WHEN AND HOW TO PROTECT? MODELING REPEATED INTERACTIONS WITH COMPUTING SERVICES UNDER UNCERTAINTY

by

KAVITA KUMARI, B.Tech

DISSERTATION
Presented to the Graduate Faculty of
The University of Texas at San Antonio
In Partial Fulfillment
Of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

COMMITTEE MEMBERS:
Murtuza Jadliwala, Ph.D., Chair
Sumit Jha, Ph.D.
Dhireesha Kudithipudi, Ph.D.
Rocky Slavin, Ph.D.
Anindya Maiti, Ph.D.

THE UNIVERSITY OF TEXAS AT SAN ANTONIO
College of Sciences
Department of Computer Science
August 2022

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Dr. Murtuza Jadliwala for his patience and immense knowledge, which he used to mold me and my research to its very best version. His support helped me develop the analytical skills much needed for my dissertation.

I would also like to acknowledge Dr. Anindya Maiti for his insightful comments and encouragement, which helped me polish my research work. I would also like to thank the committee members, Dr. Sumit Jha, and Dr. Dhireesha Kudithipudi, Dr. Rocky Slavin, for their patience and feedback, which helped me immensely to refine my dissertation proposal. Lastly, I would like to thank Dr. Ravi Sandhu for being part of my dissertation proposal committee.

I would like to acknowledge the National Science Foundation (NSF) under award number 1828071 (originally 1523960) for funding parts of this dissertation, especially Chapter 2.

Finally, I would like to acknowledge my family for their continuous support and patience, which helped me focus on my research work.

This Masters Thesis/Recital Document or Doctoral Dissertation was produced in accordance with guidelines which permit the inclusion as part of the Masters Thesis/Recital Document or Doctoral Dissertation the text of an original paper, or papers, submitted for publication. The Masters Thesis/Recital Document or Doctoral Dissertation must still conform to all other requirements explained in the Guide for the Preparation of a Masters Thesis/Recital Document or Doctoral Dissertation at The University of Texas at San Antonio. It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this Masters Thesis/Recital Document or Doctoral Dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts, which provide logical bridges between different manuscripts, are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the students contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the Masters Thesis/Recital Document or Doctoral Dissertation attest to the accuracy of this statement.

August 2022

WHEN AND HOW TO PROTECT? MODELING REPEATED INTERACTIONS WITH COMPUTING SERVICES UNDER UNCERTAINTY

Kavita Kumari, Ph.D.
The University of Texas at San Antonio, 2022

Supervising Professor: Murtuza Jadliwala, Ph.D.

Modern computing systems and web applications often interact or provide services to entities without knowing their type, i.e., it could be interacting with an honest entity wanting a genuine good response from it, or it could be interacting with a malicious entity wanting to compromise it by launching security and privacy attacks against it. Hence, designing and modeling effective defense mechanisms for such systems is a non-trivial task. The problem is even more difficult against a strategic adversary who repeatedly interacts with the target system by imitating an honest end-user/entity to infer as much information as possible, then stealthily and strategically attacking it at the appropriate opportunity. In such an interaction scenario, on the one hand, the system's goal is to strategically block the malicious entities without significantly impacting the quality of service provided to the honest entities. On the other hand, a malicious entity's objective is to stealthily compromise the target system as quickly as possible without being detected and in the most cost-efficient manner. This dissertation plans to study this classical trade-off in modern computing services and web applications by focusing on three unique use-cases of such applications/services that present this conundrum. In the first use-case, we consider the scenario of a mobile operating system attempting to regulate access to zero-permission or permission-less sensors such as accelerometers, gyroscopes, and ambient light sensors. These sensors are critical for all mobile applications, but malicious applications can misuse data from them to infer private information about users. So the mobile system must strategically decide under what conditions to share data from these sensors without knowing the type (malicious or honest) of a mobile application requesting the data. In the first research thrust of this dissertation, we address the above trade-off

