t)=(\texttt{aa},2)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\texttt{ab},2)}, \underbrace{\frac{1}{\binom{2}{0}}}_{(p,t)=(\texttt{bb},2)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\texttt{ba},2)} \right\} \\
&= \frac{1}{3}\cdot\frac{3}{2} \\
&= \frac{1}{2}\ .
\end{aligned}
$$

- *For adversary $\mathcal{A}^{nk}$:*

$$
\begin{aligned}
V_g(\pi^{nk}) &= \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \pi^{nk}_{(p,t)} \cdot g(w,(p,t)) \\
&= \max \left\{ \sum_{\substack{(p,t) \in \mathcal{X}: \\ p_1 = \mathtt{a}}} \frac{1}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}}, \sum_{\substack{(p,t) \in \mathcal{X}: \\ p_1 = \mathtt{b}}} \frac{1}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}} \right\} \\
&= \frac{1}{6} \cdot \max \Bigg\{ \underbrace{2 \cdot \frac{1}{\binom{2}{2}}}_{(p,t) \in \{(\mathtt{aa},1),(\mathtt{aa},2)\}} + \underbrace{2 \cdot \frac{1}{\binom{2}{1}}}_{(p,t) \in \{(\mathtt{ab},1),(\mathtt{ab},2)\}}, \\
&\qquad \underbrace{2 \cdot \frac{1}{\binom{2}{0}}}_{(p,t) \in \{(\mathtt{bb},1),(\mathtt{bb},2)\}} + \underbrace{2 \cdot \frac{1}{\binom{2}{1}}}_{(p,t) \in \{(\mathtt{ba},1),(\mathtt{ba},2)\}} \Bigg\} \\
&= \frac{1}{6} \cdot 3 \\
&= \frac{1}{2} \ .
\end{aligned}
$$

The prior vulnerabilities in the three cases showed in Example 3.5 corresponds to the probability of the adversary guessing correctly the target's attribute value before observing anything (e.g., the sample's histogram), so she takes "a shot in the dark", and the probability of success is $1/2$.

In QIF the process of publishing a sample's histogram can be modeled as a channel S that maps a set of possible population arrays to a set of possible samples. We formally define this channel next.

Definition 3.3.5 (Channel S). *Given a set of secrets $\mathcal{X}$ we can model a sample publication as a channel $\mathsf{S} : \mathcal{X} \to \mathbb{D}\mathcal{Y}$ that is a mapping from secrets to distributions on the the set of possible outputs $\mathcal{Y}$ (i.e., the set of all possible histograms of m people). Assuming that the order of people in the population array does not matter (i.e., any order is equally probable), we can fix the sample to be always the first m people in the population p, i.e., the sample will be just $p_{1\ldots m}$. Formally,*

$$
\mathsf{S}_{(p,t),y} = \begin{cases} 1 & , \text{ if } n_{\mathtt{a}}(p_{1\ldots m}) = y \\ 0 & , \text{ otherwise,} \end{cases} \tag{3.9}
$$

where $y \in \mathcal{Y}$ represents a histogram of a sample of size m where y people have the value $\mathtt{a}$ and $m - y$ have the value $\mathtt{b}$. The number of $\mathtt{a}$'s in the first m people of population p is $n_{\mathtt{a}}(p_{1\ldots m})$. The entry $\mathsf{S}_{(p,t),y}$ can be understood as the probability of the published histogram being y when the population is p, i.e., $Pr[y|(p,t)]$.

Example 3.6 (Running example – Channel S)**.** *Following Example 3.1 where $n=2$ and $m=1$ the channel matrix S that represents the sample publication in this scenario would be as in Table 3.3.´*

$\mathsf{S}_{(p,t),y}$	0	1
(aa,1)	0	1
(aa,2)	0	1
(ab,1)	0	1
(ab,2)	0	1
(ba,1)	1	0
(ba,2)	1	0
(bb,1)	1	0
(bb,2)	1	0

Table 3.3: Channel matrix S for $n=2$ and $m=1$. For instance, we have that $\mathsf{S}_{(\mathtt{aa},1),0}=0$ because the sample is just p_1, and as the first person in secret $(\mathtt{aa},1)$ has an $\mathtt{a}$, the sample's histogram will be "1", thus $Pr[y{=}0 \mid (p,t){=}(\mathtt{aa},1)] = 0$. And oppositely, $\mathsf{S}_{(\mathtt{aa},1),1} = Pr[y{=}1 \mid (p,t){=}(\mathtt{aa},1)] = 1$.

Example 3.7 (Running example – Posterior vulnerability)**.** *Following Example 3.1 where $n=2$ and supposing that $m=1$, adversaries from $\mathcal{G}^f$, a gain function g for attribute inference attack and channel S, we can calculate the posterior vulnerability as following.*

- *For adversary $\mathcal{A}^{in}$:*

$$V_g[\pi^{in} \rhd \mathsf{S}] = \sum_{y\in\mathcal{Y}} \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \pi^{in}_{(p,t)}\, \mathsf{S}_{(p,t),y}\; g(w,(p,t))$$

$$= \sum_{y=0}^{1} \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ t=1,\\ p_1=a,\\ n_a(p_1)=y}} \frac{1}{m(n+1)\binom{n}{n_a(p)}}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ t=1,\\ p_1=b,\\ n_a(p_1)=y}} \frac{1}{m(n+1)\binom{n}{n_a(p)}} \right\}$$

$$= \underbrace{\frac{1}{3}\cdot \max\left\{0, \underbrace{\frac{1}{\binom{2}{0}}}_{(p,t)=(\mathtt{bb},1)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ba},1)}\right\}}_{y=0} + \underbrace{\frac{1}{3}\cdot \max\left\{\underbrace{\frac{1}{\binom{2}{2}}}_{(p,t)=(\mathtt{aa},1)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ab},1)}, 0\right\}}_{y=1}$$

$$= 1\ .$$

- *For adversary $\mathcal{A}^{out}$:*

$$V_g[\pi^{out} \rhd \mathsf{S}] = \sum_{y \in \mathcal{Y}} \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \pi^{out}_{(p,t)}\, \mathsf{S}_{(p,t),y}\; g(w,(p,t))$$

$$= \sum_{y=0}^{1} \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ t=2,\\ p_2=\mathtt{a},\\ n_{\mathtt{a}}(p_1)=y}} \frac{1}{(n-m)(n+1)\binom{n}{n_{\mathtt{a}}(p)}}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ t=2,\\ p_2=\mathtt{b},\\ n_{\mathtt{a}}(p_1)=y}} \frac{1}{(n-m)(n+1)\binom{n}{n_{\mathtt{a}}(p)}} \right\}$$

$$= \underbrace{\frac{1}{3} \cdot \max \left\{ \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ba},2)}, \underbrace{\frac{1}{\binom{2}{0}}}_{(p,t)=(\mathtt{bb},2)} \right\}}_{y=0} + \underbrace{\frac{1}{3} \cdot \max \left\{ \underbrace{\frac{1}{\binom{2}{2}}}_{(p,t)=(\mathtt{aa},2)}, \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ab},2)} \right\}}_{y=1}$$

$$= \frac{2}{3}\ .$$

- *For adversary $\mathcal{A}^{nk}$:*

$$V_g[\pi^{nk} \rhd \mathsf{S}] = \sum_{y \in \mathcal{Y}} \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \pi^{nk}_{(p,t)}\, \mathsf{S}_{(p,t),y}\; g(w,(p,t))$$

$$= \sum_{y=0}^{1} \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_t=\mathtt{a},\\ n_{\mathtt{a}}(p_1)=y}} \frac{1}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_t=\mathtt{b},\\ n_{\mathtt{a}}(p_1)=y}} \frac{1}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}} \right\}$$

$$= \underbrace{\frac{1}{6} \cdot \max \left\{ \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ba},2)}, \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ba},1)} + \underbrace{\frac{1}{\binom{2}{0}}}_{(p,t)=(\mathtt{bb},1)} + \underbrace{\frac{1}{\binom{2}{0}}}_{(p,t)=(\mathtt{bb},2)} \right\}}_{y=0} +$$

$$\underbrace{\frac{1}{6} \cdot \max \left\{ \underbrace{\frac{1}{\binom{2}{2}}}_{(p,t)=(\mathtt{aa},1)} + \underbrace{\frac{1}{\binom{2}{2}}}_{(p,t)=(\mathtt{aa},2)} + \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ab},1)}, \underbrace{\frac{1}{\binom{2}{1}}}_{(p,t)=(\mathtt{ab},2)} \right\}}_{y=1}$$

$$= \frac{5}{6}\ .$$

The operational interpretation of posterior vulnerability in this case can be described as "*The expected probability of the adversary guessing correctly the target's attribute value after she observed the sample's histogram*". The posterior vulnerability in Example 3.7

for adversary $\mathcal{A}^{in}$ comes from the fact that the sample contains only 1 person, and as adversaries $\mathcal{A}^{in}$ knows the target is in the sample, then the probability of she guessing the target's value correctly is 1.

In the case of adversary $\mathcal{A}^{out}$, when she observes y=0, the possible secrets are in $\{(\texttt{ba}, 2), (\texttt{bb}, 2)\}$, and as $Pr[p = \texttt{bb}] = 1/3 > Pr[p = \texttt{ba}] = 1/6$ and she knows the target is outside the sample, her best guess is b, and her probability of success is then $1/3$. The same reasoning can be applied when y=1. Thus, the final probability of success will be $Pr[p_2 = \texttt{b} \mid p = \texttt{bb}] + Pr[p_2 = \texttt{a} \mid p = \texttt{aa}] = 1/3 + 1/3 = 2/3$.

Finally, for adversary $\mathcal{A}^{nk}$, the posterior vulnerability $5/6$ is the average of the posteriors of adversaries $\mathcal{A}^{in}$ and $\mathcal{A}^{out}$ weighted by the probability of the target being in the sample (which is m/n, as each participant's index in the population array is randomly selected, as pointed in Section 3.1) or not. Indeed we prove this equivalence in Theorem 4.1.3 (iii).

In the next section we present definitions that allow us to make analyses about the sample's publication utility by modeling an adversary that is trying to infer the distribution on attribute values in the population.

3.4 Adversary model for utility analysis

The analysis of privacy in sample publications only exists because there is some data about a population being publicly released, and we can name the purpose of this publication as its utility. In this work we are focusing in the sample publication of a population, in particular, a single binary attribute. One direct utility this kind of publication may have is to make data analysts able to answer the question *"What is the frequency of value* a *in the population?"*. In this section we provide formal definitions for the scenario of utility analysis in sample publications, what includes definitions about the data analyst adversary $\mathcal{A}^{ut}$.

Recall that a secret (p, t), as per Definition 3.3.1, is a pair where p is the population array and t is the target's index. For adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ described in Section 3.3 (that discuss privacy concerns) we considered they all have a single target and they want to infer their target's value. On the other hand, thinking about utility, an adversary $\mathcal{A}^{ut}$ that represents a data analyst trying to infer the distribution of the attribute in the population does not have an individual as a target. In this way we define the set of secrets $\mathcal{X}^{ut}$ that takes into account only the set of all possible populations.

Definition 3.4.1 (Set of secrets $\mathcal{X}^{ut}$)**.** *Let $n \geq 1$ be the population size and consider a binary attribute of interest with values in $\{\mathtt{a}, \mathtt{b}\}$. The set of secrets $\mathcal{X}^{ut}$ of all possible populations is defined as*

$$\mathcal{X}^{ut} = \{\mathtt{a}, \mathtt{b}\}^n, \tag{3.10}$$

where a secret $p \in \mathcal{X}^{ut}$ is a binary array of size n, and p_i is the attribute value of the i-th person in the array.

Example 3.8 (Running example – Set of secrets $\mathcal{X}^{ut}$)**.** *Consider a scenario in which there is a population with $n = 2$ individuals and a binary attribute of interest with possible value in $\{\mathtt{a}, \mathtt{b}\}$. The set of secrets $\mathcal{X}^{ut}$ that is the set of all possible populations is*

$$\mathcal{X}^{ut} = \{\mathtt{aa}, \mathtt{ab}, \mathtt{ba}, \mathtt{bb}\} \ .$$

In the next two sections we define formally the adversary's prior knowledge for the two groups of adversaries that make different assumptions about the population dataset (Section 3.2).

3.4.1 Group of adversaries $\mathcal{G}^f$

Similar to the prior distributions defined for the three adversaries in privacy analysis, here we assume the adversary knows the population and sample's sizes are n and m, respectively, and the assumption about the uniform distribution on possible frequencies of value $\mathtt{a}$ in the population. Therefore the prior distribution π^{ut} that represents the prior knowledge of the data analyst adversary $\mathcal{A}^{ut}$ is defined as following.

Definition 3.4.2 (Prior distribution π^{ut})**.** *Given the set of secrets $\mathcal{X}^{ut}$, the prior knowledge of the adversary that represents a data analyst trying to infer the frequency of value $\mathtt{a}$ in the population is the prior distribution π^{ut}, defined as*

$$\pi_p^{ut} = \frac{1}{(n+1)\binom{n}{n_{\mathtt{a}}(p)}} \ . \tag{3.11}$$

Example 3.9 (Running example – Prior distribution π^{ut})**.** *Following Example 3.8 where the population size is $n=2$, the prior distribution π^{ut} would assume the following probabilities:*

$$\pi^{ut}_{aa} = \pi^{ut}_{bb} = 1/3,$$
$$\pi^{ut}_{ab} = \pi^{ut}_{ba} = 1/6.$$

The probabilities showed in Example 3.9 comes directly from the adversary's assumption about the uniform distribution on possible frequencies. See Figure 3.2 fore more details.

3.4.2 Group of adversaries $\mathcal{G}^d$

Considering the assumption about adversaries in $\mathcal{G}^d$, in order to make an analysis about utility in a sample publication, we define the data analyst's prior distribution $\hat{\pi}^{ut}$ as following.

Definition 3.4.3 (Prior distribution $\hat{\pi}^{ut}$)**.** *Given the set of secrets $\mathcal{X}^{ut}$, the prior distribution $\hat{\pi}^{ut}$ that represents the data analyst adversary that assumes a uniform distribution on datasets is defined as follows. For all $p \in \mathcal{X}^{ut}$:*

$$\hat{\pi}^{ut}_p := \frac{1}{2^n} \ .$$

In the next section we define precisely the adversary guesses about the distribution on attribute values in the population and we propose a loss function for measuring the adversary's success. Also, as we are using a new set of secrets $\mathcal{X}^{ut}$ (defined in Section 3.4) for utility analysis, we also define a new channel S^{ut} for representing the sample release in this case.

3.4.3 Adversary's actions and system channel

We define the success of $\mathcal{A}^{ut}$ looking at her uncertainty about the proportion of $\mathtt{a}$'s in the population. More specifically, we calculate how far her guess about the frequency of

a's in the population is from the real frequency. Formally, we define the following loss function ℓ.

Definition 3.4.4 (Loss function ℓ – Utility). *Let $\mathcal{X}^{ut}$ be the set of secrets and $\mathcal{W} = \{0/n, 1/n, \ldots, n/n\}$ be the set of actions where $w \in \mathcal{W}$ represents the adversary guessing that w per cent of the population has the value $\mathtt{a}$. The loss function $\ell : \mathcal{W} \times \mathcal{X}^{ut} \to [0,1]$ that indicates the distance between the adversary's guess and the real frequency in the population is*

$$\ell(w,p) = \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right|. \tag{3.12}$$

Looking at Equation (3.12) it is possible to see that the further the adversary's guess is from the real frequency, the more the adversary "losses". Thus when the adversary's guess is the exact real frequency in the population, the loss will be 0. The loss will be 1 when she guesses that nobody in the population has a's, and actually everyone has, or the opposite, when she guesses that everyone has a's and actually everyone has b's.

Example 3.10 (Running example – Loss function ℓ). *Recall Example 3.8 where the population size is $n = 2$. The adversary $\mathcal{A}^{ut}$ can guess that there are 0, 1 or 2 people in the population with value $\mathtt{a}$. Thus the loss function for this scenario is defined as in Table 3.4.*

$\ell(w,p)$	aa	ab	ba	bb
$0/2$	1	$1/2$	$1/2$	0
$1/2$	$1/2$	0	0	$1/2$
$2/2$	0	$1/2$	$1/2$	1

Table 3.4: Loss function ℓ from Example 3.10 that represents the adversary's error when guessing the frequency of value a in the population.

For scenarios related to privacy concerns (adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$and $\mathcal{A}^{nk}$), as discussed in Section 3.3.3, the sample publication can be modeled as a channel S (Definition 3.3.5). In a similar way, for the data analyst adversary $\mathcal{A}^{ut}$ that we have just introduced in the beginning of this chapter, in order to adapt the definitions for the new set of secrets $\mathcal{X}^{ut}$, we define a new channel S^{ut} that maps population datasets to distributions on sample histograms.

Definition 3.4.5 (Channel S^{ut}). *Given the set of secrets $\mathcal{X}^{ut}$, a sample publication can be modeled as a channel $\mathsf{S}^{ut} : \mathcal{X}^{ut} \to \mathbb{D}\mathcal{Y}$ is the set of possible outputs (i.e., the set of all possible histograms of m people). Assuming that the order of people in the population array does not matter (i.e., any order is equally probable), we can fix the sample to be always the first m people in the population p, i.e., the sample will be just $p_{1...m}$. Formally,*

$$\mathsf{S}^{ut}_{p,y} = \begin{cases} 1 & , \text{ if } n_{\mathtt{a}}(p_{1...m}) = y \\ 0 & , \text{ otherwise,} \end{cases} \tag{3.13}$$

where $y \in \mathcal{Y}$ represents a histogram of a sample of size m where y people have the value $\mathtt{a}$ and $m - y$ have the value $\mathtt{b}$, and $n_{\mathtt{a}}(p_{1...m})$ is the number of $\mathtt{a}$'s in the first m people of population x. The entry $\mathsf{S}^{ut}_{p,y}$ can be understood as the probability of the published histogram being y when the population is p, i.e., $Pr[y|p]$.

Example 3.11 (Running example – Channel S^{ut}). *Recall the scenario of Example 3.8 where $n = 2$ and consider that the sample size is $m = 1$. The channel S^{ut} that represents the sample publication will be as in Table 3.5.*

$\mathsf{S}^{ut}_{p,y}$	0	1
aa	0	1
aa	0	1
ab	0	1
ab	0	1
ba	1	0
ba	1	0
bb	1	0
bb	1	0

Table 3.5: Channel matrix S^{ut} for $n = 2$ and $m = 1$. For instance, we have that $\mathsf{S}^{ut}_{\mathtt{aa},0} = 0$ because the sample is just p_1, and as the first person in secret aa has an a, the sample's histogram will be "1", thus $Pr[y{=}0 \mid p{=}\mathtt{aa}] = 0$. And oppositely, $\mathsf{S}^{ut}_{\mathtt{aa},1} = Pr[y{=}1 \mid p{=}\mathtt{aa}] = 1$.

Besides the models provided in this chapter for adversaries related to privacy and utility concerns, we present in the next chapter contributions on equations for prior and posterior vulnerabilities for those scenarios. In particular, we provide closed formulas for most vulnerability equations. Those formulas are relevant and necessary specially in scenarios with large datasets and/or samples in which the calculation of vulnerabilities are computationally prohibitive by the fact that the number of secrets is exponential on the population and sample's sizes (see Sections 3.3 and 3.4 for more details).

Chapter 4

Vulnerabilities and uncertainties of publishing a sample

In this chapter we present the contributions related to prior and posterior vulnerabilities and utility losses for the models described in the last chapter. Those contributions include closed formulas which are very useful when quantifying vulnerabilities and utility losses in large datasets/samples. All definitions from the framework of QIF used in this chapter are detailed in Section 2.2.

We first show in Section 4.1 equations that describe the success of attribute inference attack executed by adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ both before and after observing the sample release. Next, in Section 4.2, we present equations for the data analyst adversary $\mathcal{A}^{ut}$ and in the last section we discuss interpretation of those equations as well as the insights and conclusions we are able to make using them.

4.1 Attribute inference attack

We divide the presentation of results for attribute inference attack in two sections. The first, Section 4.1.1, shows formulas for prior vulnerability and the second one, Section 4.1.2, formulas for posterior vulnerability. All definitions related to this attack are described in Section 3.3.

4.1.1 Results on prior vulnerability

We start by showing the results for prior vulnerabilities for adversaries in $\mathcal{G}^f$. Theorem 4.1.1 states that the prior vulnerability is $1/2$ when the prior knowledge is π^{in},

π^{out} or π^{nk}. Recall that the operational interpretation of this result can be described as "*The expected probability of the adversary guessing correctly the target's attribute value before observing the sample is* $1/2$".

Theorem 4.1.1 (Prior vulnerability – adversaries in $\mathcal{G}^f$)**.** *Given the prior distributions π^{in}, π^{out} and π^{nk} on the set of secrets $\mathcal{X}$, and the gain function g for attribute inference attack, the prior vulnerability, i.e., the expected probability of adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$, respectively, inferring the target's attribute value, is*

$$V_g(\pi^{in}) = V_g(\pi^{out}) = V_g(\pi^{nk}) = 1/2. \tag{4.1}$$

Proof.

Adversary $\mathcal{A}^{in}$ and prior distribution π^{in}:

$$V_g(\pi^{in}) = \max_{w \in \mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \pi^{in}_{(p,t)} \cdot g(w,(p,t))$$

Def. of π^{in} and g:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t = w}} \frac{1}{m(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$

Split cases when w=a and w=b:

$$= \frac{1}{m(n+1)} \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t = \mathtt{a}}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t = \mathtt{b}}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} \right\}$$

We need to define how many secrets $x \in \mathcal{X}$ satisfy the restrictions $1 \leq t \leq m \wedge p_t = \mathtt{a}$ (in the left summation inside the max) and $1 \leq t \leq m \wedge p_t = \mathtt{b}$ (in the right summation inside the max). The number of secrets that satisfy these restrictions is $\sum_{t=1}^{m} \sum_{i=0}^{n-1} \binom{n-1}{i}$. The first summation on t goes over all possible indexes for the target. Once p_t is fixed, the other $n-1$ positions in the population array p can be any combination, and i is the number of a's in $p \backslash p_t$, i.e., $i = n_{\mathtt{a}}(p_{1\ldots t-1,t+1,\ldots n})$. Finally, $p_t = \mathtt{a}$ implies $\binom{n}{n_{\mathtt{a}}(p)}^{-1} = \binom{n}{i+1}^{-1}$, and $p_t = \mathtt{b}$ implies $\binom{n}{n_{\mathtt{a}}(p)}^{-1} = \binom{n}{i}^{-1}$.

$$= \frac{1}{m(n+1)} \max \left\{ \sum_{t=1}^{m} \sum_{i=0}^{n-1} \binom{n-1}{i} \binom{n}{i+1}^{-1}, \sum_{t=1}^{m} \sum_{i=0}^{n-1} \binom{n-1}{i} \binom{n}{i}^{-1} \right\} \tag{4.2}$$

$$
\begin{aligned}
&= \frac{1}{m(n+1)} \max\left\{ m\sum_{i=0}^{n-1}\frac{i+1}{n}, m\sum_{i=0}^{n-1}\frac{n-i}{n} \right\} \\
&= \frac{1}{n+1} \max\left\{ \sum_{i=0}^{n-1}\frac{i+1}{n}, \sum_{i=0}^{n-1}\frac{n-i}{n} \right\} \\
&= \frac{1}{n+1} \max\left\{ \frac{1}{n}\left(\sum_{i=0}^{n-1} i + \sum_{i=0}^{n-1} 1\right), \frac{1}{n}\left(\sum_{i=0}^{n-1} n - \sum_{i=0}^{n-1} i\right) \right\} \\
&= \frac{1}{n+1} \max\left\{ \frac{1}{n}\left(\frac{(n-1)n}{2} + n\right), \frac{1}{n}\left(n^2 - \frac{(n-1)n}{2}\right) \right\} \\
&= \frac{1}{n+1} \max\left\{ \frac{n-1}{2} + 1, n - \frac{n-1}{2} \right\} \\
&= \frac{1}{n+1} \max\left\{ \frac{n+1}{2}, \frac{n+1}{2} \right\} \\
&= \frac{1}{n+1} \cdot \frac{n+1}{2} \\
&= \frac{1}{2}.
\end{aligned}
\tag{4.3}
$$

Adversary $\mathcal{A}^{out}$ and prior distribution π^{out}:

$$V_g(\pi^{out}) = \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \pi^{out}_{(p,t)} \cdot g(w,(p,t))$$

Def. of π^{out} and g:

$$= \max_{w\in\mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=w}} \frac{1}{(n-m)(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$

Split cases when w=a and w=b:

$$= \frac{1}{(n-m)(n+1)} \max\left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=\mathtt{a}}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=\mathtt{b}}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} \right\}$$

Here the reasoning is the same as in Equation (4.2), except that now the target's index can be any value between m+1 and n.

$$= \frac{1}{(n-m)(n+1)} \max\left\{ \sum_{t=m+1}^{n}\sum_{i=0}^{n-1} \binom{n-1}{i}\binom{n}{i+1}^{-1}, \sum_{t=m+1}^{n}\sum_{i=0}^{n-1} \binom{n-1}{i}\binom{n}{i}^{-1} \right\}$$

$$= \frac{1}{(n-m)(n+1)} \max\left\{(n-m)\sum_{i=0}^{n-1}\frac{i+1}{n}, (n-m)\sum_{i=0}^{n-1}\frac{n-i}{n}\right\}$$
$$= \frac{1}{n+1} \max\left\{\sum_{i=0}^{n-1}\frac{i+1}{n}, \sum_{i=0}^{n-1}\frac{n-i}{n}\right\}. \tag{4.4}$$

Equation (4.4) is the same as Equation (4.3) that has already been proven to be 1/2.

Adversary $\mathcal{A}^{nk}$ and prior distribution π^{nk}:

$$V_g(\pi^{nk}) = \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \pi^{nk}_{(p,t)} \cdot g(w,(p,t))$$

Def. of π^{nk} and g:

$$= \max_{w\in\mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}\\ p_t=w}} \frac{1}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$

Split cases when w=a and w=b:

$$= \frac{1}{n(n+1)} \max\left\{\sum_{\substack{(p,t)\in\mathcal{X}\\ p_t=\mathtt{a}}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}, \sum_{\substack{(p,t)\in\mathcal{X}\\ p_t=\mathtt{b}}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}\right\}$$

Here the reasoning is the same as in Equation (4.2), except that now the target's index can be any value between 1 and n.

$$= \frac{1}{n(n+1)} \max\left\{\sum_{t=1}^{n}\sum_{i=0}^{n-1}\binom{n-1}{i}\binom{n}{i+1}^{-1}, \sum_{t=1}^{n}\sum_{i=0}^{n-1}\binom{n-1}{i}\binom{n}{i}^{-1}\right\}$$
$$= \frac{1}{n(n+1)} \max\left\{n\sum_{i=0}^{n-1}\frac{i+1}{n}, n\sum_{i=0}^{n-1}\frac{n-i}{n}\right\}$$
$$= \frac{1}{n+1} \max\left\{\sum_{i=0}^{n-1}\frac{i+1}{n}, \sum_{i=0}^{n-1}\frac{n-i}{n}\right\}. \tag{4.5}$$

Equation (4.5) is the same as Equation (4.3), that has already been proven to be 1/2.

Thus, we have shown that $V_g(\pi^{in}) = V_g(\pi^{out}) = V_g(\pi^{nk}) = 1/2$. □

Now we analyze the prior vulnerability for adversaries in $\mathcal{G}^d$, whose prior knowledge is formalized by prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$. This result is showed in Theorem 4.1.2.

Theorem 4.1.2 (Prior vulnerability – adversaries in $\mathcal{G}^d$)**.** *Given the prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ on the set of secrets $\mathcal{X}$, and the gain function g for attribute inference attack, the prior vulnerability, i.e., the expected probability of adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$, respectively, inferring the target's attribute value, is*

$$V_g(\hat{\pi}^{in}) = V_g(\hat{\pi}^{out}) = V_g(\hat{\pi}^{nk}) = 1/2. \tag{4.6}$$

Proof.

Adversary $\mathcal{A}^{in}$ and prior distribution $\hat{\pi}^{in}$:

$$V_g(\hat{\pi}^{in}) = \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \hat{\pi}^{in}_{(p,t)} \cdot g(w, (p,t))$$

Def. of $\hat{\pi}^{in}$ and g:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = w}} \frac{1}{m 2^n}$$

Split cases when w=a and w=b:

$$= \frac{1}{m 2^n} \max \left\{ \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = \mathtt{a}}} 1, \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = \mathtt{b}}} 1 \right\}$$

We need to define how many secrets $p \in \mathcal{X}^{ut}$ satisfy the restrictions $1 \leq t \leq m \ \wedge \ p_t = \mathtt{a}$ (in the left summation inside the max) and $1 \leq t \leq m \ \wedge \ p_t = \mathtt{b}$ (in the right summation inside the max). For both cases the number of secrets is $\sum_{t=1}^{m} 2^{n-1}$. The summation goes over all possible indexes for the adversary's target and 2^{n-1} is the number of all combinations of values people in indexes from 2 to n can assume.

$$\begin{aligned} &= \frac{1}{m 2^n} \max \left\{ \sum_{t=1}^{m} 2^{n-1}, \sum_{t=1}^{m} 2^{n-1} \right\} \\ &= \frac{1}{m 2^n} \cdot m 2^{n-1} \\ &= \frac{1}{2} \ . \end{aligned} \tag{4.7}$$

Adversary $\mathcal{A}^{out}$ and prior distribution $\hat{\pi}^{out}$:

$$V_g(\hat{\pi}^{out}) = \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \hat{\pi}^{out}_{(p,t)} \cdot g(w, (p,t))$$

Def. of $\hat{\pi}^{out}$ and g:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=w}} \frac{1}{(n-m)2^n}$$

Split cases when w=a and w=b:

$$= \frac{1}{(n-m)2^n} \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=\mathtt{a}}} 1, \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=\mathtt{b}}} 1 \right\}$$

Here the reasoning is the same as in Equation (4.7), except that now the target's index can be any value between m+1 and n.

$$\begin{aligned} &= \frac{1}{(n-m)2^n} \max \left\{ \sum_{t=m+1}^{n} 2^{n-1}, \sum_{t=m+1}^{n} 2^{n-1} \right\} \\ &= \frac{1}{(n-m)2^n} \cdot (n-m)2^{n-1} \\ &= \frac{1}{2} \ . \end{aligned}$$

Adversary $\mathcal{A}^{nk}$ and prior distribution $\hat{\pi}^{nk}$:

$$V_g(\hat{\pi}^{nk}) = \max_{w \in \mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \hat{\pi}^{nk}_{(p,t)} \cdot g(w,(p,t))$$

Def. of $\hat{\pi}^{nk}$ and g:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=w}} \frac{1}{n2^n}$$

Split cases when w=a and w=b:

$$= \frac{1}{n2^n} \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq n,\\ p_t=\mathtt{a}}} 1, \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq n,\\ p_t=\mathtt{b}}} 1 \right\}$$

Here the reasoning is the same as in Equation (4.7), except that now the target's index can be any value between 1 and n.

$$\begin{aligned} &= \frac{1}{n2^n} \max \left\{ \sum_{t=1}^{n} 2^{n-1}, \sum_{t=1}^{n} 2^{n-1} \right\} \\ &= \frac{1}{n2^n} \cdot n2^{n-1} = \frac{1}{2} \ . \end{aligned}$$

Thus we have proved that $V_g(\hat{\pi}^{in}) = V_g(\hat{\pi}^{out}) = V_g(\hat{\pi}^{nk}) = 1/2$. □

In the next section we study what happens with the adversary's knowledge when she observes the released sample's histogram. We analyze how this information update the adversary's knowledge, what will be her best guess and her probability of success in guessing the target's attribute value.

4.1.2 Results on posterior vulnerability

We now present some lemmas that reduces summations on binomial coefficients to closed formulas that will be helpful in the proofs for posterior vulnerabilities (Theorems 4.1.3 and 4.1.4). Similar to the last section we start by the posterior vulnerability for adversaries in $\mathcal{G}^f$. The proofs of all lemmas can be found in Appendix A.

Lemma 4.1.1 (Summation on binomials 1)**.** *Let $1 \leq y \leq m \leq n$ be integers. The following equivalence remains:*

$$\sum_{k=0}^{n-m} \binom{m-1}{y-1}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \frac{y(n+1)}{m(m+1)} \tag{4.8}$$

and analogously:

$$\sum_{k=0}^{n-m} \binom{m-1}{y}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \frac{(m-y)(n+1)}{m(m+1)}. \tag{4.9}$$

Lemma 4.1.2 (Ordinary generating function)**.** *The following equivalence remains:*

$$\sum_{i=0}^{\infty} \binom{k+i}{k} x^i = \frac{1}{(1-x)^{k+1}}. \tag{4.10}$$

Proof. The function in the left side of the equality is an Ordinary Generating Function (OGF) of the sequence $1, k+1, \binom{k+2}{2}, \binom{k+3}{3}, \ldots$. The equality above can be easily shown using some operations on OGFs. For more details see Chapter 3 of [27]. □

Lemma 4.1.3 (Summation on binomials 2)**.** *Let $1 \leq y \leq m \leq n$ be integers. The following equivalence remains:*

$$\sum_{k=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{k}\binom{n}{y+k+1}^{-1} = \frac{(n+1)(y+1)}{(m+1)(m+2)}, \tag{4.11}$$

and analogously:

$$\sum_{k=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{k}\binom{n}{y+k}^{-1} = \frac{(n+1)(m-y+1)}{(m+1)(m+2)}. \tag{4.12}$$

Lemma 4.1.4 (Summations on m)**.** *Let $m \geq 1$. We have that*

$$\sum_{i=0}^{\lfloor m/2 \rfloor} m - i + \sum_{i=\lfloor m/2 \rfloor + 1}^{m} i = \binom{m+1}{2} + \left\lfloor \frac{(m+1)^2}{4} \right\rfloor. \tag{4.13}$$

Now we show in Lemma 4.1.5 that, when the prior distribution on secrets is π^{in}, π^{out} or π^{nk}, the probability distribution on outputs (sample histograms) is a uniform distribution.

Lemma 4.1.5 (Marginal on $\mathcal{Y}$ for π^{in}, π^{out} and π^{nk})**.** *Given the prior distributions π^{in}, π^{out} and π^{nk} on the set of secrets $\mathcal{X}$ and the channel S, the probability of a sample's histogram $y \in \mathcal{Y}$ being the output is*

$$Pr[y] = \frac{1}{m+1}. \tag{4.14}$$

Before going through posterior vulnerability, we show in Lemma 4.1.6 the vulnerability of a given output y, i.e., the adversary's probability of correctly guessing the target's output when the published sample histogram is y.

Lemma 4.1.6 (Vulnerability of a specific output y, adversaries in $\mathcal{G}^f$)**.** *Let $\mathcal{X}$ be the set of secrets, π^{in}, π^{out} and π^{nk} be prior distributions on $\mathcal{X}$, g be the gain function for attribute inference attack and S be the channel. Given that the adversary observed some output y, the posterior vulnerability given y is*

(i)

$$V_g(\delta^{in,y}) = \max\left\{\frac{y}{m}, \frac{m-y}{m}\right\}. \tag{4.15}$$

(ii)

$$V_g(\delta^{out,y}) = \max\left\{\frac{y+1}{m+2}, \frac{m-y+1}{m+2}\right\} \tag{4.16}$$

(iii)

$$V_g(\delta^{nk,y}) = \frac{n + \max\{ny + 2y - m, nm - (ny + 2y - m)\}}{n(m+2)} \tag{4.17}$$

where $\delta^{in,y}$, $\delta^{out,y}$ and $\delta^{nk,y}$ are the inner distributions when y is observed and when π^{in}, π^{out} and π^{nk} are, respectively, the prior distributions. These vulnerabilities can be understood as $Pr[X \mid Y = y]$ with X being the set of secrets and Y begin the set of sample histograms.

Finally we present in Theorem 4.1.3 closed formulas for posterior vulnerabilities.

Theorem 4.1.3 (Posterior vulnerability for prior distributions π^{in}, π^{out} and π^{nk})**.** *Given the prior distributions π^{in}, π^{out} and π^{nk}, the gain function g for attribute inference attack and the channel S, the corresponding posterior vulnerabilities are*

(i) for adversary $\mathcal{A}^{in}$:

$$V_g[\pi^{in} \triangleright \mathsf{S}] = \frac{3}{4} + \frac{1}{4(\lfloor \frac{m}{2} \rfloor + \lceil \frac{m+1}{2} \rceil)}, \tag{4.18}$$

(ii) for adversary $\mathcal{A}^{out}$:

$$V_g[\pi^{out} \triangleright \mathsf{S}] = \frac{3}{4} - \frac{1}{4\left(\lfloor \frac{m+1}{2} \rfloor + \lceil \frac{m}{2} \rceil + 1\right)}, \tag{4.19}$$

(iii) for adversary $\mathcal{A}^{nk}$:

$$V_g[\pi^{nk} \triangleright \mathsf{S}] = \frac{m}{n} \cdot V_g[\pi^{in} \triangleright \mathsf{S}] + \frac{n-m}{n} \cdot V_g[\pi^{out} \triangleright \mathsf{S}], \tag{4.20}$$

where m/n is the prior probability of a person being selected to be part of the sample and $(m-n)/n$ is the prior probability of a person not being selected to be in the sample.

We first present the proof for Equation (4.18).

Proof.

$$V_g[\pi^{in} \triangleright \mathsf{S}] = \sum_{y \in \mathcal{Y}} Pr[y] \cdot V_g(\delta^{in,y})$$

Def. of S and by Lemmas 4.1.5 and 4.1.6 (i):

$$\begin{aligned}
&= \sum_{y=0}^{m} \frac{1}{m+1} \cdot \max\left\{\frac{y}{m}, \frac{m-y}{m}\right\} \\
&= \frac{1}{m(m+1)} \sum_{y=0}^{m} \max\{y, m-y\} \\
&= \frac{1}{m(m+1)} \left(\sum_{y=0}^{\lfloor \frac{m}{2} \rfloor} m - y + \sum_{\lfloor \frac{m}{2} \rfloor + 1}^{m} y \right)
\end{aligned}$$

By Lemma 4.1.4:

$$\begin{aligned}
&= \frac{1}{m(m+1)} \cdot \left(\binom{m+1}{2} + \left\lfloor \frac{(m+1)^2}{4} \right\rfloor \right) \\
&= \frac{1}{2} + \left\lfloor \frac{(m+1)^2}{4} \right\rfloor \cdot \frac{1}{m(m+1)}
\end{aligned}$$

When m is odd, $\left\lfloor \frac{(m+1)^2}{4} \right\rfloor = \frac{(m+1)^2}{4}$, then

$$\begin{aligned} &= \frac{1}{2} + \frac{(m+1)^2}{4} \cdot \frac{1}{m(m+1)} \\ &= \frac{1}{2} + \frac{m+1}{4m} \\ &= \frac{3}{4} + \frac{1}{4m}. \end{aligned}$$

When m is even, $\left\lfloor \frac{(m+1)^2}{4} \right\rfloor = \frac{m(m+2)}{4}$, then

$$\begin{aligned} &= \frac{1}{2} + \frac{m(m+2)}{4} \cdot \frac{1}{m(m+1)} \\ &= \frac{1}{2} + \frac{m+2}{4(m+1)} \\ &= \frac{1}{2} + \frac{m+1}{4(m+1)} + \frac{1}{4(m+1)} \\ &= \frac{3}{4} + \frac{1}{4(m+1)}. \end{aligned}$$

Rewriting:

$$V_g[\pi^{in} \rhd \mathsf{S}] = \begin{cases} \frac{3}{4} + \frac{1}{4m} & \text{, if } m \text{ is odd} \\ \frac{3}{4} + \frac{1}{4(m+1)} & \text{, if } m \text{ is even.} \end{cases}$$

Unifying for an arbitrary m:

$$V_g[\pi^{in} \rhd \mathsf{S}] = \frac{3}{4} + \frac{1}{4(\lfloor \frac{m}{2} \rfloor + \lceil \frac{m+1}{2} \rceil)}.$$

□

Now for Equation (4.19).