by modeling the strategic interactions between mobile applications and a defense mechanism (or a mobile system) using a two-player discrete-time, imperfect information game called the Signaling game. The second use-case that we consider comprises of a black-box machine learning model that provides a label to each query sent by an end-user and an explanation (or attribution) for that label. Such explanations/attributions can be very useful for honest users in understanding model decisions. However, malicious users can misuse repeated explanations to reveal private model information such as parameters and training data. So the model must strategically decide under what conditions to stop sharing explanations with end users without knowing their type (malicious or honest). In the second research thrust, we address this trade-off by modeling the dynamics of explanation variance generated by a system (comprising of an ML model and the corresponding explanation technique) for predictions/labels related to queries sent by end-users. Specifically, we model the interactions between an end-user and the system, where the variance of the explanations generated by the system evolve according to a *stochastic differential equation (SDE)*, as a *two-player continuous-time Signaling Game*. Such a modeling and analysis exercise helps us determine the optimal explanation variance threshold for an attacker to launch explanation-based threshold attacks against the system. The third use-case that we consider is a federated learning scenario, where multiple clients (malicious or honest) cooperatively learn a global system model (computed by some server) in a distributed or decentralized fashion. In such a distributed learning scenario, malicious clients want to cheat the server (computing the global model) by stealthily sending false/incorrect updates, while the server wants to detect such malicious updates in a timely fashion so that it does not corrupt the global model. In the third research thrust, we address this trade-off by designing a Bayesian defense mechanism on the server-side. Specifically, we employ concepts from non-parametric Bayesian modeling to compute a probabilistic measure that can be leveraged in the detection phase (of the malicious updates) with the aim to decouple it from the local clients' training strategies such as data distribution, attack strategy of the malicious clients or the number of clients selected in a federated learning training round.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Modeling and designing effective defense mechanisms for modern computing services and applications is often non-trivial because of an adversary that repeatedly interacts with the service provider and strategically hides its actions within the actions of honest users of the service. Moreover, the service/application provider is unaware of the type of user it interacts with. It could be interacting with a malicious user who can perfectly imitate an honest user, whose eventual goal is to attack the provider stealthily, or it could be interacting with an honest user who is requesting a legitimate service from the provider. In other words, the application or service provider operates with imperfect information about the type of user that interacts with it. In such an interaction scenario, on the one hand, the service provider would like to strategically serve the honest users while preventing malicious users from achieving their goal of compromising it. On the other hand, the malicious user or adversary wants to compromise the target system as efficiently as possible (in terms of cost and time). This dissertation plans to study this classical trade-off between a service/application provider and a malicious user interacting with it by considering three specific application use-cases of such interactions as shown in Figure 1.1. Each use case highlights a different nature/perspective of the interaction, which presents a unique research challenge that is addressed in this dissertation.

1.1 Analyzing Defense Strategies Against Mobile Information Leakages: A Game-Theoretic Approach.

In this case, we consider a mobile system that provides mobile applications access to on-board zero-permission sensors. Zero-permission sensors are critical sensors on mobile systems (e.g., accelerometers, gyroscopes, and ambient light sensors) that do not require explicit user- or system-defined permissions for applications to access them. Research literature has already shown how data sampled from these sensors can maliciously employ inference attacks against sensitive information such as user keystrokes, activities, and locations [76, 156, 179, 251, 262, 269, 285, 286,

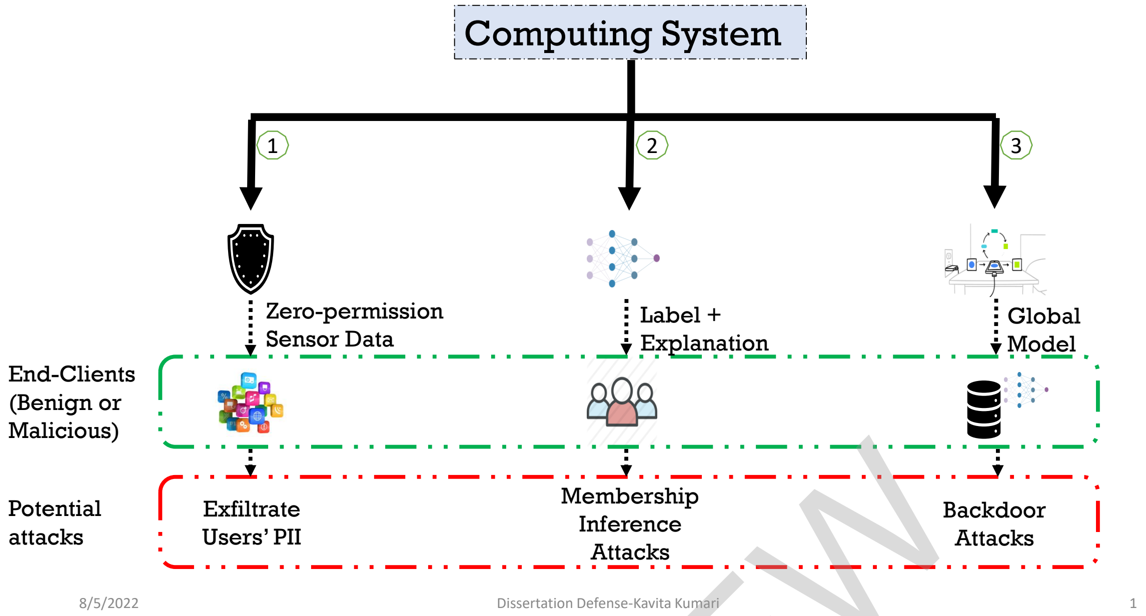
Computing System
1
2
3
Zero-permission Sensor Data
Label + Explanation
Global Model
End-Clients (Benign or Malicious)
Potential attacks
Exfiltrate Users' PII
Membership Inference Attacks
Backdoor Attacks
8/5/2022
Dissertation Defense-Kavita Kumari
1