Proof.

$$V_g[\pi^{out} \rhd \mathsf{S}] = \sum_{y \in \mathcal{Y}} Pr[y] \cdot V_g(\delta^{out,y})$$

Def. of S and by Lemmas 4.1.5 and 4.1.6 (ii):

$$\begin{aligned} &= \sum_{y=0}^{m} \frac{1}{m+1} \cdot \max\left\{ \frac{y+1}{m+2}, \frac{m-y+1}{m+2} \right\} \\ &= \frac{1}{(m+1)(m+2)} \sum_{y=0}^{m} \max\{y, m-y\} + 1 \end{aligned}$$

$$= \frac{1}{(m+1)(m+2)} \sum_{y=0}^{\lfloor \frac{m}{2} \rfloor} m - y + \sum_{y=\lfloor \frac{m}{2} \rfloor + 1}^{m} y + \sum_{y=0}^{m} 1$$

By Lemma 4.1.4:

$$= \frac{1}{(m+1)(m+2)} \left(\binom{m+1}{2} + \left\lfloor \frac{(m+1)^2}{4} \right\rfloor + (m+1) \right)$$
$$= \frac{1}{(m+1)(m+2)} \left(\frac{m(m+1)}{2} + \left\lfloor \frac{(m+1)^2}{4} \right\rfloor + (m+1) \right)$$
$$= \frac{m}{2(m+2)} + \frac{\left\lfloor \frac{(m+1)^2}{4} \right\rfloor}{(m+1)(m+2)} + \frac{1}{m+2}$$
$$= \frac{1}{2} + \frac{\left\lfloor \frac{(m+1)^2}{4} \right\rfloor}{(m+1)(m+2)}$$

When m is odd:

$$= \frac{1}{2} + \frac{(m+1)^2}{4(m+1)(m+2)}$$
$$= \frac{1}{2} + \frac{m+1}{4(m+1)}$$
$$= \frac{3m+5}{4(m+2)}$$
$$= \frac{3}{4} - \frac{1}{4(m+2)}.$$

When m is even:

$$= \frac{1}{2} + \frac{m(m+2)}{4(m+1)(m+2)}$$
$$= \frac{1}{2} + \frac{m}{4(m+1)}$$
$$= \frac{3m+2}{4(m+1)}$$
$$= \frac{3}{4} - \frac{1}{4(m+1)}.$$

Rewriting:

$$V_g[\pi^{out} \rhd \mathsf{S}] = \begin{cases} \frac{3}{4} - \frac{1}{4(m+2)} & \text{, if } m \text{ is odd} \\ \frac{3}{4} - \frac{1}{4(m+1)} & \text{, if } m \text{ is even.} \end{cases}$$

Unifying for an arbitrary m:

$$V_g[\pi^{out} \rhd \mathsf{S}] = \frac{3}{4} - \frac{1}{4\left(\left\lfloor\frac{m+1}{2}\right\rfloor + \left\lceil\frac{m}{2}\right\rceil + 1\right)}.$$

□

And finally for Equation (4.20).

Proof.

$$V_g[\pi^{unk} \rhd \mathsf{S}] = \sum_{y\in\mathcal{Y}} Pr[y] \cdot V_g(\delta^{unk,y})$$

Def. of S and by Lemmas 4.1.5 and 4.1.6 (iii):

$$= \sum_{y=0}^{m} \frac{1}{m+1} \cdot \frac{n + \max\{ny + 2y - m, nm - (ny + 2y - m)\}}{n(m+2)}$$

$$= \frac{1}{n(m+1)(m+2)} \sum_{y=0}^{m} n + \max\{ny + 2y - m, n + nm - ny - 2y + m\}$$

$$= \frac{1}{n(m+1)(m+2)} \left(n(m+1) + \sum_{y=0}^{m} \max\{y(n+2) - m, n(m-y) - 2y + m\} \right)$$

To remove the max, split the summation in two cases. The left part $y(n+2) - m \geq n(m-y) - 2y + m$ when $m \geq \left\lfloor\frac{m}{2}\right\rfloor$. Then

$$= \frac{n(m+1)}{n(m+1)(m+2)} +$$

$$\frac{1}{n(m+1)(m+2)} \left(\sum_{y=0}^{\left\lfloor\frac{m}{2}\right\rfloor} n(m-y) - 2y + m + \sum_{y=\left\lfloor\frac{m}{2}\right\rfloor+1}^{n} y(n+2) - m \right)$$

$$= \frac{n(m+1)}{n(m+1)(m+2)} +$$

$$\frac{1}{n(m+1)(m+2)} \left(n \sum_{y=0}^{\left\lfloor\frac{m}{2}\right\rfloor} (m-y) - 2 \sum_{y=0}^{\left\lfloor\frac{m}{2}\right\rfloor} y + \sum_{y=0}^{\left\lfloor\frac{m}{2}\right\rfloor} m + \right.$$

$$\left. (n+2) \sum_{y=\left\lfloor\frac{m}{2}\right\rfloor+1}^{n} y - \sum_{y=\left\lfloor\frac{m}{2}\right\rfloor+1}^{n} m \right)$$

$$
\begin{aligned}
&= \frac{n(m+1)}{n(m+1)(m+2)} + \\
&\quad \frac{1}{n(m+1)(m+2)} \left(\frac{n(2m - \lfloor \frac{m}{2} \rfloor)(\lfloor \frac{m}{2} \rfloor + 1)}{2} - \frac{2 \lfloor \frac{m}{2} \rfloor (\lfloor \frac{m}{2} \rfloor + 1)}{2} + \frac{2m(\lfloor \frac{m}{2} \rfloor + 1}{2} + \right. \\
&\qquad \left. \frac{(n+2)(m + \lfloor \frac{m}{2} \rfloor + 1)(m - \lfloor \frac{m}{2} \rfloor)}{2} - \frac{2m(m - \lfloor \frac{m}{2} \rfloor)}{2} \right) \\
&= \frac{n(m+1)}{n(m+1)(m+2)} + \\
&\quad \frac{1}{2n(m+1)(m+2)} \left(\left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) \cdot \left(n \left(2m - \left\lfloor \frac{m}{2} \right\rfloor \right) - 2 \left\lfloor \frac{m}{2} \right\rfloor + 2m \right) \right. \\
&\qquad \left. + \left(m - \left\lfloor \frac{m}{2} \right\rfloor \right) \cdot \left((n+2) \left(m + \left\lfloor \frac{m}{2} \right\rfloor + 1 \right) - 2m \right) \right) \\
&= \frac{n(m+1)}{n(m+1)(m+2)} + \\
&\quad \frac{1}{2n(m+1)(m+2)} \left(\left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) \cdot \left(2nm - n \left\lfloor \frac{m}{2} \right\rfloor - 2 \left\lfloor \frac{m}{2} \right\rfloor + 2m \right) \right. \\
&\qquad \left. + \left(m - \left\lfloor \frac{m}{2} \right\rfloor \right) \cdot \left(nm + n \left\lfloor \frac{m}{2} \right\rfloor + n + 2 \left\lfloor \frac{m}{2} \right\rfloor + 2 \right) \right) \\
&= \frac{n(m+1)}{n(m+1)(m+2)} + \\
&\quad \frac{1}{2n(m+1)(m+2)} \left(2nm \left\lfloor \frac{m}{2} \right\rfloor - 2n \left\lfloor \frac{m}{2} \right\rfloor^2 - 4 \left\lfloor \frac{m}{2} \right\rfloor^2 + 4m \left\lfloor \frac{m}{2} \right\rfloor \right. \qquad (4.21) \\
&\qquad \left. + 3nm - 2n \left\lfloor \frac{m}{2} \right\rfloor - 4 \left\lfloor \frac{m}{2} \right\rfloor + nm^2 + 4m \right)
\end{aligned}
$$

When m is even, $\lfloor \frac{m}{2} \rfloor = m/2$, therefore Equation (4.21) becomes

$$
\begin{aligned}
&\frac{n(m+1)}{n(m+1)(m+2)} + \\
&\frac{1}{2n(m+1)(m+2)} \left(2nm \left(\frac{m}{2} \right) - 2n \left(\frac{m}{2} \right)^2 - 4 \left(\frac{m}{2} \right)^2 + 4m \left(\frac{m}{2} \right) + 3nm \right. \\
&\qquad \left. - 2n \left(\frac{m}{2} \right) - 4 \left(\frac{m}{2} \right) + nm^2 + 4m \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n(m+1)}{n(m+1)(m+2)} + \\
&\quad \frac{1}{2n(m+1)(m+2)} \left(nm^2 - \frac{nm^2}{2} + m^2 + 2nm + nm^2 + 2m \right) \\
&= \frac{2n(m+1) + \frac{3nm^2}{2} + m^2 + 2nm + 2m}{2n(m+1)(m+2)} \\
&= \frac{\frac{3nm^2}{2} + m^2 + 4nm + 2n + 2m}{2n(m+1)(m+2)} \\
&= \frac{(m+2)(3nm + 2m + 2n)}{4n(m+1)(m+2)} \\
&= \frac{3nm + 2m + 2n}{4n(m+1)} \\
&= \frac{m(3m+4) + (n-m)(3m+2)}{4n(m+1)} \\
&= \frac{m}{n} \cdot \left(\frac{3m+4}{4(m+1)} \right) + \frac{n-m}{n} \cdot \left(\frac{3m+2}{4(m+1)} \right) \\
&= \frac{m}{n} \cdot \left(\frac{3}{4(m+1)} + \frac{1}{4m} \right) + \frac{n-m}{n} \cdot \left(\frac{3}{4} - \frac{1}{4(m+1)} \right) \\
&= \frac{m}{n} \cdot V_g[\pi^{in} \rhd \mathsf{S}] + \frac{n-m}{n} \cdot V_g[\pi^{out} \rhd \mathsf{S}] \ .
\end{aligned}
$$

When m is odd, $\left\lfloor \frac{m}{2} \right\rfloor = \frac{m-1}{2}$, therefore Equation (4.21) becomes

$$
\begin{aligned}
&\frac{n(m+1)}{n(m+1)(m+2)} + \\
&\frac{1}{2n(m+1)(m+2)} \left(2nm \left(\frac{m-1}{2} \right) - 2n \left(\frac{m-1}{2} \right)^2 - 4 \left(\frac{m-1}{2} \right)^2 \right. \\
&\qquad + 4m \left(\frac{m-1}{2} \right) + 3nm - 2n \left(\frac{m-1}{2} \right) - 4 \left(\frac{m-1}{2} \right) \\
&\qquad \left. + nm^2 + 4m \right) \\
&= \frac{n(m+1)}{n(m+1)(m+2)} + \\
&\frac{1}{2n(m+1)(m+2)} \left(\frac{2nm(m-1)}{2} - \frac{2n(m-1)^2}{2} - \frac{4(m-1)^2}{4} + \frac{4m(m-1)}{2} \right. \\
&\qquad \left. + 3nm - \frac{2n(m-1)}{2} - \frac{4(m-1)}{2} + nm^2 + 4m \right) \\
&= \frac{n(m+1)}{n(m+1)(m+2)} + \frac{1}{2n(m+1)(m+2)} \left(\frac{3nm^2}{2} + 2nm + m^2 + \frac{n}{2} + 2m + 1 \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{4n(m+1) + 3nm^2 + 4nm + 2m^2 + n + 4m + 2}{4n(m+1)(m+2)} \\
&= \frac{3nm^2 + 2m^2 + 8nm + 5n + 4m + 2}{4n(m+1)(m+2)} \\
&= \frac{(m+1)(3nm + 2m + 5n + 2)}{4n(m+1)(m+2)} \\
&= \frac{3nm + 2m + 5n + 2}{4n(m+2)} \\
&= \frac{(3m+1)(m+2) + 3nm + 5n - 3m^2 - 5m}{4n(m+2)} \\
&= \frac{3m+1}{4n} + \frac{3nm + 5n - 3m^2 - 5m}{4n(m+2)} \\
&= \frac{m}{n} \cdot \left(\frac{3m+1}{4m}\right) + \frac{n-m}{n} \cdot \left(\frac{3m+5}{4(m+2)}\right) \\
&= \frac{m}{n} \cdot \left(\frac{3}{4} + \frac{1}{4m}\right) + \frac{n-m}{n} \cdot \left(\frac{3}{4} - \frac{1}{4(m+2)}\right) \\
&= \frac{m}{n} \cdot V_g[\pi^{in} \rhd \mathsf{S}] + \frac{n-m}{n} \cdot V_g[\pi^{out} \rhd \mathsf{S}] \ .
\end{aligned}
$$

□

Now we analyze the posterior vulnerabilities for adversaries in $\mathcal{G}^d$. We first present two lemmas that will be useful in the proof of Theorem 4.1.4. These lemmas present closed formulas for the marginal distribution on sample histograms (i.e., $Pr[y]$) and for the vulnerability of a specific output y (i.e., $Pr[X \mid Y = y]$).

Lemma 4.1.7 (Marginal on $\mathcal{Y}$ for $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$)**.** *Given the set of secrets $\mathcal{X}$, the prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ and channel S, the marginal probability distribution on $\mathcal{Y}$ is*

$$Pr[y] = \binom{m}{y} 2^{-m}. \tag{4.22}$$

Lemma 4.1.8 (Vulnerability of a specific output y)**.** *Let $\mathcal{X}$ be the set of secrets, $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ be prior distributions on $\mathcal{X}$, g be the gain function for attribute inference attack and S be the channel. Given that the adversary observed some output y, the posterior vulnerability given y is*

(i)

$$V_g(\hat{\delta}^{in,y}) = \max\left\{\frac{y}{m}, \frac{m-y}{m}\right\}, \tag{4.23}$$

(ii)

$$V_g(\hat{\delta}^{out,y}) = \frac{1}{2}, \tag{4.24}$$

(iii)

$$V_g(\hat{\delta}^{nk,y}) = \frac{1}{n}\left(\frac{n-m}{2} + \max\{y, m-y\}\right). \tag{4.25}$$

where $\hat{\delta}^{in,y}$, $\hat{\delta}^{out,y}$ and $\hat{\delta}^{nk,y}$ are the inner distributions when $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ are, respectively, the prior distributions, and when y is observed, i.e., $Pr[X \mid Y = y]$.

Finally we present in Theorem 4.1.4 closed formulas for posterior vulnerabilities.

Theorem 4.1.4 (Posterior vulnerability for prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$)**.** *Given the prior distribution $\hat{\pi}^{in}$, the gain function g for attribute inference attack and the channel S, the posterior vulnerability $V_g[\hat{\pi}^{in} \triangleright \mathsf{S}]$ that represents the expected probability of $\mathcal{A}^{in}$ guessing correctly the target's attribute value is*

(i) for adversary $\mathcal{A}^{in}$:

$$V_g[\hat{\pi}^{in} \triangleright \mathsf{S}] = \frac{1}{2} + \frac{1}{2^m}\binom{m-1}{\lfloor \frac{m-1}{2} \rfloor}, \tag{4.26}$$

(ii) for adversary $\mathcal{A}^{out}$:

$$V_g[\hat{\pi}^{out} \triangleright \mathsf{S}] = V_g(\hat{\pi}^{out}) = \frac{1}{2}, \tag{4.27}$$

(iii) for adversary $\mathcal{A}^{nk}$:

$$V_g[\hat{\pi}^{nk} \triangleright \mathsf{S}] = \frac{m}{n} \cdot V_g[\hat{\pi}^{in} \triangleright \mathsf{S}] + \frac{m-n}{n} \cdot V_g[\hat{\pi}^{out} \triangleright \mathsf{S}]. \tag{4.28}$$

We start by the proof of Equation (4.26).

Proof.

$$V_g[\hat{\pi}^{in} \triangleright \mathsf{S}] = \sum_{y \in \mathcal{Y}} Pr[y] \cdot V_g(\hat{\delta}^{in,y})$$

By Lemmas 4.1.7 and 4.1.8 (i):

$$\begin{aligned} &= \sum_{y=0}^{m} \binom{m}{y} 2^{-m} \cdot \max\left\{\frac{y}{m}, \frac{m-y}{m}\right\} \\ &= \frac{1}{2^m} \sum_{y=0}^{m} \binom{m}{y} \frac{\max\{y, m-y\}}{m} \end{aligned} \tag{4.29}$$

by Theorem 19 in [21]:

$$= \frac{1}{2} + \frac{1}{2^m}\binom{m-1}{\lfloor \frac{m-1}{2} \rfloor} .$$

□

Now for Equation (4.27).

Proof.

$$V_g[\hat{\pi}^{out} \rhd \mathsf{S}] = \sum_{y \in \mathcal{Y}} Pr[y] \cdot V_g(\hat{\delta}^{out,y})$$

By Lemmas 4.1.7 and 4.1.8 (ii):

$$\begin{aligned} &= \sum_{y=0}^{m} \binom{m}{y} 2^{-m} \cdot \frac{1}{2} \\ &= \frac{1}{2^{m+1}} \cdot \sum_{y=0}^{m} \binom{m}{y} \\ &= \frac{1}{2^{m+1}} \cdot 2^m \\ &= \frac{1}{2} . \end{aligned}$$

□

And finally for Equation (4.28).

Proof.

$$V_g[\hat{\pi}^{unk} \rhd \mathsf{S}] = \sum_{y \in \mathcal{Y}} Pr[y] \cdot V_g(\hat{\delta}^{unk,y})$$

By Lemmas 4.1.7 and 4.1.8 (iii):

$$\begin{aligned} &= \sum_{y=0}^{m} \binom{m}{y} 2^{-m} \cdot \frac{1}{n} \left(\frac{n-m}{2} + \max\{y, m-y\} \right) \\ &= \frac{1}{n2^m} \sum_{y=0}^{m} \binom{m}{y} \left(\frac{n-m}{2} + \max\{y, m-y\} \right) \end{aligned}$$

Split the summation:

$$= \frac{1}{n2^m} \left[\frac{n-m}{2} \sum_{y=0}^{m} \binom{m}{y} + \sum_{y=0}^{m} \binom{m}{y} \max\{y, m-y\} \right]$$

Using the same result of Equation (4.29), but with some algebraic manipulation:

$$= \frac{1}{n2^m}\left[\frac{n-m}{2}\cdot 2^m + m\left(2^{m-1} + \binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}\right)\right]$$
$$= \frac{n-m}{2n} + \frac{m}{n2^m}\left(2^{m-1} + \binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}\right)$$
$$= \frac{n-m}{n}\cdot\frac{1}{2} + \frac{m}{n}\left(\frac{1}{2} + \frac{1}{2^m}\binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}\right)$$

Change order of summation:

$$= \frac{m}{n}\left(\frac{1}{2} + \frac{1}{2^m}\binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}\right) + \frac{n-m}{n}\cdot\frac{1}{2}$$
$$= \frac{m}{n}\cdot V_g[\hat{\pi}^{in} \triangleright \mathsf{S}] + \frac{n-m}{n}\cdot V_g[\hat{\pi}^{out} \triangleright \mathsf{S}]\ .$$

□

In the next section we present the results for the data analyst adversary $\mathcal{A}^{ut}$.

4.2 Data analyst and sample's utility

We introduce in this section the results related to adversary $\mathcal{A}^{ut}$ that represents a data analyst trying to infer the distribution on the binary attribute in the population. We start by presenting the prior utility loss whose operational interpretation can be understood as *"The expected error of the adversary trying to infer the frequency of value* $\mathtt{a}$ *in the population"*. After that, in Section 4.1.2, we present results on posterior utility loss that describes the adversary's expected error when she is trying to infer the frequency of value a in the population after observing the sample's histogram. All definition related to this adversary are detailed in Section 3.4.

4.2.1 Results on prior utility loss

We start this section by proving four lemmas that will be helpful in the proofs of prior utility loss. The main results are presented in Theorems 4.2.1 and 4.2.2. The proofs of all lemmas can be found in Appendix B.