Figure 1.1: Interactions between the System and the End Clients.

288, 302, 306, 350, 355]. So, the research question here is how a mobile operating system can strategically protect against a malicious application that could attempt to misuse the data sampled from zero-permission sensors, while simultaneously continuing to serve the honest applications optimally. The mobile system has incomplete information about the type of application it interacts with because applications with both honest and malicious intent can request data from zero-permission sensors. We answer the above question using game theory. Specifically, we model the strategic interactions between mobile applications and a defense mechanism using a two-player, imperfect information game, called the Signaling game. Using this game model, we derive equilibria strategies (specifically, Perfect Bayesian Nash Equilibrium) for both the mobile system and the interacting applications. These equilibrium strategies denote the mutual best responses of the interacting entities in this imperfect information environment. We further perform numerical simulations to evaluate how the game evolves under different game parameters in both single-stage and repeated scenarios.

1.2 How much Explanation is Enough? Towards a Game-theoretic Understanding of Explanations of ML Models.

In this case, we consider a classification system, comprising of a black-box machine learning model and an explanation method, that provides a label and its explanation in response to a query sent by an end-user. Providing explanations increases the plausibility of the predictions generated by the model. However, it also provides an attack vector for an adversary who can craft malicious inputs and feed them to the system to compromise it. The research literature has already shown the feasibility of privacy threats in the form of membership inference and other adversarial attacks arising from model explanations. Our focus here is in the former direction, specifically on one fundamental adversarial attack called the Explanation-based Membership Inference Attack (or MIA), where an adversary attempts to determine whether a data-point belongs to the training dataset by leveraging the variance threshold of the gradient-based explanations. However, it is not trivial to compute a variance threshold, especially in a repeated interaction scenario. Despite the intuition that the length and the pattern of repeated interactions with a black-box model and the relevant explanations may significantly impact the leakage of private information from the model, there have been no prior efforts to understand this phenomenon formally. Hence, the research questions that arise in this scenario are: How long should an adversary interact with the target model to compromise the system? Can the target model detect such malicious interactions in a timely fashion to prevent membership inference? and, How can the target model strategically serve both honest and malicious users in such a setting? We answer the above questions by formally modeling the dynamics of explanation variance generated by a system (comprising of an ML model and the corresponding explanation technique) for predictions/labels related to queries sent by end-users. Specifically, we model the interactions between an end-user and the system, where the variance of the explanations generated by the system evolve according to a stochastic differential equation (SDE), as a two-player continuous-time Signaling Game. Next, we characterize the Markov Perfect Equilibrium of this stochastic game and further prove the existence and uniqueness of Markov

Perfect Equilibrium by using concepts from stochastic calculus and the study of ordinary differential equations. We also conduct extensive numerical analysis for four gradient-based explanation methods, namely, Integrated Gradients, Gradient*Input, ε-LRP, and Guided Backpropagation, to determine how the game evolves and the MIA accuracy for an attacker in these scenarios.

1.3 Designing defense mechanisms against backdoors in Federated Learning.

In this case, we consider the Federated learning (FL) use-case. FL enables multiple decentralized clients (malicious or honest) to cooperatively learn a global model/function. The research literature has already shown that FL is prone to adversarial attacks as the malicious client (or clients) can insert backdoors into the global server model during the training process, which can result in poor test performance of the global model on all or some subsets of predictive tasks. Backdoor attacks aim to corrupt the global server model by injecting malicious data points (adversarial triggers) into it during the training process,thus forcing a classifier to misclassify on some data points during the test time. Accordingly, many defense mechanisms have also been introduced in the literature to defend against malicious clients (or clients) who can corrupt the global model by inserting backdoors. However, FL systems are still not robust enough against backdoor attacks as developed defense mechanisms work only under specific conditions. For example, some defense mechanisms do not work when multiple diverse backdoors are simultaneously inserted by a malicious client (or clients). Also, other defense mechanisms clip weights and add noise to negate the effect of malicious model updates, thus reducing the benign performance of the global server model. Therefore, the overarching research question is: How to design a defense mechanism that can effectively defend the trained model against any malicious client (or clients) trying to insert backdoors into the global model? In addition, the proposed defense mechanism should detect backdoors, both when inserted in a single training round or during multiple training rounds. In this work, we take a completely different approach and present BayBFed, a novel framework against backdoor attacks in FL that utilizes Bayesian non-parametric (BNP) modeling techniques.

It functions in two steps. First, it computes an (alternate) probabilistic measure over the clients' weights to keep track of their deviations. Second, a detection algorithm leverages this probabilistic measure to decouple itself from the aforementioned assumptions (as it does not have to administer the client weights *directly*). Specifically, we utilize two BNP extensions: (i) a *Hierarchical Beta-Bernoulli* process to draw a probabilistic measure given the clients' weights, and (ii) an adaptation of the *Chinese Restaurant Process (CRP)*, which we call CRP-Jensen, which is a clustering algorithm that can leverage the above computed probabilistic measure to detect and filter out malicious updates in FL. We extensively evaluate our BNP-inspired defense approach on five popular benchmark datasets: CIFAR10, Reddit, IoT intrusion detection, FMNIST, and MNIST, and show that the designed defense can effectively eliminate malicious updates without deteriorating the benign performance of the model.

CHAPTER 2: ANALYZING DEFENSE STRATEGIES AGAINST MOBILE INFORMATION LEAKAGES: A GAME-THEORETIC APPROACH.

**This chapter has previously appeared in Conference Decision and Game Theory for Security (GameSec 2019) and was published as Lecture Notes in Computer Science(), vol 11836. Springer, Cham. It was co-authored by Murtuza Jadliwala, Anindya Maiti, and Mohammad Hossein Manshaei, and has been reproduced here with minor revisions.*

2.1 Introduction

Modern mobile and wearable devices, equipped with state-of-the-art sensing and communication capabilities, enable a variety of novel context-based applications such as social networking, activity tracking, wellness monitoring and home automation. The presence of a diverse set of on-board sensors, however, also provide an additional attack surface to applications intending to infer personal user information in an unauthorized fashion. In order to thwart such privacy threats, most modern mobile operating systems (including, Android and iOS) have introduced stringent access controls on front-end or user-accessible sensors, such as microphone, camera and GPS. As a result, the focus of adversarial applications has now shifted to employing on-board sensors that are not guarded by strong user or system-defined access control policies. Examples of such back-end or user-inaccessible sensors include accelerometer, gyroscope, power meter and ambient light sensor, and we refer to these as *zero-permission sensors*. As all installed applications have access to them by default, and that they cannot be actively disengaged by users on an application-specific basis, these zero-permission sensors pose a significant privacy threat to mobile device users, as it has been extensively studied in the security literature [76, 139, 156, 179, 251, 260–262, 269, 285, 286, 288, 302, 306, 318, 350, 355, 405, 407].

At the same time, development of efficient and effective protection mechanisms against such privacy threats is still an open problem [77]. One of the main reasons why zero-permission sensors have limited or no access control policies associated with them is because they are required by a majority of applications (accessed by means of a common set of libraries or APIs) primarily for efficient and user-friendly operation on the device's small and constrained form factor and display. For instance, gyroscope data is used by applications to re-position front-ends (or GUIs) depending device orientation, while an ambient light sensor is used to update on-screen brightness. Thus, a straightforward approach of completely blocking access or reducing the frequency at which applications can sample data from these sensors is not feasible, as it will significantly impact their usability. Alternatively, having a static access control policy for each application is also not practical as it will become increasingly complex for users to manage these policies. Moreover, such an approach will not protect against applications that gain legitimate access to these sensors (based on such static policies). Given that all applications (with malicious intentions or not) can request access to these sensors without violating any system security policy, an important challenge for a defense mechanism is to differentiate between authentic sensor access requests and requests that could be potentially misused.