Lemma 4.2.1 (Guessing symmetry when n is even)**.** *Let* $p \geq 1$ *and* $0 \leq k \leq 2p$*. Let also*

$$f(k) = k^2 - 2kp \ . \tag{4.30}$$

We have that

$$f(k) = f(2p - k).$$

Lemma 4.2.2 (Guessing symmetry when n is odd)**.** *Let $p \geq 1$ and $0 \leq k \leq 2p + 1$. Let also*

$$f'(k) = k^2 - 2kp - k \ . \tag{4.31}$$

We have that

$$f'(k) = f'(2p + 1 - k).$$

Lemma 4.2.3 (Sum of differences when n is even)**.** *Let $n \geq 2$ be even. We have that*

$$\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| = \frac{n(n+2)}{4} \ , \tag{4.32}$$

where the minimum in Equation (4.32) happens when $k = \frac{n}{2}$.

Lemma 4.2.4 (Sum of differences when n id odd)**.** *Let $n \geq 1$ be odd. We have that*

$$\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| = \frac{(n+1)^2}{4} \ , \tag{4.33}$$

where the minimum in Equation (4.33) happens when $k = \frac{n-1}{2}$.

Recall that π^{ut} represents the data analyst's prior knowledge for an adversary in $\mathcal{G}^f$. Using lemmas showed above we present a closed formula for prior utility loss $U_\ell(\pi^{ut})$ in Theorem 4.2.1.

Theorem 4.2.1 (Prior utility loss for π^{ut})**.** *Given the prior distribution π^{ut} and the loss function ℓ, the prior vulnerability is*

$$U_\ell(\pi^{ut}) = \frac{1}{4} + \frac{1}{4\left(\left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n+1}{2} \right\rceil\right)}. \tag{4.34}$$

Proof.

$$\begin{aligned}
U_\ell(\pi^{ut}) &= \min_{w \in \mathcal{W}} \sum_{p \in \mathcal{X}^{ut}} \pi^{ut}_p \cdot \ell(w, p) \\
&= \min_{w \in \mathcal{W}} \sum_{p \in \mathcal{X}^{ut}} \frac{\left| w - \frac{n_{\mathtt{a}}(p)}{n} \right|}{(n+1)\binom{n}{n_{\mathtt{a}}(p)}} \\
&= \frac{1}{n+1} \min_{w \in \mathcal{W}} \sum_{p \in \mathcal{X}^{ut}} \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right| \cdot \binom{n}{n_{\mathtt{a}}(p)}^{-1}
\end{aligned}$$

We can group secrets $p \in \mathcal{X}^{ut}$ by $n_{\mathtt{a}}(p) \in [0, n]$, and we have that $|p : n_{\mathtt{a}}(p) = i| = \binom{n}{i}$.

$$\begin{aligned}
&= \frac{1}{n+1} \min_{w \in \mathcal{W}} \sum_{i=0}^{n} \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right| \cdot \binom{n}{i} \cdot \binom{n}{i}^{-1} \\
&= \frac{1}{n+1} \min_{w \in \mathcal{W}} \sum_{i=0}^{n} \left| w - \frac{i}{n} \right| \\
&= \frac{1}{n+1} \min_{0 \le k \le n} \sum_{i=0}^{n} \left| \frac{k}{n} - \frac{i}{n} \right| \\
&= \frac{1}{n(n+1)} \min_{0 \le k \le n} \sum_{i=0}^{n} |k - i| \,.
\end{aligned}$$

For n even and by Lemma 4.2.3:

$$\begin{aligned}
\frac{1}{n(n+1)} \min_{0 \le k \le n} \sum_{i=0}^{n} |k - i| &= \frac{1}{n(n+1)} \cdot \frac{n(n+2)}{4} \\
&= \frac{n+2}{4(n+1)} \\
&= \frac{1}{4} + \frac{1}{4(n+1)}.
\end{aligned}$$

For n odd and by Lemma 4.2.4:

$$\begin{aligned}
\frac{1}{n(n+1)} \min_{0 \le k \le n} \sum_{i=0}^{n} |k - i| &= \frac{1}{n(n+1)} \cdot \frac{(n+1)^2}{4} \\
&= \frac{n+1}{4n} \\
&= \frac{1}{4} + \frac{1}{4n}.
\end{aligned}$$

Rewriting:

$$U_\ell(\pi^{ut}) = \begin{cases} \frac{1}{4} + \frac{1}{4(n+1)} & \text{, when } n \text{ is even} \\ \frac{1}{4} + \frac{1}{4n} & \text{, when } n \text{ is odd.} \end{cases}$$

Unifying for an arbitrary n:

$$U_\ell(\pi^{ut}) = \frac{1}{4} + \frac{1}{4\left(\left\lfloor \frac{n}{2} \right\rfloor + \left\lceil \frac{n+1}{2} \right\rceil\right)}.$$

□

We next present a closed formula for adversaries in $\mathcal{G}^d$. Recall that in the case of adversary $\mathcal{A}^{ut}$, this prior knowledge is defined by $\hat{\pi}^{ut}$.

Theorem 4.2.2 (Prior utility loss for $\hat{\pi}^{ut}$)**.** *Given the prior distribution $\hat{\pi}^{ut}$ on the set of secrets $\mathcal{X}^{ut}$ and the loss function ℓ, the prior vulnerability, i.e., the expected probability of the data analyst adversary inferring the frequency of value $\mathtt{a}$ in the population is*

$$U_\ell(\hat{\pi}^{ut}) = \frac{1}{n2^n} \min_{0\leq k\leq n} \sum_{i=0}^{n} \binom{n}{i} |k-i| \ .$$

Proof.

$$\begin{aligned} U_\ell(\hat{\pi}^{ut}) &= \min_{w\,\in\,\mathcal{W}} \sum_{p\,\in\,\mathcal{X}^{ut}} \hat{\pi}^{ut}_p \cdot \ell(w,p) \\ &= \frac{1}{2^n} \min_{w\,\in\,\mathcal{W}} \sum_{p\,\in\,\mathcal{X}^{ut}} \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right| \end{aligned}$$

Following the definition of $\mathcal{W}$, we can replace $\min\limits_{w\,\in\,\mathcal{W}}$ by $\min\limits_{0\leq k\leq n}$ and rewrite each action w as k/n. Also we can group secrets $p \in \mathcal{X}^{ut}$ by $n_{\mathtt{a}}(p) \in [0,n]$, and we have that $|p : n_{\mathtt{a}}(p) = i| = \binom{n}{i}$.

$$\begin{aligned} &= \frac{1}{2^n} \min_{0\leq k\leq n} \sum_{i=0}^{n} \binom{n}{i} \left| \frac{k}{n} - \frac{i}{n} \right| \\ &= \frac{1}{n2^n} \min_{0\leq k\leq n} \sum_{i=0}^{n} \binom{n}{i} |k-i| \ . \end{aligned}$$

□

In the next section we present results for posterior utility loss.

4.2.2 Results on posterior utility loss

We start this section by proving two lemmas that state the distribution on sample histograms and the utility loss of a specific output y. After that we present an equation for posterior utility loss for the data analyst in $\mathcal{G}^f$.

Lemma 4.2.5 (Marginal on $\mathcal{Y}$ for π^{ut})**.** *Given the prior distribution π^{ut} on the set of secrets $\mathcal{X}^{ut}$ and the channel S^{ut}, the probability of a sample's histogram $y \in \mathcal{Y}$ being the output is*

$$Pr[y] = \frac{1}{m+1} \ . \tag{4.35}$$

Lemma 4.2.6 (Utility loss for a specific output y)**.** *Let π^{ut} be a prior distribution on the set of secrets $\mathcal{X}^{ut}$, g be the gain function for attribute inference attack and S^{ut} be the channel. Given that the adversary observed some output y, the posterior vulnerability given y is*

$$U_\ell(\delta^{y,ut}) = \frac{m+1}{n(n+1)} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y} \binom{n-m}{y'} \binom{n}{y+y'}^{-1} \cdot |k-y-y'| \ .$$

where $\delta^{y,ut} \in \mathbb{D}\mathcal{X}^{ut}$ *is the inner distribution when* π^{ut} *is the prior distribution and* y *is observed (i.e.,* $Pr[X|Y = y]$*).*

Theorem 4.2.3 (Posterior utility loss for π^{ut})**.** *Given the prior distribution* π^{ut}*, the loss function* ℓ *and the channel* S^{ut}*, the posterior vulnerability is*

$$U_\ell[\pi^{ut} \rhd \mathsf{S}^{ut}] = \frac{1}{n(n+1)} \sum_{y=0}^{m} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y} \binom{n-m}{y'} \binom{n}{y+y'}^{-1} |k-y-y'|. \quad (4.36)$$

Proof.

$$U_\ell[\pi^{ut} \rhd \mathsf{S}^{ut}] = \sum_{y\in\mathcal{Y}} Pr[y]\, U_\ell(\delta^{y,ut})$$

By Lemmas 4.2.5 and 4.2.6:

$$= \frac{1}{m+1} \sum_{y=0}^{m} \frac{m+1}{n(n+1)} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y} \binom{n-m}{y'} \binom{n}{y+y'}^{-1} \cdot |k-y-y'|$$

$$= \frac{1}{n(n+1)} \sum_{y=0}^{m} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y} \binom{n-m}{y'} \binom{n}{y+y'}^{-1} \cdot |k-y-y'| \ .$$

□

We now present two lemmas for the marginal distribution on outputs and utility loss given an output y for adversaries in $\mathcal{G}^d$. In the end of this section Theorem 4.2.4 presents an equation for the posterior utility loss.

Lemma 4.2.7 (Marginal on $\mathcal{Y}$ for $\hat{\pi}^{ut}$)**.** *Given the set of secrets* $\mathcal{X}^{ut}$*, the prior* $\hat{\pi}^{ut}$*, the loss function* ℓ *and channel* S^{ut}*, we have that the marginal distribution on outputs* $\mathcal{Y}$ *is*

$$Pr[y] = \binom{m}{y} 2^{-m} \ . \quad (4.37)$$

Lemma 4.2.8 (Utility loss for a specific output y)**.** *Given the set of secrets* $\mathcal{X}^{ut}$*, the prior* $\hat{\pi}^{ut}$*, the loss function* ℓ *and channel* S^{ut}*, and given that the adversary observed the output* y*, the posterior vulnerability given this observation is*

$$U_\ell(\hat{\delta}^{y,ut}) = \frac{1}{n2^{n-m}} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} |k-y-y'| \ ,$$

where $\hat{\delta}^{y,ut} \in \mathbb{D}\mathcal{X}^{ut}$ *is the inner distribution when* $\hat{\pi}^{ut}$ *is the prior distribution and* y *is observed (i.e.,* $Pr[X|Y = y]$*).*

Theorem 4.2.4 (Posterior utility loss for $\hat{\pi}^{ut}$)**.** *Given the prior distribution $\hat{\pi}^{ut}$ on the set of secrets $\mathcal{X}^{ut}$, the loss function ℓ and the channel S^{ut}, the posterior vulnerability, i.e., the expected probability of the data analyst adversary inferring the frequency of value $\mathtt{a}$ in the population after observing the sample's histogram is*

$$U_\ell[\hat{\pi}^{ut} \rhd \mathsf{S}^{ut}] = \frac{1}{n2^n} \sum_{y=0}^{m} \binom{m}{y} \min_{0 \leq k \leq n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} |k - y - y'| \ .$$

Proof.

$$U_\ell[\hat{\pi}^{ut} \rhd \mathsf{S}^{ut}] = \sum_{y \in \mathcal{Y}} Pr[y] U_\ell(\hat{\delta}^{y,ut})$$

By Lemmas 4.2.7 and 4.2.8:

$$= \sum_{y=0}^{m} \binom{m}{y} 2^{-m} \cdot \frac{2^m}{n2^n} \cdot \min_{0 \leq k \leq n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} |k - y - y'|$$

$$= \frac{1}{n2^n} \sum_{y=0}^{m} \binom{m}{y} \min_{0 \leq k \leq n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} |k - y - y'| \ .$$

□

4.3 Discussion of results

In this section we discuss interpretations of equations presented in the last two sections and show graphs that compare their behavior when we vary the parameters (e.g., the population and sample sizes).

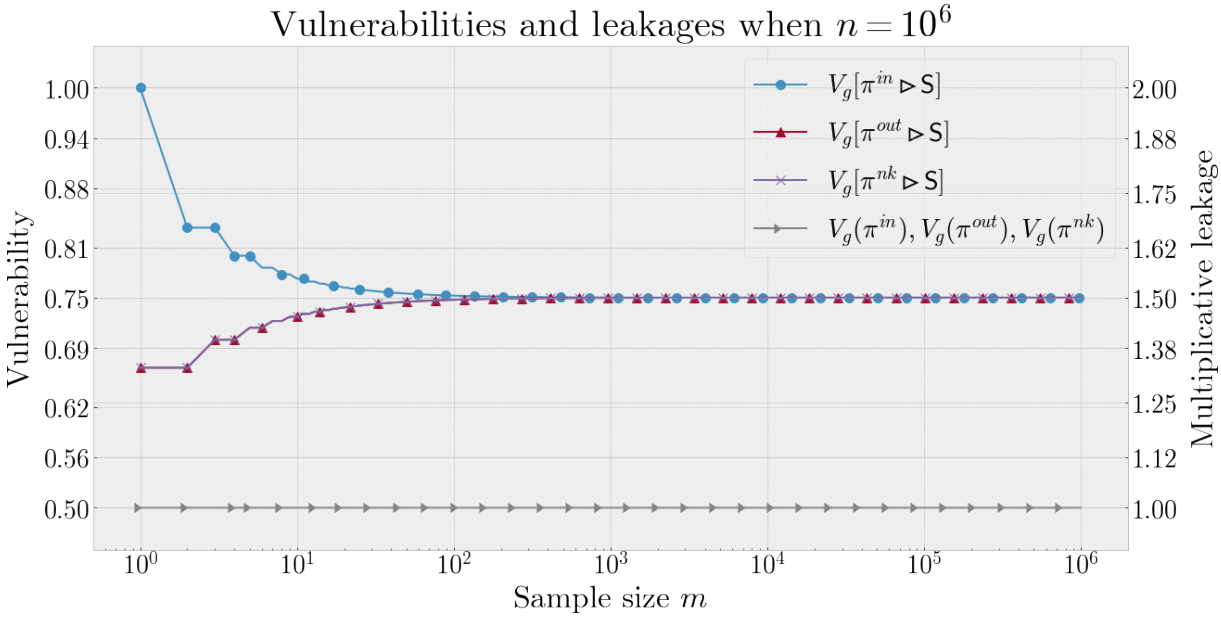
Prior vulnerability for adversaries in both $\mathcal{G}^f$ and $\mathcal{G}^d$. We start the discussion looking at the results related to the prior vulnerability of attribute inference attack performed by adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$. Theorems 4.1.1 and 4.1.2 showed that the prior vulnerability is $1/2$ regardless the adversary's prior knowledge (i.e., regardless the adversary being in $\mathcal{G}^f$ or $\mathcal{G}^d$). Before getting any information other than those described by π^{in}, π^{out}, π^{nk}, $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ or $\hat{\pi}^{nk}$, the adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ will deduce that $Pr[p_t = \mathtt{a}] = Pr[p_t = \mathtt{b}] = 1/2$, i.e., the probability of their target having value $\mathtt{a}$or $\mathtt{b}$for the sensitive attribute is the same. This is a direct consequence of both assumptions of a uniform distribution on frequency of value $\mathtt{a}$ in the population and a uniform distribution on datasets. As these assumptions induce a uniform prior on $\{\mathtt{a}, \mathtt{b}\}$ for the target's at-

tribute value, the adversary's guess could be either `a` or `b`, and their expected probability of success guessing the target's attribute value will be $1/2$.

This result corroborates the intuition that before getting any data about the population from the data curator, the adversary has a very low success rate when trying to infer the attribute value of a target, or from the point of view of the data curator, releasing no data causes no damage to privacy.

Privacy for adversaries in $\mathcal{G}^f$. Moving to posterior vulnerability, let us first analyze the group of adversaries with prior knowledge π^{in}, π^{out} and π^{nk}. Figure 4.1 shows the behavior of posterior vulnerability for a population with 1 million people when the sample size m grows.

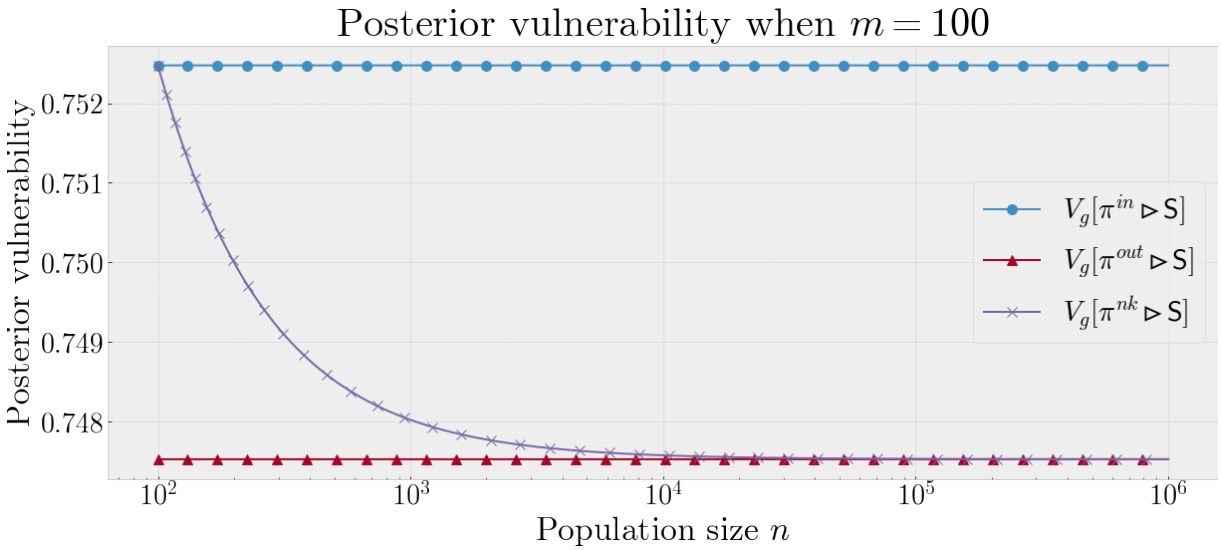
Figure 4.1: Vulnerabilities and multiplicative leakages for $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ in $\mathcal{G}^f$, fixing n and varying m. In the right side of the y axis we can also see by which factor the adversary is increasing her chance of guessing correctly the target's attribute after observing the sample. For instance, $m = 1$ implies $V_g[\pi^{in} \triangleright \mathsf{S}] = 1$ because there is only one person in the sample and the adversary knows that is the target, thus her chance of success will be doubled, i.e., $\mathcal{L}^{\times}_g(\pi^{in}, \mathsf{S}) = 2$.



Font: Elaborated by the author.

The results showed in Figure 4.1 brings us an important insight about the reliability of participating in a statistical publication. They confirmed a reasonable intuition that a person that was selected to be in the sample has always a higher probability to have his attribute value inferred by an adversary than a person that was not selected to be in the sample. Because of that, a person may be inclined to refuse to answer a research that would collect his data arguing that if he is outside the sample he will be more protected against attribute inferences. This analysis corroborates the motivation for studies involving mitigation methods for privacy in statistical publications.

Figure 4.2: Posterior vulnerability for $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ in $\mathcal{G}^f$, fixing m and varying n. Note that the x axis starts at $n = 10^2$ because the sample size $m \leq n$.



Font: Elaborated by the author.