In order to begin addressing this long-standing open problem, we take a clean-slate approach by first formally (albeit, realistically) modeling the strategic interactions between (honest or potentially malicious) mobile applications and an on-board defense mechanism that cannot differentiate between their (sensor access) requests. We employ *game-theory* as a vehicle for modeling and analyzing these interactions. Specifically, we model the following scenario. A defense mechanism on a mobile operating system receives requests to access zero-permission sensors from two different *types* of applications: *honest* and *malicious*. Each of these applications could send either a *normal* or a *suspicious* request for access to on-board zero-permission sensors. A request could be classified as suspicious or normal (non-suspicious) based on the context, frequency or amount of requested sensor data. Although honest applications would typically make normal requests, they could also make suspicious requests depending on application- or context-specific operations and requirements to improve overall application performance and usability. The goal of malicious applications, on the other hand, is to successfully infer private user data from these requests. Normal requests would give them some (probably, not enough) data to carry out these privacy threats, however, suspicious requests could give them additional critical data either to amplify or increase the success probability of their attacks. The defense mechanism, on receiving the request, has one of the following two potential responses: (i) *accept* the request and release the requested sensor data, or (ii) *block* the request preventing any data being released to the requesting application. It should be noted that the defense mechanism does not know the type of the application (i.e., honest or malicious) sending a particular request (i.e., suspicious or non-suspicious), as all mobile applications can currently request zero-permission sensor data without raising a flag or violating any policy. In other words, the defense mechanism has *imperfect information* on the type of application sending the request. The requesting application, on the other hand, has perfect information about its type and potential strategies of the defense mechanism. Given this scenario, the following are the main technical contributions of this paper:

1. We first formally model the strategic interactions between mobile applications and a defense mechanism (outlined above) using a *two-player*, *imperfect-information* game, called the *signaling game* [95]. We refer to it as the *Sensor Access Signaling Game*.

2. Next, we solve the Sensor Access Signaling Game by deriving both the pure- and mixed-strategy *Perfect Bayesian Nash Equilibria (PBNE)* strategy profiles possible in the game.

3. Finally, by means of numerical simulations, we examine how the obtained game solutions or equilibria evolve with respect to different system (or game) parameters in both the *single-stage* and *repeated* (more practical) scenarios.

Our game-theoretic model, and the related preliminary results, is the first clean-slate attempt to formally model the problem of protecting zero-permission sensors on mobile platforms against privacy threats from strategic applications and adversaries (with unrestricted access to it). Our hope is that this model will act as a good starting point for designing efficient, effective and incentive-compatible strategies for protecting against such threats.

2.2 System Model

System Model. Our system (Figure 2.1) comprises of two key entities residing on a user's (mobile) device. The first is *applications* (*APP*) that utilize, and thus, need access to, data from zero-permission sensors. We consider two *types* of applications: *Honest (HA)* and *Malicious (MA)*. Honest applications provide some useful service to the end-user with the help of zero-permission sensor data, while malicious applications would like to infer personal/private information about the user in the guise of offering some useful service. Both honest and malicious applications can request sensor data in a manner which may look normal/non-suspicious or suspicious (details next), regardless of their intentions or use-cases. The second entity is a sensor access regulator, which we refer to as the *Defense Mechanism (DM)*. All sensor access requests (by all applications) must pass through and processed by the *DM*. The *ideal* functionality that the *DM* would like to achieve is to block sensor requests coming from *MA*s, while allowing requests from *HA*s. As noted earlier, the *DM* itself does not know the type (i.e., honest or malicious) of application requesting sensor access - otherwise the job of the *DM* is trivial. This is also a practical assumption as currently all applications can access these sensors without violating any system/user-defined policy (to clarify, there is currently no way to set access control policies for zero-permission sensors on most mobile platforms). As the *DM* has no way of certainly knowing an application's true intentions (and thus, its type), it must rely on the received request (suspicious or non-suspicious, as described next) and its belief about the requesting application's type to determine whether it poses a threat to user privacy or not.

Suspicious and Non-Suspicious Requests. Zero-permission sensor access requests by the applications (to the *DM*) can be classified as either *suspicious* ($\mathcal{S}$) or *non-suspicious* ($\mathcal{NS}$). Such a classification (generally, system-defined) can be accomplished using contextual information available to both the applications and the defense mechanism, such as, frequency, time, sampling rate, and relevance (according to the advertised type of service offered by the application) of these requests. Although there are several efforts in the literature in the direction of determining sensor over-privileges in mobile platforms [138, 178], we abstract away this detail to keep our model gen-

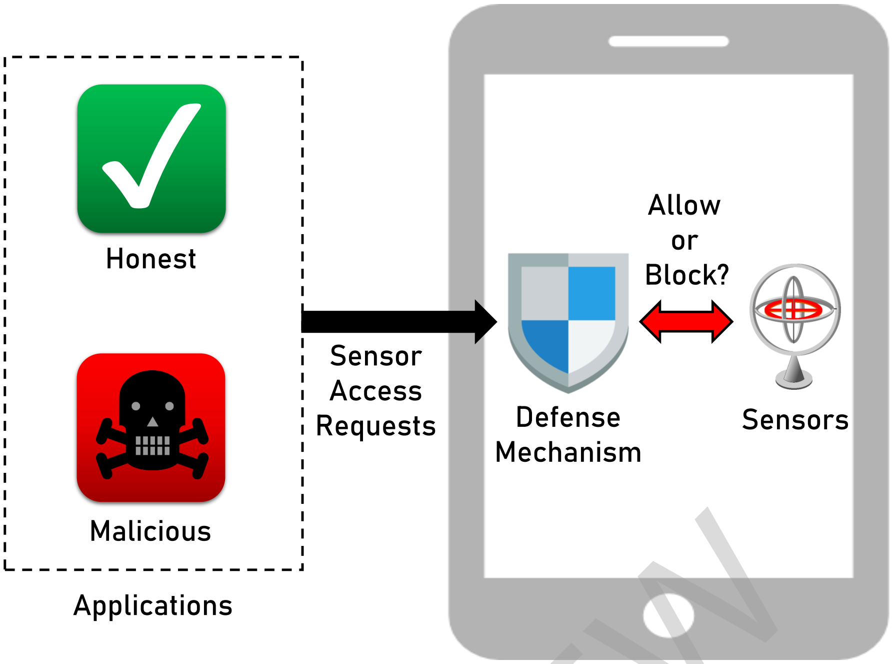

Figure 2.1: System Model

eral. We, however, assume that malicious applications are able to masquerade themselves perfectly as honest applications (in terms of the issued sensor requests), which is easy to accomplish when the target of these applications is zero-permission sensors.

Other System Parameters. The strategic interactions between the (honest or malicious) *APP* and *DM* can be characterized using several system parameters which we summarize in Table 2.1. In addition to identifying these parameters, we also establish the relationship between these parameters by considering realistic network and system constraints as discussed next. For example, if the cost of an application processing a successful $\mathcal{S}$ request (i.e., $c^{\mathcal{S}}$) or $\mathcal{NS}$ request (i.e., $c^{\mathcal{NS}}$) is expressed in terms of the CPU utilization (of the application), then it is clear that $c^{\mathcal{S}} \geq c^{\mathcal{NS}}$ because suspicious requests would usually solicit fine-grained (high sampling rate) sensor data compared to non-suspicious requests, thus requiring more processing time. By a similar rationale, $\psi^{\mathcal{S}} \geq \psi^{\mathcal{NS}}$, where $\psi^{\mathcal{S}}$ and $\psi^{\mathcal{NS}}$ are the costs to a *DM* (or the system) for processing a $\mathcal{S}$ or $\mathcal{NS}$ request, respectively. Now, the cost to the *HA* in terms of loss in usability when its request is blocked by *DM* (i.e., γ) and benefit for the *HA* in terms of gain in usability when its request is allowed by the *DM* (i.e., σ) are inversely proportional ($\gamma \propto 1/\sigma$). Similarly, benefit to the *MA* when it's request is al-