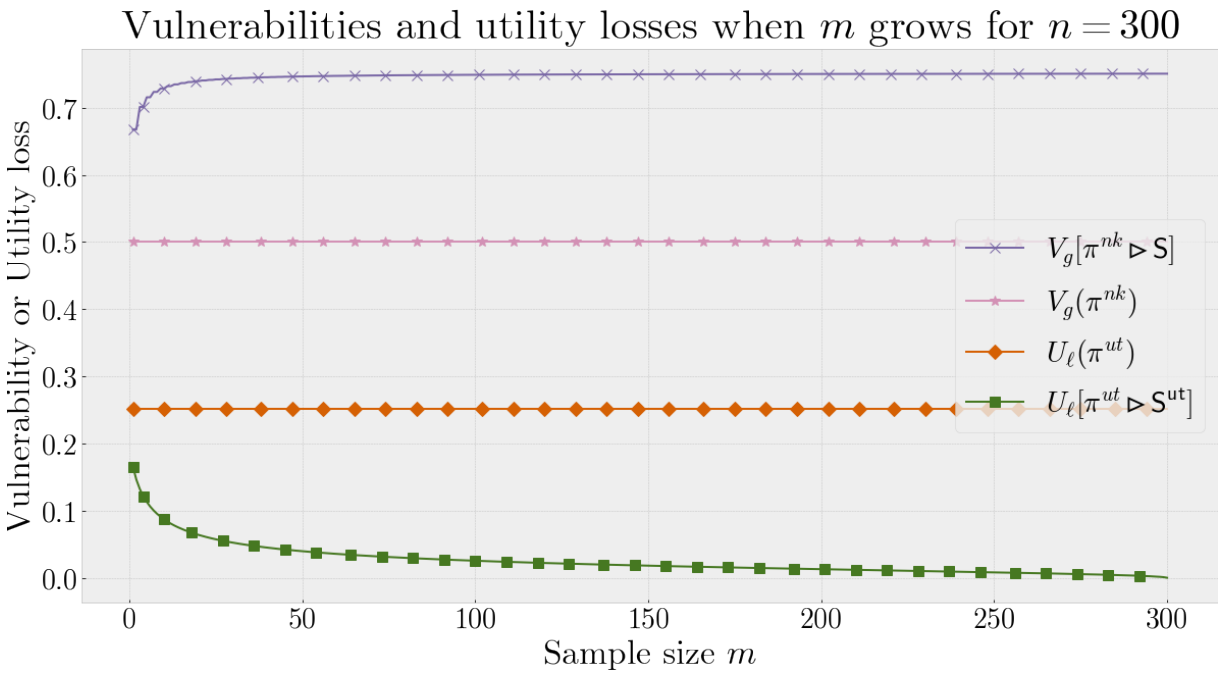
Figure 4.2 shows the behavior of posterior vulnerability when we fix the sample size and vary the number of people in the population. It is possible to see that the impact of the population size n in posterior vulnerability is very low when m is fixed. For adversaries $\mathcal{A}^{in}$ and $\mathcal{A}^{out}$ the impact is nonexistent – indeed Equations (4.18) and (4.19) in Theorem 4.1.3 depend only on the sample size m. The population size n appears in posterior vulnerability for adversary $\mathcal{A}^{nk}$, however the value of $V_g[\pi^{nk} \triangleright \mathsf{S}]$ does not change more than 1% for n varying from 10^2 to 10^6, as showed in Figure 4.2.

Utility for adversaries in $\mathcal{G}^f$. At the time we have written this thesis we were not able to find a closed formula for posterior utility of adversary $\mathcal{A}^{ut}$. The adversary's guess depends on n, m and the number of a's observed in the sample, however, the best for the adversary given these three parameters is not clear for us. Although we still have not found a closed formula for the posterior utility loss, looking at the graph in Figure 4.3 we are still able to describe the behavior of $U_\ell[\pi^{ut} \triangleright \mathsf{S}]$, also comparing it with the vulnerabilities for adversary $\mathcal{A}^{nk}$.

There is a clear trade-off between privacy and utility that can be observed. We see that, as the sample size m grows, the expected probability of $\mathcal{A}^{nk}$ inferring the attribute value of his target also grows, and on the other hand the data analyst's expected error decreases. Besides, the rate that those vulnerabilities/utility losses increase (or decrease in the case of $\mathcal{A}^{ut}$) are quite similar. The variation in posterior vulnerability/utility loss for $1 \leq m \leq 100$ is much higher than the variation for $100 < m \leq 500$, as showed in Table 4.1.

It is possible to see in the proof of Theorem 4.2.1 that the best guess for the data analyst adversary for the frequency of value a in the population without observing

Figure 4.3: Vulnerabilities and utility losses for $\mathcal{A}^{nk}$ and $\mathcal{A}^{ut}$ in $\mathcal{G}^f$ for $n = 500$ and varying the sample size m.



Font: Elaborated by the author.

m	$V_g[\pi^{nk} \triangleright \mathsf{S}]$	$U_\ell[\pi^{ut} \triangleright \mathsf{S}]$
1	66.74%	19.55%
100	74.85%	2.79%
500	75.05%	0%

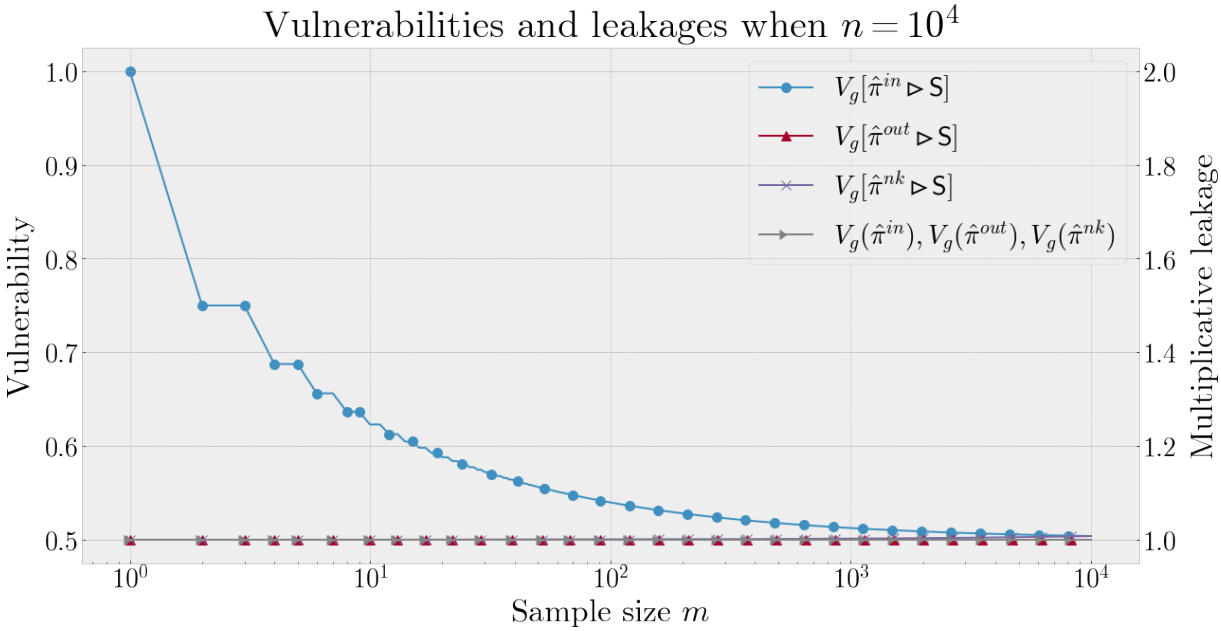
Table 4.1: Posterior vulnerability/utility loss for $n = 10^6$ and for different sample sizes m.

anything is $n/2$. This becomes from the fact we assumed before that all possible frequencies are equally probable. Thus, the expected error she will get guessing $n/2$ is at lest 25% from above or below, and this error decreases as the population size grows.

Privacy for adversaries in $\mathcal{G}^d$. Now we turn our attention to adversaries in $\mathcal{G}^d$. First we analyze the behavior of posterior vulnerability and the multiplicative leakage for adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$.

Figure 4.4 shows the posterior vulnerabilities for adversaries in $\mathcal{G}^d$, and we see that the participant in the sample is the one that suffers more damage to his privacy, i.e., the adversary has a higher expected probability of guessing correctly his attribute value. This fact also happens for adversaries in $\mathcal{G}^f$. Following Figures 4.1 and 4.4, it is clear that, for a fixed n and as m grows, while the posterior vulnerability $V_g[\hat{\pi}^{in} \triangleright \mathsf{S}]$ converges to $1/2$, $V_g[\pi^{in} \triangleright \mathsf{S}]$ converges to $3/4$, what implies in their multiplicative leakages converging to 1 and 1.5, respectively. One conclusion about this fact is that, for large samples (i.e., for values of m close to n), adversaries $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ in $\mathcal{G}^f$ has higher expected probability of success than those in $\mathcal{G}^d$.

Figure 4.4: Vulnerabilities and multiplicative leakage for $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$ in $\mathcal{G}^d$, fixing n and varying m.



Font: Elaborated by the author.

Another interesting fact is that individuals outside the sample has no damage to their privacy for adversaries in $\mathcal{G}^d$, i.e., $V_g[\hat{\pi}^{out} \triangleright \mathsf{S}] = V_g(\hat{\pi}^{out}) = 1/2$. Suppose that $\mathcal{A}^{out}$ observed $n_{\mathtt{a}}(p_{1\ldots m}) = y$ $\mathtt{a}$'s in the sample. Given that $\mathcal{A}^{out}$ knows the target is outside the sample, she will calculate which attribute value – $\mathtt{a}$ or $\mathtt{b}$ – is the most probable for people outside the sample, and she will find that

$$Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = k \mid y] = \frac{\binom{n-m}{k}}{2^{n-m}} \ ,$$

for $0 \leq k \leq n-m$. Considering the symmetry of binomials, i.e., $\binom{n}{k} = \binom{n}{n-k}$, the probability

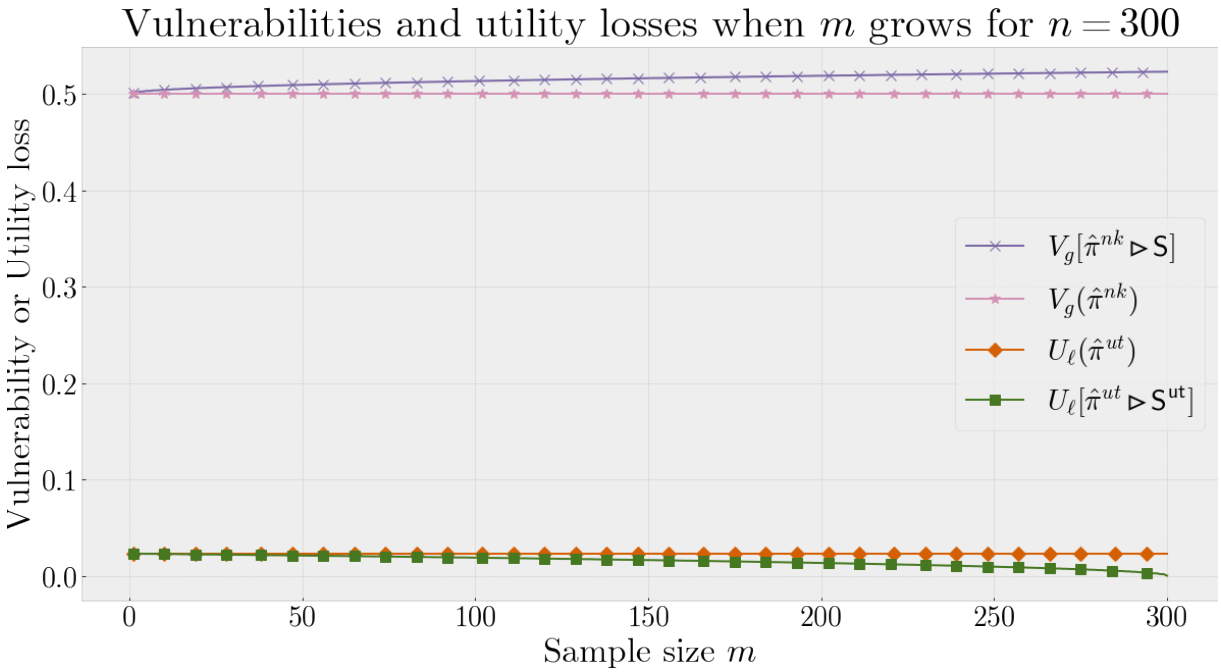
$$\begin{aligned} Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = 0 \mid y] &= Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = n-m \mid y] \\ Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = 1 \mid y] &= Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = n-m-1 \mid y] \\ &\vdots \\ Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = \lfloor (n-m)/2 \rfloor \mid y] &= Pr[n_{\mathtt{a}}(p_{m+1\ldots n}) = \lfloor (n-m+1)/2 \rfloor \mid y] \ . \end{aligned}$$

Thus the probability of the target's value $Pr[p_t = \mathtt{a}] = Pr[p_t = \mathtt{b}] = 1/2$, and therefore her expected guess will be $1/2$, regardless her guess for the target's attribute value.

Utility for adversaries in $\mathcal{G}^d$. We can observe in Figure 4.5 that the shape of the curves of posterior vulnerability $V_g[\hat{\pi}^{nk} \triangleright \mathsf{S}]$ and posterior utility loss $U_\ell[\hat{\pi}^{ut} \triangleright \mathsf{S}]$ are "mirrored", i.e., while $V_g[\hat{\pi}^{nk} \triangleright \mathsf{S}]$ increases as m grows, $U_\ell[\hat{\pi}^{ut} \triangleright \mathsf{S}]$ decreases. It again corroborates with the well known trade-off between privacy and utility found in the literature.

Nonetheless, it is possible to see in Figures 4.5 and 4.3 that the distance between prior and posterior vulnerabilities for adversary $\mathcal{A}^{nk}$ is smaller for adversaries in $\mathcal{G}^{d}$ compared to adversaries in $\mathcal{G}^{f}$. It may indicate that sample publications are more vulnerable to attribute inference attack from adversaries in $\mathcal{G}^{f}$ than those in $\mathcal{G}^{d}$.

Figure 4.5: Vulnerabilities and utility losses for $\mathcal{A}^{nk}$ and $\mathcal{A}^{ut}$ in $\mathcal{G}^{d}$, fixing n and varying m.



Font: Elaborated by the author.

Chapter 5

Conclusions

In this work we have presented a model to analyze attribute inference attack in sample publications of a single binary attribute, as well as the utility of that publication for data analysts that are interested to understand the distribution of an attribute in a certain population. The framework of QIF allowed us formalize the sample release itself and the bevahior and gains of all adversaries studied. The model enabled us to answer the following two questions:

(i) *"What is the expected probability of an adversary guessing correctly the attribute value of a single target when she observes a sample from the population?"*, and

(ii) *"When a data analyst observes a sample and tries to guess the distribution of a binary attribute in the population, how far her guess will be from the real distribution in expected values?"*.

Question (i) is related to attribute inference attack, and it was answered for adversaries with three different prior knowledge about the presence of the target in the sample, formalized as $\mathcal{A}^{in}$, $\mathcal{A}^{out}$ and $\mathcal{A}^{nk}$, and for two different assumptions about how the data was generated, formalized as groups $\mathcal{G}^f$ and $\mathcal{G}^d$ (Sections 3.3.2 and 3.3.1). On the other hand, question (ii) is related to the utility of a sample release, and it was answered for data analysts with two different assumptions about how the data was generated, formalized as groups $\mathcal{G}^f$ and $\mathcal{G}^d$ (Sections 3.4.2 and 3.4.1).

Besides we have derived closed formulas for prior and posterior vulnerability of attribute inference attack and for prior utility loss of a data analyst in $\mathcal{G}^f$. Those formulas enable us to evaluate privacy and utility of large sample/datasets, what is computationally infeasible to do using the original equations that defines the prior and posterior vulnerabilities and utility losses.

As a future work, this thesis can be considered a start point in the development of a general model that assess privacy and utility levels of a wider set of data releases such as microdata, histograms, query answers, etc.

Another expansion of this thesis includes modeling mitigation methods of privacy breaches (e.g., differential privacy) in order to compare, quantitatively, their efficiency in protecting privacy and keeping a reasonable utility level in easy terms that can be

explainable to data curators. The work could be used as a reference by institutions to motivate the usage of some mitigation methods as well as guiding their process of publicly releasing data.

References

[1] WWDC 2016. Wwdc 2016 keynote, 2016. Apple Worldwide Developers Conference 2016, 2016. Available at https://www.apple.com/apple-events/#june-2016.

[2] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

[3] Mário S. Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. *The Science of Quantitative Information Flow*. Information Security and Cryptography. Springer International Publishing, Cham, Switzerland, 2020.

[4] Mário S. Alvim, Natasha Fernandes, Annabelle McIver, and Gabriel H. Nunes. On Privacy and Accuracy in Data Releases (Invited Paper). In Igor Konnov and Laura Kovács, editors, *31st International Conference on Concurrency Theory (CONCUR 2020)*, volume 171 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 1:1–1:18, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[5] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. Improved summation from shuffling. *arXiv preprint arXiv:1909.11225*, 2019.

[6] Borja Balle, James Bell, Adria Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, pages 638–667. Springer, 2019.

[7] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459, 2017.

[8] Albert Cheu, Adam D. Smith, Jonathan R. Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In Yuval Ishai and Vincent Rijmen, editors, *Advances in Cryptology - EUROCRYPT 2019 - 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19-23, 2019, Proceedings, Part I*, volume 11476 of *Lecture Notes in Computer Science*, pages 375–403. Springer, 2019.

[9] Tore Dalenius. Towards a methodology for statistical disclosure control. 1977.

[10] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting Telemetry Data Privately. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3571–3580, 2017.

[11] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[12] George Duncan and Diane Lambert. The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2):207–217, 1989.

[13] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[14] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[15] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[16] EU. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. Available at https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[17] Benjamin CM Fung, Ke Wang, Ada Wai-Chee Fu, and S Yu Philip. *Introduction to privacy-preserving data publishing: Concepts and techniques*. Chapman and Hall/CRC, 2010.

[18] Aris Gkoulalas-Divanis and Grigorios Loukides. *Medical data privacy handbook*. Springer, 2015.

[19] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*, volume 2. Wiley New York, 2012.

[20] Inep. Instituto nacional de estudos e pesquisas educacionais anísio teixeira - institucional, 2022. Available at https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional.

[21] Mireya Jurado, Mário Alvim, Ramon Gonze, and Catuscia Palamidessi. Analyzing the shuffle model through the lens of quantitative information flow. Technical report, 2023.

[22] Gregory J Matthews and Ofer Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5:1–29, 2011.

[23] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

[24] Gabriel Henrique Lopes Gomes Alves Nunes. A formal quantitative study of privacy in the publication of official educational censuses in Brazil. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, April 2021.

[25] Jessie Pease and Julien Freudiger. Engineering privacy for your users. Apple Worldwide Developers Conference 2016, 2016. Available at https://developer.apple.com/videos/play/wwdc2016/709/.

[26] Jocelyn Quaintance and Henry W Gould. *Combinatorial identities for Stirling numbers: the unpublished notes of HW Gould*. World Scientific, 2015.

[27] Robert Sedgewick and Philippe Flajolet. *An introduction to the analysis of algorithms*. Pearson Education India, 2013.

[28] Neil J. A. Sloane and The OEIS Foundation Inc. The on-line encyclopedia of integer sequences, 2020.

[29] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[30] Leon Willenborg and Ton De Waal. *Elements of statistical disclosure control*, volume 155. Springer Science & Business Media, 2012.

Appendix A

Proofs of Lemmas – Privacy

Here we present the proofs of all lemmas related to privacy analysis in Section 4.1.2.