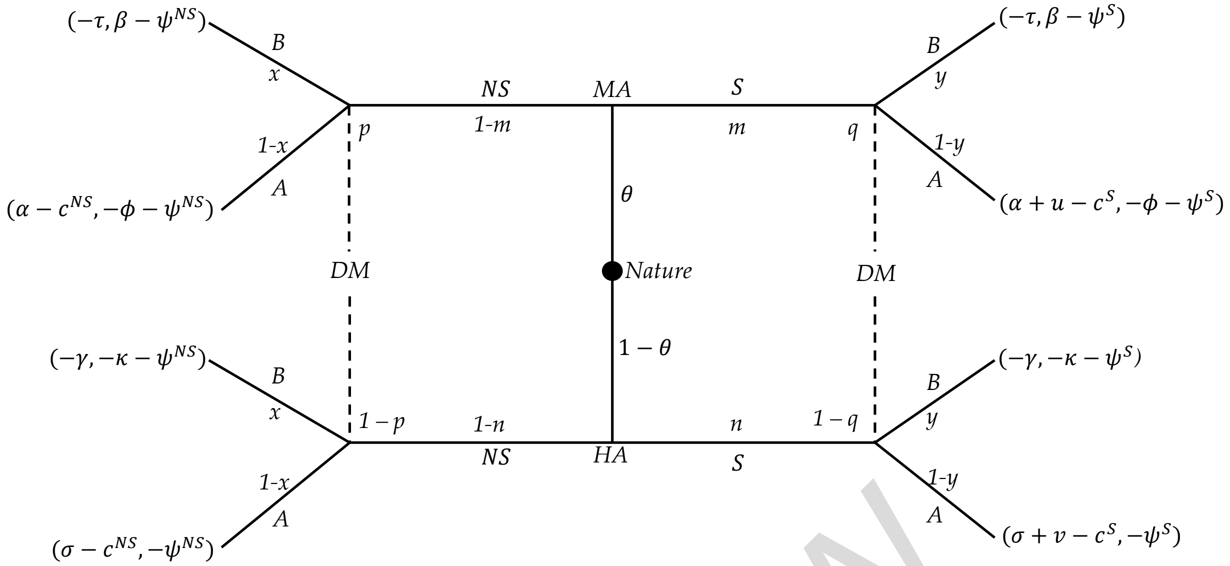
$(-\tau, \beta - \psi^{NS})$
B
x
1-x
A
$(\alpha - c^{NS}, -\phi - \psi^{NS})$
NS
1-m
MA
S
m
p
q
θ
DM
Nature
DM
$1-\theta$
$(-\tau, \beta - \psi^{S})$
B
y
1-y
A
$(\alpha + u - c^{S}, -\phi - \psi^{S})$
$(-\gamma, -\kappa - \psi^{NS})$
B
x
$1-p$
1-n
NS
HA
n
S
$1-q$
1-x
A
$(\sigma - c^{NS}, -\psi^{NS})$
$(-\gamma, -\kappa - \psi^{S})$
B
y
1-y
A
$(\sigma + v - c^{S}, -\psi^{S})$

Figure 2.2: Extensive form of the Sensor Access Signaling Game $\mathbb{G}_D$.

lowed by *DM* (α) can be expressed in terms of monetary gains. An acute example would be if *MA* is able to successfully infer user's banking credentials using sensor data [251, 261, 350, 405], and uses it for theft. A more clement example of monetary gain could be through selling contextual data (inferred from sensor data) to advertising companies, without user's consent. Accordingly, *MA* is set back with a proportional cost (τ) if its request is rejected by *DM*, i.e., $\alpha \propto \tau$. On the other hand, *DM*'s cost of allowing a *MA*'s request (ϕ) versus benefit to the *DM* for blocking *MA*'s request (β) are also inversely proportional ($\phi \propto 1/\beta$). *DM*'s cost of allowing a *MA*'s request is essentially borne by the user, but since the *DM* is working in the best interest of the user, we combine their costs and benefits. Consequently, in case *DM* blocks an *HA*'s request, it incurs a cost (κ) representing loss of utility/usability for the user. Lastly, we also capture the *difference in benefits* for *MA* and *HA*, in case they send out a $\mathcal{S}$ versus $\mathcal{NS}$ request, as *u* and *v*, respectively. In essence, *u* denotes the gain in benefit due to *MA*'s better inference accuracy caused by sensor data obtained from $\mathcal{S}$, and *v* denotes the improvement of *HA*'s utility/usability due to sensor data obtained from $\mathcal{S}$. We also assume that these different (discrete) costs and benefits are appropriately scaled and normalized such that their absolute values lie in the same range of real values. Next, we outline the signaling game formulation to capture the strategic interaction between the mobile applications

Table 2.1: System entities and parameters.