Lemma 4.1.1 (Summation on binomials 1)**.** *Let $1 \leq y \leq m \leq n$ be integers. The following equivalence remains:*

$$\sum_{k=0}^{n-m} \binom{m-1}{y-1}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \frac{y(n+1)}{m(m+1)} \tag{4.8}$$

and analogously:

$$\sum_{k=0}^{n-m} \binom{m-1}{y}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \frac{(m-y)(n+1)}{m(m+1)}. \tag{4.9}$$

Proof. First, for Equation (4.8):

$$\sum_{k=0}^{n-m} \binom{m-1}{y-1}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \frac{y(n+1)}{m(m+1)}$$

Note: $\binom{m}{y}\frac{y}{m} = \binom{m-1}{y-1}$

$$\sum_{k=0}^{n-m} \binom{m}{y}\binom{n-m}{k}\binom{n}{y+k}^{-1} \cdot \frac{y}{m} = \frac{y(n+1)}{m(m+1)}$$

$$\sum_{k=0}^{n-m} \binom{m}{y}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \frac{n+1}{m+1} \tag{A.1}$$

Note: $\frac{n+1}{m+1} = \binom{n+1}{m+1}\binom{n}{m}^{-1}$.

$$\sum_{k=0}^{n-m} \binom{m}{y}\binom{n-m}{k}\binom{n}{y+k}^{-1} = \binom{n+1}{m+1}\binom{n}{m}^{-1}$$

To factorials.

$$\sum_{k=0}^{n-m} \frac{(n-m)!}{k!(n-m-k)!}\frac{m!}{y!(m-y)!}\frac{(y+k)!(n-y-k)!}{n!} = \binom{n+1}{m+1}\binom{n}{m}^{-1}$$

Isolate $\binom{n}{m}^{-1}$.

$$\binom{n}{m}^{-1} \sum_{k=0}^{n-m} \frac{1}{k!(n-m-k)!} \frac{1}{y!(m-y)!} \frac{(y+k)!(n-y-k)!}{1} = \binom{n+1}{m+1} \binom{n}{m}^{-1}$$

Cancel.

$$\sum_{k=0}^{n-m} \frac{1}{k!(n-m-k)!} \frac{1}{y!(m-y)!} \frac{(y+k)!(n-y-k)!}{1} = \binom{n+1}{m+1}$$

Re-arrange.

$$\sum_{k=0}^{n-m} \frac{(y+k)!}{k!y!} \frac{(n-y-k)!}{(n-m-k)!(m-y)!} = \binom{n+1}{m+1}$$

To binomials.

$$\sum_{k=0}^{n-m} \binom{y+k}{y} \binom{n-(y+k)}{m-y} = \binom{n+1}{m+1}$$

Define $i = y + k$

$$\cdot \sum_{i=y}^{n-m+y} \binom{i}{y} \binom{n-i}{m-y} = \binom{n+1}{m+1}$$

Expand summation range. Recall $\binom{n}{k} = 0$ if $k > n$. For $0 \leq i < y$, $\binom{i}{y} = 0$ because $i < y$. Similarly, for $n - m + y < i \leq n$, $\binom{n-i}{m-y} = 0$ because $n - i$ can at most be $m - y - 1$, which is less than $m - y$.

$$\sum_{i=0}^{n} \binom{i}{y} \binom{n-i}{m-y} = \binom{n+1}{m+1}$$

By Vandermonde Convolution [26]:

$$\binom{n+1}{m+1} = \binom{n+1}{m+1}$$

And for Equation (4.9):

$$\sum_{k=0}^{n-m} \binom{m-1}{y} \binom{n-m}{k} \binom{n}{y+k}^{-1} = \frac{(m-y)(n+1)}{m(m+1)}$$

Note: $\binom{m}{y} \frac{m-y}{m} = \binom{m-1}{y}$

$$\sum_{k=0}^{n-m} \binom{m}{y} \binom{n-m}{k} \binom{n}{y+k}^{-1} \cdot \frac{m-y}{m} = \frac{(m-y)(n+1)}{m(m+1)}$$

$$\sum_{k=0}^{n-m} \binom{m}{y} \binom{n-m}{k} \binom{n}{y+k}^{-1} = \frac{n+1}{m+1} \tag{A.2}$$

The equality in Equation (A.2) is the same as the equality in Equation (A.1), which we have already demonstrated to be true, then we conclude our proof. □

Lemma 4.1.3 (Summation on binomials 2)**.** *Let $1 \leq y \leq m \leq n$ be integers. The following equivalence remains:*

$$\sum_{k=0}^{n-m-1} \binom{m}{y} \binom{n-m-1}{k} \binom{n}{y+k+1}^{-1} = \frac{(n+1)(y+1)}{(m+1)(m+2)}, \tag{4.11}$$

and analogously:

$$\sum_{k=0}^{n-m-1} \binom{m}{y} \binom{n-m-1}{k} \binom{n}{y+k}^{-1} = \frac{(n+1)(m-y+1)}{(m+1)(m+2)}. \tag{4.12}$$

Proof. For the equality in Equation (4.11), let's first reduce it:

$$\begin{aligned}\sum_{k=0}^{n-m-1} \binom{m}{y} \binom{n-m-1}{k} \binom{n}{y+k+1}^{-1} &= \binom{m}{y} \sum_{k=0}^{n-m-1} \left(\frac{(n-m-1)!}{k!(n-m-1-k)!} \cdot \right. \\ &\qquad \left. \frac{(y+k+1)!(n-y-k-1)!}{n!} \right) \\ &= \frac{(y+1)m!(n-m-1)!}{n!} \cdot \\ &\qquad \underbrace{\sum_{k=0}^{n-m-1} \binom{y+1+k}{y+1} \binom{n-y-k-1}{m-y}}_{\mathcal{A}(n-m-1)}\end{aligned} \tag{A.3}$$

We note that $\mathcal{A}(n-m-1)$ can be rewritten as follows:

$$\mathcal{A}(n-m-1) = \sum_{k=0}^{n-m-1} \underbrace{\binom{y+1+k}{y+1}}_{a(k)} \underbrace{\binom{m-y+n-m-1-k}{m-y}}_{b(n-m-1-k)} \tag{A.4}$$

We have:

$$\mathcal{A}(\ell) = \sum_{k=0}^{\ell} a(k) b(\ell - k) \tag{A.5}$$

Hence we can see $\mathcal{A}(\ell)$ as a term coefficient in the following Cauchy product (discrete convolution of two infinite power series):

$$\sum_{\ell=0}^{\infty} \mathcal{A}(\ell) x^{\ell} = \left(\sum_{i=0}^{\infty} a(i) x^i \right) \cdot \left(\sum_{j=0}^{\infty} b(j) x^j \right) \tag{A.6}$$

Using Lemma 4.1.2:

(i)

$$\sum_{i=0}^{\infty} a(i) x^i = \sum_{i=0}^{\infty} \binom{y+1+i}{y+1} x^i = \frac{1}{(1-x)^{y+2}}$$

(ii)

$$\sum_{j=0}^{\infty} b(i)x^j = \sum_{j=0}^{\infty} \binom{m-y+j}{m-y} x^j = \frac{1}{(1-x)^{m-y+1}}$$

Using the property of generating function, i.e., that the generating function of a product is the product of the generating functions (Equation (A.6)):

$$\begin{aligned}
\sum_{\ell=0}^{\infty} \mathcal{A}(\ell)x^\ell &= \frac{1}{(1-x)^{y+2}} \cdot \frac{1}{(1-x)^{m-y+1}} \\
&= \frac{1}{(1-x)^{m+3}} \\
&= \sum_{\ell=0}^{\infty} \binom{m+2+\ell}{m+2} x^\ell \ . \qquad \text{(Lemma 4.1.2)}
\end{aligned}$$

Hence, considering the $\ell = n-m-1$ power term (Equation (A.4)):

$$\begin{aligned}
\mathcal{A}(\ell) &= \binom{m+2+n-m-1}{m+2} \\
&= \binom{n+1}{m+2}.
\end{aligned}$$

Backing to Equation (A.3):

$$\begin{aligned}
\sum_{k=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{k}\binom{n}{y+k+1}^{-1} &= \frac{(y+1)m!(n-m-1)!}{n!} \cdot \mathcal{A}(n-m-1) \\
&= \frac{(y+1)m!(n-m-1)!}{n!} \cdot \binom{n+1}{m+2} \\
&= \frac{(y+1)m!(n-m-1)!}{n!} \cdot \frac{(n+1)!}{(m+2)!(n-m-1)!} \\
&= \frac{(n+1)(y+1)}{(m+1)(m+2)}.
\end{aligned}$$

We can apply the same reasoning to Equation (4.12). For that case the terms of the Cauchy product are

$$\begin{aligned}
a'(i) &= \binom{y+i}{i} \ , \text{ and} \\
b'(j) &= \binom{m-y+1+j}{m-y+1} \ ,
\end{aligned}$$

and we use the following generating functions (Lemma 4.1.2):

$$\begin{aligned}
\sum_{i=0}^{\infty} a'(i)x^i &= \frac{1}{(1-x)^{y+1}} \ , \\
\sum_{j=0}^{\infty} b'(j)x^j &= \frac{1}{(1-x)^{m-y+2}} \ .
\end{aligned}$$

Reducing Equation (4.12), we have:

$$\sum_{k=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{k}\binom{n}{y+k}^{-1} = \binom{m}{y}\cdot\frac{y!(m-y+1)!(n-m+1)!}{n!}\cdot \underbrace{\sum_{k=0}^{n-m+1}\binom{y+k}{k}\binom{n-y-k}{m-y+1}}_{\mathcal{B}(n-m-1)}$$

and

$$\begin{aligned}\sum_{\ell=0}^{\infty}\mathcal{B}(\ell)x^{\ell} &= \left(\sum_{i=0}^{\infty}a'(i)x^{i}\right)\cdot\left(\sum_{j=0}^{\infty}b'(j)x^{j}\right)\\ &= \frac{1}{(1-x)^{y+1}}\cdot\frac{1}{(1-x)^{m-y+2}}\\ &= \frac{1}{(1-x)^{m+3}}.\end{aligned}$$

Hence:

$$\mathcal{B}(n-m-1) = \binom{m+2+n-m-1}{m+2} = \binom{n+1}{m+2}$$

Backing to Equation (4.12):

$$\begin{aligned}\sum_{k=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{k}\binom{n}{y+k}^{-1} &= \binom{m}{y}\cdot\frac{y!(m-y+1)!(n-m+1)!}{n!}\cdot\mathcal{B}(n-m-1)\\ &= \frac{m!}{y!(m-y)!}\cdot\frac{y!(m-y+1)!(n-m+1)!}{n!}\binom{n+1}{m+2}\\ &= \frac{m!(m-y+1)(n-m+1)!}{n!}\cdot\frac{(n+1)!}{(m+2)!(n-m-1)!}\\ &= \frac{(n+1)(m-y+1)}{(m+1)(m+2)}.\end{aligned}$$

□

Lemma 4.1.4 (Summations on m)**.** *Let $m \geq 1$. We have that*

$$\sum_{i=0}^{\lfloor m/2\rfloor} m-i+\sum_{i=\lfloor m/2\rfloor+1}^{m} i = \binom{m+1}{2} + \left\lfloor\frac{(m+1)^2}{4}\right\rfloor. \tag{4.13}$$

Proof. Note the right side is the triangular numbers plus quarter-squares, which is listed in the Online Encyclopedia of Integer Sequences (OEIS) [28] under integer sequence

A001859.

$$
\begin{aligned}
&= \sum_{i=0}^{\lfloor m/2 \rfloor} m - i + \sum_{i=\lfloor m/2 \rfloor + 1}^{m} i \\
&= \sum_{i=0}^{\lfloor m/2 \rfloor} m - \sum_{i=0}^{\lfloor m/2 \rfloor} i + \sum_{i=\lfloor m/2 \rfloor + 1}^{m} i \\
&= \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) m - \sum_{i=1}^{\lfloor m/2 \rfloor} i + \left(\sum_{i=1}^{m} i - \sum_{i=1}^{\lfloor m/2 \rfloor} i \right) \\
&= \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) m - \frac{\lfloor m/2 \rfloor (\lfloor m/2 \rfloor + 1)}{2} + \left(\frac{m(m+1)}{2} - \frac{\lfloor m/2 \rfloor (\lfloor m/2 \rfloor + 1)}{2} \right) \\
&= \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) m + \frac{m(m+1)}{2} - 2\frac{\lfloor m/2 \rfloor (\lfloor m/2 \rfloor + 1)}{2} \\
&= \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) m + \frac{m(m+1)}{2} - \left\lfloor \frac{m}{2} \right\rfloor \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) \left(m - \left\lfloor \frac{m}{2} \right\rfloor \right) + \frac{m(m+1)}{2} \\
&= \left\lceil \frac{m+1}{2} \right\rceil \left\lfloor \frac{m+1}{2} \right\rfloor + \frac{m(m+1)}{2} \\
&= \left\lfloor \frac{(m+1)^2}{4} \right\rfloor + \binom{m+1}{2}.
\end{aligned}
$$

□

Lemma 4.1.5 (Marginal on $\mathcal{Y}$ for π^{in}, π^{out} and π^{nk})**.** *Given the prior distributions π^{in}, π^{out} and π^{nk} on the set of secrets $\mathcal{X}$ and the channel S, the probability of a sample's histogram $y \in \mathcal{Y}$ being the output is*

$$Pr[y] = \frac{1}{m+1}. \tag{4.14}$$

Proof. Let's use the prior distribution π^{in} to construct the proof.

$$Pr[y] = \sum_{(p,t) \in \mathcal{X}} Pr[(p,t)] Pr[y|(p,t)] \tag{A.7}$$

Def. of π^{in} and S:

$$= \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \le t \le m, \\ n_{\mathtt{a}}(p_{1 \ldots m}) = y}} \frac{1}{m(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$

We need to count how many secrets $(p,t)\in\mathcal{X}$ satisfy the restrictions $1\leq t\leq m$ and $n_{\mathtt{a}}(p_{1\ldots m}) = y$. In the first m elements of p we have y $\mathtt{a}$'s, so $\binom{m}{y}$ different combinations. The other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, so $\binom{n-m}{y'}$ different combinations. Finally, $n_{\mathtt{a}}(p_{1\ldots m}) = y$ and $n_{\mathtt{a}}(p_{m+1\ldots n}) = y'$ implies $n_{\mathtt{a}}(p) = y + y'$.

$$= \frac{1}{m(n+1)} \sum_{t=1}^{m} \sum_{y'=0}^{n-m} \binom{m}{y} \binom{n-m}{y'} \binom{n}{y+y'}^{-1}$$

By Lemma 4.1.1:

$$\begin{aligned} &= \frac{1}{m(n+1)} \cdot \sum_{t=1}^{m} \frac{n+1}{m+1} \\ &= \frac{1}{m(n+1)} \cdot \frac{m(n+1)}{m+1} \\ &= \frac{1}{m+1}. \end{aligned} \tag{A.8}$$

Note that the proof above is also valid for prior distributions π^{out} and π^{nk}. The only difference between these three priors are the range of t, that are $\{1,\ldots,m\}$, $\{m+1,\ldots,n\}$ and $\{1,\ldots,n\}$, for π^{in}, π^{out} and π^{nk}, respectively. The size of these ranges are all canceled out in Equation (A.8). □

Lemma 4.1.6 (Vulnerability of a specific output y, adversaries in $\mathcal{G}^f$)**.** *Let $\mathcal{X}$ be the set of secrets, π^{in}, π^{out} and π^{nk} be prior distributions on $\mathcal{X}$, g be the gain function for attribute inference attack and S be the channel. Given that the adversary observed some output y, the posterior vulnerability given y is*

(i)

$$V_g(\delta^{in,y}) = \max\left\{\frac{y}{m}, \frac{m-y}{m}\right\}. \tag{4.15}$$

(ii)

$$V_g(\delta^{out,y}) = \max\left\{\frac{y+1}{m+2}, \frac{m-y+1}{m+2}\right\} \tag{4.16}$$

(iii)

$$V_g(\delta^{nk,y}) = \frac{n + \max\{ny+2y-m, nm-(ny+2y-m)\}}{n(m+2)} \tag{4.17}$$

where $\delta^{in,y}$, $\delta^{out,y}$ and $\delta^{nk,y}$ are the inner distributions when y is observed and when π^{in}, π^{out} and π^{nk} are, respectively, the prior distributions. These vulnerabilities can be understood as $Pr[X \mid Y = y]$ with X being the set of secrets and Y begin the set of sample histograms.

Proof.

(i) **Adversary $\mathcal{A}^{in}$ and prior distribution π^{in}:**

$$V_g(\delta^{in,y}) = \max_{w \in \mathcal{W}} \sum_{(p,t)\in\mathcal{X}} Pr[(p,t)|y] \cdot g(w,(p,t))$$

Bayes' theorem:

$$= \max_{w \in \mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \frac{Pr[(p,t)]Pr[y|(p,t)]}{Pr[y]} \cdot g(w,(p,t))$$

Def. of π^{in}, S, g and by Lemma 4.1.5:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t=w}} \frac{\mathsf{S}_{x,y} \cdot (m+1)}{m(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$

$$= \frac{m+1}{m(n+1)} \cdot \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t=w,\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}$$

Split cases when $w{=}\mathtt{a}$ and $w{=}\mathtt{b}$:

$$= \frac{m+1}{m(n+1)} \cdot \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t=\mathtt{a},\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}, \sum_{\substack{(p,t)\in\mathcal{X}:\\ 1\leq t\leq m,\\ p_t=\mathtt{b},\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} \right\} \quad \text{(A.9)}$$

We need to count how many secrets $(p,t){\in}\mathcal{X}$ satisfy the restrictions $1{\leq}t{\leq}m \wedge p_t{=}\mathtt{a}\wedge$ $n_{\mathtt{a}}(p_{1\ldots m}){=}y$ in the left summation inside the max and $1{\leq}t{\leq}m \wedge p_t{=}\mathtt{b} \wedge n_{\mathtt{a}}(p_{1\ldots m}){=}y$ in the right summation inside the max. Each secret x is a tuple (p,t), where p is the population array and t is the target's index.

- In the left summation, in the first m elements of p there are y $\mathtt{a}$'s, and as $p_t{=}\mathtt{a}$, there will be $y-1$ $\mathtt{a}$'s in the other $m-1$ positions, so we have $\binom{m-1}{y-1}$ possible combinations.
- In the right summation the reasoning is similar the the left one, except that now $p_t = \mathtt{b}$, so there will be y $\mathtt{a}$'s in the other $m-1$ positions, so $\binom{m-1}{y}$ possible combinations.

For both summations the other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, so $\binom{n-m}{y'}$ different combinations.. Finally, $n_{\mathtt{a}}(p_{1\ldots m}) = y$ and $n_{\mathtt{a}}(p_{m+1\ldots n}) = y'$ implies $n_{\mathtt{a}}(p) = y + y'$.

$$= \frac{m+1}{m(n+1)} \cdot \max \left\{ \sum_{t=1}^{m} \sum_{y'=0}^{n-m} \binom{m-1}{y-1} \binom{n-m}{y'} \binom{n}{y+y'}^{-1}, \sum_{t=1}^{m} \sum_{y'=0}^{n-m} \binom{m-1}{y} \binom{n-m}{y'} \binom{n}{y+y'}^{-1} \right\} \tag{A.10}$$

By Lemma 4.1.1:

$$= \frac{m+1}{m(n+1)} \cdot \max \left\{ \sum_{t=1}^{m} \frac{y(n+1)}{m(m+1)}, \sum_{t=1}^{m} \frac{(m-y)(n+1)}{m(m+1)} \right\}$$
$$= \frac{1}{m} \cdot \max \left\{ \sum_{t=1}^{m} \frac{y}{m}, \sum_{t=1}^{m} \frac{m-y}{m} \right\}$$
$$= \frac{1}{m} \cdot \max \left\{ m \cdot \frac{y}{m}, m \cdot \frac{m-y}{m} \right\}$$
$$= \max \left\{ \frac{y}{m}, \frac{m-y}{m} \right\}.$$

(ii) **Adversary $\mathcal{A}^{out}$ and prior distribution π^{out}:**

$$V_g(\delta^{out,y}) = \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} Pr[(p,t)|y] \cdot g(w,(p,t))$$

Bayes' theorem:

$$= \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \frac{Pr[(p,t)]Pr[y|(p,t)]}{Pr[y]} \cdot g(w,(p,t))$$

Def. of π^{out}, S, g and by Lemma 4.1.5:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t) \in \mathcal{X}: \\ m+1 \le t \le n, \\ p_t = w}} \frac{\mathsf{S}_{x,y} \cdot (m+1)}{(n-m)(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$
$$= \frac{m+1}{(n-m)(n+1)} \cdot \max_{w \in \mathcal{W}} \sum_{\substack{(p,t) \in \mathcal{X}: \\ m+1 \le t \le n, \\ p_t = w, \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}$$

Split cases when w=$\mathtt{a}$ and w=$\mathtt{b}$:

$$= \frac{m+1}{(n-m)(n+1)} \cdot \max \left\{ \sum_{\substack{(p,t) \in \mathcal{X}: \\ m+1 \le t \le n, \\ p_t = \mathtt{a}, \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}, \sum_{\substack{(p,t) \in \mathcal{X}: \\ m+1 \le t \le n, \\ p_t = \mathtt{b}, \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} \right\}$$

We need to count how many secrets $(p,t)\in\mathcal{X}$ satisfy the restrictions $m{+}1{\leq}t{\leq}n \wedge p_t{=}\mathtt{a} \wedge n_{\mathtt{a}}(p_{1\ldots m}){=}y$ in the left summation inside the max and $m{+}1{\leq}t{\leq}n \wedge p_t{=}\mathtt{b} \wedge n_{\mathtt{a}}(p_{1\ldots m}){=}y$ in the right summation inside the max. Each secret x is a tuple (p,t), where p is the population array and t is the target's index. In the first m elements of p, there are y $\mathtt{a}$'s, so $\binom{m}{y}$ possible combinations. For the other $n{-}m$ people in $p_{m+1\ldots n}$, they can have any value (except by p_t), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$ except by p_t, therefore y' goes from 0 to $n{-}m{-}1$, then $\binom{n-m-1}{y'}$ possible combinations. Also:

- In the left summation, as $p_t = \mathtt{a}$, $n_{\mathtt{a}}(p) = y + y' + 1$.
- In the left summation, as $p_t = \mathtt{b}$, $n_{\mathtt{a}}(p) = y + y'$.

$$= \frac{m+1}{(n-m)(n+1)} \cdot \max\left\{ \sum_{t=m+1}^{n} \sum_{y'=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{y'}\binom{n}{y+y'+1}^{-1}, \sum_{t=m+1}^{n} \sum_{y'=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{y'}\binom{n}{y+y'}^{-1} \right\} \tag{A.11}$$

By Lemma 4.1.3:

$$\begin{aligned} &= \frac{m+1}{(n-m)(n+1)} \cdot \max\left\{ \sum_{t=m+1}^{n} \frac{(n+1)(y+1)}{(m+1)(m+2)}, \sum_{t=m+1}^{n} \frac{(n+1)(m-y+1)}{(m+1)(m+2)} \right\} \\ &= \frac{1}{n-m} \cdot \max\left\{ \frac{(n-m)(y+1)}{m+2}, \frac{(n-m)(m-y+1)}{m+2} \right\} \\ &= \max\left\{ \frac{y+1}{m+2}, \frac{m-y+1}{m+2} \right\}. \end{aligned} \tag{A.12}$$

(iii) **Adversary $\mathcal{A}^{nk}$ and prior distribution π^{nk}:**

$$V_g(\delta^{nk,y}) = \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} Pr[(p,t)|y] \cdot g(w,(p,t))$$

Bayes' theorem:

$$= \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \frac{Pr[(p,t)]Pr[y|(p,t)]}{Pr[y]} \cdot g(w,(p,t))$$

Def. of π^{nk}, S, g and by Lemma 4.1.5:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}: \\ p_t = w}} \frac{\mathsf{S}_{x,y} \cdot (m+1)}{n(n+1)\binom{n}{n_{\mathtt{a}}(p)}}$$

$$= \frac{m+1}{n(n+1)} \cdot \max_{w \in \mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}: \\ p_t = w \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}$$

Split cases when w=$\mathtt{a}$ and w=$\mathtt{b}$:

$$= \frac{m+1}{n(n+1)} \cdot \max \left\{ \sum_{\substack{(p,t)\in\mathcal{X}: \\ p_t = \mathtt{a} \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1}, \sum_{\substack{(p,t)\in\mathcal{X}: \\ p_t = \mathtt{b} \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} \right\}$$

We need to count how many secrets $(p,t)\in\mathcal{X}$ satisfy the restrictions p_t=$\mathtt{a} \wedge n_{\mathtt{a}}(p_{1\ldots m})$=$y$ in the left summation inside the max and p_t=$\mathtt{b}$ $\wedge$ $n_{\mathtt{a}}(p_{1\ldots m})$=$y$ in the right summation inside the max. Each secret x is a tuple (p,t), where p is the population array and t is the target's index. We can divide the counting in four cases:

(1) The adversary's guess is $\mathtt{a}$ and $1 \leq t \leq m$,

(2) The adversary's guess is $\mathtt{a}$ and $m{+}1 \leq t \leq n$,

(3) The adversary's guess is $\mathtt{b}$ and $1 \leq t \leq m$, and

(4) The adversary's guess is $\mathtt{b}$ and $m{+}1 \leq t \leq n$.

For cases (1) and (3) we can use the same reasoning we have used in Equation (A.10), and for cases (2) and (4) we can use the same reasoning we have used in Equation (A.11). The sum of summations on the left inside the max represents cases (1) and (2), and the sum of summations on the right inside the max represents cases (3) and (4).

$$
\begin{aligned}
&= \frac{m+1}{n(n+1)} \cdot \max\left\{ \sum_{t=1}^{m}\sum_{y'=0}^{n-m} \binom{m-1}{y-1}\binom{n-m}{y'}\binom{n}{y+y'}^{-1} + \right. \\
&\qquad \sum_{t=m+1}^{n}\sum_{y'=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{y'}\binom{n}{y+y'+1}^{-1}, \\
&\qquad \sum_{t=1}^{m}\sum_{y'=0}^{n-m} \binom{m-1}{y}\binom{n-m}{y'}\binom{n}{y+y'}^{-1} + \\
&\qquad \left. \sum_{t=m+1}^{n}\sum_{y'=0}^{n-m-1} \binom{m}{y}\binom{n-m-1}{y'}\binom{n}{y+y'}^{-1} \right\} \\
&= \frac{1}{n} \cdot \max\left\{ m \cdot \frac{y}{m} + \frac{(n-m)(y+1)}{m+2}, m \cdot \frac{m-y}{m} + \frac{(n-m)(m-y+1)}{m+2} \right\} \\
&= \frac{1}{n(m+2)} \cdot \max\big\{ y(m+2) + (n-m)(y+1), \\
&\qquad (m-y)(m+2) + (n-m)(m-y+1) \big\} \\
&= \frac{1}{n(m+2)} \cdot \max\big\{ my + 2y + ny + n - my - m, \\
&\qquad m^2 + 2m - my - 2y + nm - ny + n - m^2 + my - m \big\} \\
&= \frac{1}{n(m+2)} \cdot \max\{ n + ny + 2y - m, n + nm - (ny + 2y - m) \} \\
&= \frac{n + \max\{ ny + 2y - m, nm - (ny + 2y - m) \}}{n(m+2)}.
\end{aligned}
$$

□

Lemma 4.1.7 (Marginal on $\mathcal{Y}$ for $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$)**.** *Given the set of secrets $\mathcal{X}$, the prior distributions $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ and channel S, the marginal probability distribution on $\mathcal{Y}$ is*

$$
Pr[y] = \binom{m}{y} 2^{-m}. \tag{4.22}
$$

Proof. Let's use the prior distribution $\hat{\pi}^{in}$ to construct the proof.

$$
\begin{aligned}
Pr[y] &= \sum_{(p,t)\,\in\,\mathcal{X}} Pr[(p,t)] Pr[y|(p,t)] \\
&= \sum_{(p,t)\,\in\,\mathcal{X}} \hat{\pi}^{in} \cdot \mathsf{S}_{x,y} \\
&= \sum_{\substack{(p,t)\,\in\,\mathcal{X}: \\ 1 \le t \le m, \\ n_{\mathtt{a}}(x^{p}_{1\ldots m}) = y}} \frac{1}{m 2^n}
\end{aligned}
$$

We need to count how many secrets $(p,t)\in\mathcal{X}$ satisfy the restrictions $1\leq t\leq m$ and $n_{\mathtt{a}}(p_{1\ldots m}) = y$. Each secret x is a tuple (p,t), where p is the population array and t is the target's index. In the first m elements of p we have y $\mathtt{a}$'s, so $\binom{m}{y}$ different combinations. The other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, so $\binom{n-m}{y'}$ different combinations.

$$\begin{aligned}
&= \frac{1}{m2^n}\sum_{t=1}^{m}\sum_{y'=0}^{n-m}\binom{m}{y}\binom{n-m}{y'} \\
&= \frac{1}{m2^n}\cdot m\binom{m}{y}\sum_{y'=0}^{n-m}\binom{n-m}{y'} \\
&= \frac{1}{2^n}\cdot\binom{m}{y}2^{n-m} \\
&= \binom{m}{y}2^{-m}\ .
\end{aligned} \tag{A.13}$$

Note that the proof above is also valid for prior distributions $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$. The only difference between these three priors are the range of t, that are $\{1,\ldots,m\}$, $\{m{+}1,\ldots,n\}$ and $\{1,\ldots,n\}$, for $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$, respectively. The size of these ranges are all canceled out in Equation (A.13). □

Lemma 4.1.8 (Vulnerability of a specific output y)**.** *Let $\mathcal{X}$ be the set of secrets, $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ be prior distributions on $\mathcal{X}$, g be the gain function for attribute inference attack and S be the channel. Given that the adversary observed some output y, the posterior vulnerability given y is*

(i)

$$V_g(\hat{\delta}^{in,y}) = \max\left\{\frac{y}{m}, \frac{m-y}{m}\right\}, \tag{4.23}$$

(ii)

$$V_g(\hat{\delta}^{out,y}) = \frac{1}{2}, \tag{4.24}$$

(iii)

$$V_g(\hat{\delta}^{nk,y}) = \frac{1}{n}\left(\frac{n-m}{2} + \max\{y, m-y\}\right). \tag{4.25}$$

where $\hat{\delta}^{in,y}$, $\hat{\delta}^{out,y}$ and $\hat{\delta}^{nk,y}$ are the inner distributions when $\hat{\pi}^{in}$, $\hat{\pi}^{out}$ and $\hat{\pi}^{nk}$ are, respectively, the prior distributions, and when y is observed, i.e., $Pr[X \mid Y = y]$.

Proof.

(i) **Adversary $\mathcal{A}^{in}$ and prior distribution $\hat{\pi}^{in}$:**

$$V_g(\hat{\delta}^{in,y}) = \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} Pr[(p,t)|y] \cdot g(w,(p,t))$$

Bayes' theorem:

$$= \max_{w \in \mathcal{W}} \sum_{(p,t) \in \mathcal{X}} \frac{Pr[(p,t)]Pr[y|(p,t)]}{Pr[y]} \cdot g(w,(p,t))$$

Def. of $\hat{\pi}^{in}$, S, g and by Lemma 4.1.7:

$$= \max_{w \in \mathcal{W}} \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = w}} \frac{2^m \cdot \mathsf{S}_{(p,t),y}}{m 2^n \binom{m}{y}}$$

$$= \frac{2^m}{m 2^n \binom{m}{y}} \cdot \max_{w \in \mathcal{W}} \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = w, \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} 1$$

Split cases when w=$\mathtt{a}$ and w=$\mathtt{b}$:

$$= \frac{2^m}{m 2^n \binom{m}{y}} \cdot \max \left\{ \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = \mathtt{a}, \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} 1, \sum_{\substack{(p,t) \in \mathcal{X}: \\ 1 \leq t \leq m, \\ p_t = \mathtt{b}, \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} 1 \right\}$$

We need to count how many secrets $(p,t) \in \mathcal{X}$ satisfy the restrictions $1 \leq t \leq m \wedge p_t = \mathtt{a} \wedge n_{\mathtt{a}}(p_{1\ldots m}) = y$ in the left summation inside the max and $1 \leq t \leq m \wedge p_t = \mathtt{b} \wedge n_{\mathtt{a}}(p_{1\ldots m}) = y$ in the right summation inside the max.

- In the left summation, in the first m elements of p there are y $\mathtt{a}$'s, and as $p_t = \mathtt{a}$, there will be $y-1$ $\mathtt{a}$'s in the other $m-1$ positions, so we have $\binom{m-1}{y-1}$ possible combinations.
- In the right summation the reasoning is similar the the left one, except that now $p_t = \mathtt{b}$, so there will be y $\mathtt{a}$'s in the other $m-1$ positions, so $\binom{m-1}{y}$ possible combinations.

For both summations the other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, so $\binom{n-m}{y'}$ different combinations.

$$
\begin{aligned}
&= \frac{2^m}{m2^n\binom{m}{y}} \cdot \max\left\{ \sum_{t=1}^{m}\sum_{y'=0}^{n-m} \binom{m-1}{y-1}\binom{n-m}{y'}, \right. \\
&\qquad\qquad \left. \sum_{t=1}^{m}\sum_{y'=0}^{n-m} \binom{m-1}{y}\binom{n-m}{y'} \right\} \\
&= \frac{2^m}{m2^n\binom{m}{y}} \cdot \max\left\{ m\binom{m-1}{y-1}\sum_{y'=0}^{n-m} \binom{n-m}{y'}, \right. \\
&\qquad\qquad \left. m\binom{m-1}{y}\sum_{y'=0}^{n-m} \binom{n-m}{y'} \right\} \\
&= \frac{2^m}{m2^n\binom{m}{y}} \max\left\{ m\binom{m-1}{y-1}2^{n-m}, m\binom{m-1}{y}2^{n-m} \right\} \\
&= \frac{1}{\binom{m}{y}} \max\left\{ \binom{m-1}{y-1}, \binom{m-1}{y} \right\} \\
&= \frac{1}{\binom{m}{y}} \max\left\{ \binom{m}{y}\cdot\frac{y}{m}, \binom{m}{y}\cdot\frac{m-y}{m} \right\} \\
&= \max\left\{ \frac{y}{m}, \frac{m-y}{m} \right\} .
\end{aligned}
\tag{A.14}
$$

(ii) **Adversary $\mathcal{A}^{out}$ and prior distribution $\hat{\pi}^{out}$:**

$$V_g(\hat{\delta}^{out,y}) = \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} Pr[(p,t)|y]\cdot g(w,(p,t))$$

Bayes' theorem:

$$= \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \frac{Pr[(p,t)]Pr[y|(p,t)]}{Pr[y]}\cdot g(w,(p,t))$$

Def. of $\hat{\pi}^{out}$, S, g and by Lemma 4.1.7:

$$
\begin{aligned}
&= \max_{w\in\mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=w}} \frac{2^m\cdot \mathsf{S}_{(p,t),y}}{(n-m)2^n\binom{m}{y}} \\
&= \frac{2^m}{(n-m)2^n\binom{m}{y}} \cdot \max_{w\in\mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=w,\\ n_{\mathsf{a}}(p_{1\ldots m})=y}} 1
\end{aligned}
$$

Split cases when w=a and w=b:

$$
= \frac{2^m}{(n-m)2^n\binom{m}{y}} \cdot \max\left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=\mathsf{a},\\ n_{\mathsf{a}}(p_{1\ldots m})=y}} 1, \sum_{\substack{(p,t)\in\mathcal{X}:\\ m<t\leq n,\\ p_t=\mathsf{b},\\ n_{\mathsf{a}}(p_{1\ldots m})=y}} 1 \right\}
$$

We need to count how many secrets $(p,t)\in\mathcal{X}$ satisfy the restrictions $m<t\leq n \wedge p_t=\mathtt{a}\wedge n_{\mathtt{a}}(p_{1\ldots m})=y$ in the left summation inside the max and $m<t\leq n \wedge p_t=\mathtt{b}\wedge n_{\mathtt{a}}(p_{1\ldots m})=y$ in the right summation inside the max. In the first m elements of p, there are y $\mathtt{a}$'s, so $\binom{m}{y}$ possible combinations. For the other $n-m$ people in $p_{m+1\ldots n}$, they can have any value (except by p_t), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$ except by p_t, therefore y' goes from 0 to $n-m-1$, then $\binom{n-m-1}{y'}$ possible combinations.

$$\begin{aligned}
&= \frac{2^m}{(n-m)2^n\binom{m}{y}} \cdot \max\left\{ \sum_{t=m+1}^{n}\sum_{y'=0}^{n-m-1}\binom{m}{y}\binom{n-m-1}{y'}, \right. \\
&\qquad\qquad \left. \sum_{t=m+1}^{n}\sum_{y'=0}^{n-m-1}\binom{m}{y}\binom{n-m-1}{y'}\right\} \\
&= \frac{2^m}{(n-m)2^n\binom{m}{y}} \cdot \max\left\{ (n-m)\binom{m}{y}\sum_{y'=0}^{n-m-1}\binom{n-m-1}{y'}, \right. \\
&\qquad\qquad \left. (n-m)\binom{m}{y}\sum_{y'=0}^{n-m-1}\binom{n-m-1}{y'}\right\} \\
&= \frac{1}{2^{n-m}} \cdot \max\{2^{n-m-1}, 2^{n-m-1}\} \\
&= \frac{1}{2} \,.
\end{aligned} \tag{A.15}$$

(iii) **Adversary $\mathcal{A}^{nk}$ and prior distribution $\hat{\pi}^{nk}$:**

$$V_g(\hat{\delta}^{nk,y}) = \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} Pr[(p,t)|y]\cdot g(w,(p,t))$$

Baye's theorem:

$$= \max_{w\in\mathcal{W}} \sum_{(p,t)\in\mathcal{X}} \frac{Pr[(p,t)]Pr[y|(p,t)]}{Pr[y]}\cdot g(w,(p,t))$$

Def. of $\hat{\pi}^{nk}$, S, g and by Lemma 4.1.7:

$$\begin{aligned}
&= \max_{w\in\mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_t=w}} \frac{2^m\cdot \mathsf{S}_{(p,t),y}}{n2^n\binom{m}{y}} \\
&= \frac{2^m}{n2^n\binom{m}{y}} \cdot \max_{w\in\mathcal{W}} \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_t=w\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} 1
\end{aligned}$$

Split cases when $w=\mathtt{a}$ and $w=\mathtt{b}$:

$$= \frac{2^m}{n2^n\binom{m}{y}} \cdot \max\left\{ \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_t=\mathtt{a}\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} 1, \sum_{\substack{(p,t)\in\mathcal{X}:\\ p_t=\mathtt{b}\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} 1 \right\}$$

We need to count how many secrets $(p,t)\in\mathcal{X}$ satisfy the restrictions $p_t{=}\mathtt{a} \wedge n_{\mathtt{a}}(p_{1\ldots m}){=}y$ in the left summation inside the max and $p_t{=}\mathtt{b} \wedge n_{\mathtt{a}}(p_{1\ldots m}){=}y$ in the right summation inside the max. We can divide the counting in four cases:

(1) The adversary's guess is $\mathtt{a}$ and $1 \leq t \leq m$,

(2) The adversary's guess is $\mathtt{a}$ and $m < t \leq n$,

(3) The adversary's guess is $\mathtt{b}$ and $1 \leq t \leq m$, and

(4) The adversary's guess is $\mathtt{b}$ and $m < t \leq n$.

For cases (1) and (3) we can use the same reasoning we have used in Equation (A.14), and for cases (2) and (4) we can use the same reasoning we have used in Equation (A.15). The sum of summations on the left inside the max represents cases (1) and (2), and the sum of summations on the right inside the max represents cases (3) and (4).

$$
\begin{aligned}
&= \frac{2^m}{n2^n\binom{m}{y}} \cdot \max\left\{ \sum_{t=1}^{m}\sum_{y'=0}^{n-m}\binom{m-1}{y-1}\binom{n-m}{y'} + \right. \\
&\qquad\qquad \sum_{t=m+1}^{n}\sum_{y'=0}^{n-m-1}\binom{m}{y}\binom{n-m-1}{y'}, \\
&\qquad\qquad \sum_{t=1}^{m}\sum_{y'=0}^{n-m}\binom{m-1}{y}\binom{n-m}{y'} + \\
&\qquad\qquad \left. \sum_{t=m+1}^{n}\sum_{y'=0}^{n-m-1}\binom{m}{y}\binom{n-m-1}{y'} \right\} \\
&= \frac{2^m}{n2^n\binom{m}{y}} \cdot \max\left\{ m\binom{m-1}{y-1}2^{n-m} + (n-m)\binom{m}{y}2^{n-m-1}, \right. \\
&\qquad\qquad \left. m\binom{m-1}{y}2^{n-m} + (n-m)\binom{m}{y}2^{n-m-1} \right\} \\
&= \frac{2^m}{n2^n\binom{m}{y}} \cdot \max\left\{ m\binom{m}{y}\frac{y}{m}\cdot 2^{n-m} + (n-m)\binom{m}{y}2^{n-m-1}, \right. \\
&\qquad\qquad \left. m\binom{m}{y}\frac{m-y}{y}\cdot 2^{n-m} + (n-m)\binom{m}{y}2^{n-m-1} \right\} \\
&= \frac{1}{n} \cdot \max\left\{ y + \frac{n-m}{2}, m - y + \frac{n-m}{2} \right\} \\
&= \frac{1}{n}\left(\frac{n-m}{2} + \max\{y, m-y\} \right) \ .
\end{aligned}
$$

□

Appendix B

Proofs of Lemmas – Utility

Here we present the proofs of all lemmas related to utility analysis in Section 4.2.

Lemma 4.2.1 (Guessing symmetry when n is even)**.** *Let $p \geq 1$ and $0 \leq k \leq 2p$. Let also*

$$f(k) = k^2 - 2kp \ . \tag{4.30}$$

We have that

$$f(k) = f(2p - k).$$

Proof.

$$\begin{aligned} f(2p-k) &= (2p-k)^2 - 2p(2p-k) \\ &= 4p^2 - 4kp + k^2 - 4p^2 + 2kp \\ &= k^2 - 2kp \\ &= f(k) \ . \end{aligned}$$

□

Lemma 4.2.2 (Guessing symmetry when n is odd)**.** *Let $p \geq 1$ and $0 \leq k \leq 2p+1$. Let also*

$$f'(k) = k^2 - 2kp - k \ . \tag{4.31}$$

We have that

$$f'(k) = f'(2p + 1 - k).$$

Proof.

$$\begin{aligned} f'(2p+1-k) &= (2p+1-k)^2 - 2p(2p+1-k) - (2p+1-k) \\ &= 4p^2 + 2p - 2kp + 2p + 1 - k - 2kp - k + k^2 \\ &\quad - 4p^2 - 2p + 2kp - 2p - 1 + k \\ &= k^2 - 2kp - k \\ &= f'(k) \ . \end{aligned}$$

□

Lemma 4.2.3 (Sum of differences when n is even)**.** *Let $n \geq 2$ be even. We have that*

$$\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| = \frac{n(n+2)}{4} \ , \tag{4.32}$$

where the minimum in Equation (4.32) happens when $k = \frac{n}{2}$.

Proof.

$$\begin{aligned}
\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| &= \min_{0 \leq k \leq n} \left(\sum_{i=0}^{k} (k - i) + \sum_{i=k+1}^{n} (i - k) \right) \\
&= \min_{0 \leq k \leq n} \left(k \sum_{i=0}^{k} 1 - \sum_{i=0}^{k} i + \sum_{i=k+1}^{n} i - k \sum_{i=k+1}^{n} 1 \right)
\end{aligned}$$

Solving arithmetic progressions:

$$\begin{aligned}
&= \min_{0 \leq k \leq n} \left(k(k+1) - \frac{k(k+1)}{2} + \frac{(k+1+n)(n-k)}{2} - k(n-k) \right) \\
&= \min_{0 \leq k \leq n} \left(\frac{k(k+1)}{2} + \frac{kn - k^2 + n - k + n^2 - kn}{2} - \frac{2k(n-k)}{2} \right) \\
&= \min_{0 \leq k \leq n} \left(\frac{k^2 + k - k^2 + n - k + n^2 - 2kn + 2k^2}{2} \right) \\
&= \min_{0 \leq k \leq n} \left(\frac{n^2 + 2k^2 + n - 2kn}{2} \right)
\end{aligned}$$

Because n is constant:

$$= \min_{0 \leq k \leq n} \left(k^2 - kn\right) + \frac{n(n+1)}{2}. \tag{B.1}$$

Rewriting:

$$\begin{aligned}
\min_{0 \leq k \leq n} \left(k^2 - kn\right) + \frac{n(n+1)}{2} &= \frac{n(n+2)}{4} \\
\Leftrightarrow \min_{0 \leq k \leq n} \left(k^2 - kn\right) &= -\frac{n^2}{4}.
\end{aligned} \tag{B.2}$$

Looking at Equation (B.2), as n is even, let $n = 2p$ for some $p \in \mathbb{N}$. As proposed before, the minimum will happen when $k = {}^{n}/_{2} = p$. Thus we want to show:

(i) $\forall 0 \leq k < p : k^2 - 2kp \geq -p^2$, and

(ii) $\forall p < k \leq 2p : k^2 - 2kp \geq -p^2$.

For (i), let us prove by induction on p.

Base case $p = 1$

$$\forall 0 \leq k < 1 : k^2 - 2k \geq -1.$$
$$(k = 0) \Rightarrow 0^2 - 2 \cdot 0 = 0 \geq -1.$$

Induction step Assume $\forall 0 \leq k < p : k^2 - 2kp \geq -p^2$. We want to show that $\forall 0 \leq k < p+1 : k^2 - 2k(p+1) \geq -(p+1)^2$.

$$\begin{aligned} k^2 - 2k(p+1) &= k^2 - 2kp - 2k \\ &\geq -p^2 - 2k && \text{(by I.H)} \\ &\geq -p^2 - 2p - 1 && (0 \leq k < p+1) \\ &= -(p+1)^2. \end{aligned}$$

For (ii), and assuming that $f(k) = k^2 - 2kp$, we want to show that

$$\forall p < k \leq 2p : f(k) \geq -p^2.$$

We have already proved (i), that states

$$f(0) \geq -p^2, f(1) \geq -p^2, \ldots, f(p-1) \geq -p^2. \tag{B.3}$$

Using Lemma 4.2.1, we can rewrite Equation (B.3) as

$$f(2p) \geq -p^2, f(2p-1) \geq -p^2, \ldots, f(p+1) \geq -p^2,$$

which is the same thing as saying that

$$\forall p < k \leq 2p : f(k) \geq -p^2,$$

which was exactly what we wanted to prove. Therefore, proving (i) and (ii), we have shown that

$$\min_{0 \leq k \leq n} \left(k^2 - kn\right) = -\frac{n^2}{4},$$

which implies

$$\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| = \frac{n(n+2)}{4}.$$

□

Lemma 4.2.4 (Sum of differences when n id odd)**.** *Let $n \geq 1$ be odd. We have that*

$$\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| = \frac{(n+1)^2}{4}, \tag{4.33}$$

where the minimum in Equation (4.33) happens when $k = \frac{n-1}{2}$.

Proof.

By the same derivation done for Equation (B.1):

$$\min_{0 \leq k \leq n} \sum_{i=0}^{n} |k - i| = \min_{0 \leq k \leq n} \left(k^2 - kn\right) + \frac{n(n+1)}{2}. \tag{B.4}$$

Rewriting:

$$\min_{0\leq k\leq n}\left(k^2-kn\right)+\frac{n(n+1)}{2}=\frac{(n+1)^2}{4}$$
$$\Leftrightarrow \min_{0\leq k\leq n}\left(k^2-kn\right)=-\frac{(n^2-1)}{4}. \tag{B.5}$$

Looking at Equation (B.5), as n is odd, let $n=2p+1$ for some $p\in\mathbb{N}$. As proposed before, the minimum will happen when $k=\frac{n-1}{2}=p$. Thus we want to show:

(i) $\forall 0\leq k<p: k^2-k(2p+1)\geq -p(p+1)$, and

(ii) $\forall p<k\leq 2p+1: k^2-k(2p+1)\geq -p(p+1)$.

For (i), let us prove by induction on p.

Base case $p=1$

$$\forall 0\leq k<1: k^2-3k\geq -2.$$
$$(k=0)\Rightarrow 0^2-3\cdot 0=0\geq -2.$$

Induction step Assume $\forall 0\leq k<p: k^2-k(2p+1)=k^2-2kp-k\geq -p(p+1)$. We want to show that $\forall 0\leq k<p+1: k^2-k(2p+3)\geq -p^2-3p-2$.

$$\begin{aligned} k^2-k(2p+3)&=k^2-2kp-3k\\ &\geq -p^2-p-2k && \text{(by I.H)}\\ &\geq -p^2-3p-1. && (0\leq k<p+1)\end{aligned}$$

For (ii), and assuming that $f'(k)=k^2-2kp-k$, we want to show that

$$\forall p<k\leq 2p+1: f'(k)\geq -p(p+1).$$

We have already proved (i), that states

$$f'(0)\geq -p(p+1),\ f'(1)\geq -p(p+1),\ \ldots,f'(p-1)\geq -p(p+1). \tag{B.6}$$

Using Lemma 4.2.2, we can rewrite Equation (B.6) as

$$f'(2p+1)\geq -p(p+1),\ f'(2p)\geq -p(p+1),\ \ldots,\ f'(p+2)\geq -p(p+1). \tag{B.7}$$

Also note that $f'(p+1)=-p(p+1)\geq -p(p+1)$. Thus we are saying that

$$\forall p<k\leq 2p+1: f'(k)\geq -p(p+1),$$

which was exactly what we wanted to prove. Therefore, proving (i) and (ii), we have shown that

$$\min_{0\leq k\leq n}\left(k^2-kn\right)=-\frac{(n^2-1)}{4}\ ,$$

which implies

$$\min_{0\leq k\leq n}\sum_{i=0}^{n}|k-i|=\frac{(n+1)^2}{4}.$$

□

Lemma 4.2.5 (Marginal on $\mathcal{Y}$ for π^{ut})**.** *Given the prior distribution π^{ut} on the set of secrets $\mathcal{X}^{ut}$ and the channel S^{ut}, the probability of a sample's histogram $y\in\mathcal{Y}$ being the output is*

$$Pr[y]=\frac{1}{m+1}\ . \tag{4.35}$$

Proof.

$$\begin{aligned}
Pr[y] &= \sum_{p\in\mathcal{X}} Pr[p]Pr[y|p] && \text{(B.8)}\\
&= \sum_{p\,\in\,\mathcal{X}^{ut}} \frac{\mathsf{S}^{ut}_{p,y}}{(n+1)\binom{n}{n_{\mathtt{a}}(p)}} && \text{(Def. of } \pi^{ut} \text{ and } \mathsf{S}^{ut})\\
&= \frac{1}{n+1}\sum_{\substack{p\,\in\,\mathcal{X}^{ut}:\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} && \text{(Def. of } \mathsf{S}^{ut})
\end{aligned}$$

We need to count how many secrets $p\in\mathcal{X}^{ut}$ satisfy the restriction $n_{\mathtt{a}}(p_{1\ldots m})=y$. In the first m elements of x we have y $\mathtt{a}$'s, so $\binom{m}{y}$ different combinations. The other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, thus $\binom{n-m}{y'}$ different combinations. Finally, $n_{\mathtt{a}}(p_{1\ldots m})=y$ and $n_{\mathtt{a}}(p_{m+1\ldots n})=y'$ implies $n_{\mathtt{a}}(p)=y+y'$.

$$= \frac{1}{n+1}\sum_{y'=0}^{n-m}\binom{m}{y}\binom{n-m}{y'}\binom{n}{y+y'}^{-1}$$

By Lemma 4.1.1:

$$\begin{aligned}
&= \frac{1}{n+1}\cdot\frac{n+1}{m+1}\\
&= \frac{1}{m+1}.
\end{aligned}$$

□

Lemma 4.2.6 (Utility loss for a specific output y)**.** *Let π^{ut} be a prior distribution on the set of secrets $\mathcal{X}^{ut}$, g be the gain function for attribute inference attack and S^{ut} be*

the channel. Given that the adversary observed some output y, the posterior vulnerability given y is

$$U_\ell(\delta^{y,ut}) = \frac{m+1}{n(n+1)} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y}\binom{n-m}{y'}\binom{n}{y+y'}^{-1} \cdot |k-y-y'| \ .$$

where $\delta^{y,ut} \in \mathbb{D}\mathcal{X}^{ut}$ is the inner distribution when π^{ut} is the prior distribution and y is observed (i.e., $Pr[X|Y=y]$).

Proof.

$$U_\ell(\delta^{y,ut}) = \min_{w\in\mathcal{W}} \sum_{p\in\mathcal{X}^{ut}} Pr[(p,t)|y] \cdot \ell(w,p)$$

Baye's theorem:

$$= \min_{w\in\mathcal{W}} \sum_{p\in\mathcal{X}^{ut}} \frac{Pr[p]Pr[y|p]}{Pr[y]} \cdot \ell(w,p)$$

Def. of π^{ut}, S^{ut}, ℓ and by Lemma 4.2.5.

$$= \min_{w\in\mathcal{W}} \sum_{p\in\mathcal{X}^{ut}} \frac{(m+1)\mathsf{S}^{ut}_{p,y}}{(n+1)\binom{n}{n_{\mathtt{a}}(p)}} \cdot \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right|$$

Definition of S^{ut}:

$$= \frac{m+1}{n+1} \min_{w\in\mathcal{W}} \sum_{\substack{p\in\mathcal{X}^{ut}:\\ n_{\mathtt{a}}(p_{1\ldots m})=y}} \binom{n}{n_{\mathtt{a}}(p)}^{-1} \cdot \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right|$$

We need to count how many secrets $p\in\mathcal{X}^{ut}$ satisfy the restriction $n_{\mathtt{a}}(p_{1\ldots m}) = y$. In the first m elements of x we have y $\mathtt{a}$'s, so $\binom{m}{y}$ different combinations. The other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, thus $\binom{n-m}{y'}$ different combinations. Finally, $n_{\mathtt{a}}(p_{1\ldots m}) = y$ and $n_{\mathtt{a}}(p_{m+1\ldots n}) = y'$ implies $n_{\mathtt{a}}(p) = y+y'$.

$$= \frac{m+1}{n+1} \min_{w\in\mathcal{W}} \sum_{y'=0}^{n-m} \binom{m}{y}\binom{n-m}{y'}\binom{n}{y+y'}^{-1} \cdot \left| w - \frac{y+y'}{n} \right|$$

The set of actions $\mathcal{W} = \{{}^0\!/_n, \ldots, {}^n\!/_n\}$, but we can rewrite it in terms of an integer $0 \leq k \leq n$ such that $\mathcal{W} = \{{}^k\!/_n \mid 0\leq k\leq n\}$ and rewrite $\min_{w\in\mathcal{W}}$ in terms of k:

$$= \frac{m+1}{n+1} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y}\binom{n-m}{y'}\binom{n}{y+y'}^{-1} \cdot \left| \frac{k}{n} - \frac{y+y'}{n} \right|$$

Because n is constant:

$$= \frac{m+1}{n(n+1)} \min_{0\leq k\leq n} \sum_{y'=0}^{n-m} \binom{m}{y}\binom{n-m}{y'}\binom{n}{y+y'}^{-1} \cdot |k-y-y'| \ .$$

□

Lemma 4.2.7 (Marginal on $\mathcal{Y}$ for $\hat{\pi}^{ut}$)**.** *Given the set of secrets $\mathcal{X}^{ut}$, the prior $\hat{\pi}^{ut}$, the loss function ℓ and channel S^{ut}, we have that the marginal distribution on outputs $\mathcal{Y}$ is*

$$Pr[y] = \binom{m}{y} 2^{-m} \ . \tag{4.37}$$

Proof.

$$Pr[y] = \sum_{p \in \mathcal{X}^{ut}} Pr[p] Pr[y|p]$$

Def. of $\hat{\pi}^{ut}_{(p,t)}$ and S^{ut}

$$\begin{aligned} &:= \frac{1}{2^n} \sum_{p \in \mathcal{X}^{ut}} \mathsf{S}^{ut}_{p,y} \\ &= \frac{1}{2^n} \sum_{\substack{p \in \mathcal{X}^{ut}: \\ n_{\mathtt{a}}(p_{1\ldots m})=y}} 1 \end{aligned}$$

We need to count how many secrets $p \in \mathcal{X}^{ut}$ satisfy the restriction $n_{\mathtt{a}}(p_{1\ldots m}) = y$. In the first m elements of x we have y $\mathtt{a}$'s, so $\binom{m}{y}$ different combinations. The other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, thus $\binom{n-m}{y'}$ different combinations.

$$\begin{aligned} &= \frac{\binom{m}{y}}{2^n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} \\ &= \frac{\binom{m}{y} 2^{n-m}}{2^n} = \binom{m}{y} 2^{-m} \ . \end{aligned}$$

□

Lemma 4.2.8 (Utility loss for a specific output y)**.** *Given the set of secrets $\mathcal{X}^{ut}$, the prior $\hat{\pi}^{ut}$, the loss function ℓ and channel S^{ut}, and given that the adversary observed the output y, the posterior vulnerability given this observation is*

$$U_\ell(\hat{\delta}^{y,ut}) = \frac{1}{n 2^{n-m}} \min_{0 \le k \le n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} |k - y - y'| \ ,$$

where $\hat{\delta}^{y,ut} \in \mathbb{D}\mathcal{X}^{ut}$ is the inner distribution when $\hat{\pi}^{ut}$ is the prior distribution and y is observed (i.e., $Pr[X|Y = y]$).

Proof.

$$U_\ell(\hat{\delta}^{y,ut}) = \min_{w \in \mathcal{W}} \sum_{p \in \mathcal{X}^{ut}} Pr[(p,t)|y] \cdot \ell(w,p)$$

Bayes' theorem:

$$= \min_{w \in \mathcal{W}} \sum_{p \in \mathcal{X}^{ut}} \frac{Pr[p] Pr[y|p]}{Pr[y]} \cdot \ell(w,p)$$

Definition of $\hat{\pi}^{ut}$, S^{ut}, ℓ and by Lemma 4.2.7:

$$= \min_{w \in \mathcal{W}} \sum_{p \in \mathcal{X}^{ut}} \frac{2^{-n} \cdot \mathsf{S}^{ut}_{p,y}}{\binom{m}{y} 2^{-m}} \cdot \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right|$$

$$= \frac{2^m}{\binom{m}{y} 2^n} \min_{w \in \mathcal{W}} \sum_{\substack{p \in \mathcal{X}^{ut}: \\ n_{\mathtt{a}}(p_{1\ldots m}) = y}} \left| w - \frac{n_{\mathtt{a}}(p)}{n} \right|$$

We need to count how many secrets $p \in \mathcal{X}^{ut}$ satisfy the restriction $n_{\mathtt{a}}(p_{1\ldots m}) = y$. In the first m elements of x we have y $\mathtt{a}$'s, so $\binom{m}{y}$ different combinations. The other $n-m$ people can have any value ($\mathtt{a}$ or $\mathtt{b}$), so we say that there are y' $\mathtt{a}$'s in $p_{m+1\ldots n}$, such that y' goes from 0 to $n-m$, thus $\binom{n-m}{y'}$ different combinations. Finally, $n_{\mathtt{a}}(p_{1\ldots m}) = y$ and $n_{\mathtt{a}}(p_{m+1\ldots n}) = y'$ implies $n_{\mathtt{a}}(p) = y + y'$.

$$= \frac{2^m}{\binom{m}{y} 2^n} \min_{w \in \mathcal{W}} \sum_{y'=0}^{n-m} \binom{m}{y} \binom{n-m}{y'} \left| w - \frac{y + y'}{n} \right|$$

Following the definition of $\mathcal{W}$, we can replace $\min_{w \in \mathcal{W}}$ by $\min_{0 \le k \le n}$ and rewrite each action w as k/n.

$$= \frac{1}{2^{n-m}} \cdot \min_{0 \le k \le n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} \left| \frac{k}{n} - \frac{y + y'}{n} \right|$$

$$= \frac{1}{n 2^{n-m}} \cdot \min_{0 \le k \le n} \sum_{y'=0}^{n-m} \binom{n-m}{y'} |k - y - y'| \ .$$

□