Symbol	Definition
DM	Defense Mechanism
HA	Honest Application
MA	Malicious Application
θ	Probability that Nature selects *MA*
$\mathcal{S}$	Suspicious sensor request
$\mathcal{NS}$	Non-suspicious sensor request
q	Belief probability of the *DM* that the requester is of type *MA* on receiving a $\mathcal{S}$ request
p	Belief probability of the *DM* that the requester is of type *MA* on receiving a $\mathcal{NS}$ request
B	*DM* response to block a sender request
A	*DM* response to allow a sender request
$c^{\mathcal{S}}$	Cost of an application processing a successful $\mathcal{S}$ request
$c^{\mathcal{NS}}$	Cost of an application processing a successful $\mathcal{NS}$ request
γ	Cost to the *HA* when its request is blocked by *DM*
$\psi^{\mathcal{S}}$	Cost of a *DM* processing a $\mathcal{S}$ request
$\psi^{\mathcal{NS}}$	Cost of a *DM* processing a $\mathcal{NS}$ request
ϕ	Cost to the *DM* when *MA*'s request is allowed
τ	Cost to the *MA* when its request is blocked by the *DM*
κ	Cost to the *DM* when *HA*'s request is blocked
α	Benefit to the *MA* when its request is allowed by the *DM*
β	Benefit to the *DM* for blocking *MA*'s request
σ	Benefit to the *HA* when its request is allowed by the *DM*
u	Benefit difference to *MA* for sending $\mathcal{S}$ instead of $\mathcal{NS}$
v	Benefit difference to *HA* for sending $\mathcal{S}$ instead of $\mathcal{NS}$
m	probability with which *MA* plays the $\mathcal{S}$ strategy
n	probability with which *HA* plays the $\mathcal{S}$ strategy
x	probability with which *DM* plays the $\mathcal{B}$ strategy on receiving a $\mathcal{NS}$ request
y	probability with which *DM* plays the $\mathcal{B}$ strategy on receiving a $\mathcal{S}$ request

(requesting zero-permission sensor access) and the defense mechanism (attempting to regulating these requests).

Game Model. A classical signaling game [95] is a sequential two-player incomplete information game in which *Nature* starts the game by choosing the *type* of the first player or *player 1*. Player 1 is the more informed out of the two players since it knows the choice of *Nature* and can send *signals* to the less informed player, i.e., *player 2*. Player 2 is uncertain about the type of player 1, and must decide its strategic response solely based on the signal received from player 1. In other words, player 2 must decide its best response to player 1's signal without any knowledge about the type of player 1. Both players receive some utility (payoff) depending on the signal, type of player 1 and the response by player 2 (to player 1's signal). Both the players are assumed to be rational and are interested in solely maximizing their individual payoffs.

Given the above generic description of the signaling game, let us briefly describe how our zero-permission sensor access scenario naturally lends itself as a single-stage signaling game. We refer to this game as the *Sensor Access Signaling Game* and is formally represented as $\mathbb{G}_D = \langle \mathbb{P}, \mathbb{T}, \mathbb{S}, \mathbb{A}, \mathbb{U}, \theta, (p,q) \rangle$, where $\mathbb{P}$ is the set of players, $\mathbb{T}$ is the set of player 1 types, $\mathbb{S}$ is the set of player 1 signals, $\mathbb{A}$ is the set of player 2 actions, $\mathbb{U}$ is the *payoff/utility function*, θ is the *Nature's probability distribution function*, and (p,q) are player 2's *belief functions* about player 1's type. Each sensor access request by an application can be modeled as a single stage of the above signaling game. In each such stage, $\mathbb{P}$ contains two players, i.e., *APP* which is player 1 and the *DM* which is player 2. As there are two types of applications (or player 1), i.e., honest (*HA*) and malicious (*MA*), $\mathbb{T} \equiv \{HA, MA\}$. As applications can send two types of signals (or requests), i.e., suspicious ($\mathcal{S}$) and non-suspicious ($\mathcal{NS}$), $\mathbb{S} \equiv \{\mathcal{S}, \mathcal{NS}\}$. As the *DM* (or player 2) takes two types of actions depending on the received signal from player 1, i.e., Allow (*A*) or Block (*B*), $\mathbb{A} \equiv \{A, B\}$. The utility function $\mathbb{U} : \mathbb{T} \times \mathbb{S} \times \mathbb{A} \rightarrow (\mathbb{R}, \mathbb{R})$ assigns a real-valued payoff to each player (at the end of the stage) based on the benefit received and the cost borne by each player, and is outlined in the extensive form of the game depicted in Figure 2.2. The first utility in the pair is the *APP*'s utility denoted as U_{APP}, while the second utility in the pair is the *DM*'s utility denoted as U_{DM}. Lastly, let $\Gamma_{APP} = \{\mu_{APP} | \forall t_i \in \mathbb{T}, \sum_{\lambda \in \mathbb{S}} \mu_{APP}(\lambda | t_i) = 1; \forall t_i \in \mathbb{T}\}$ and $\Gamma_{DM} = \{\mu_{DM} | \forall \lambda \in \mathbb{S}, \sum_{a \in \mathbb{A}} \mu_{DM}(a|\lambda) = 1; \forall \lambda \in \mathbb{S}\}$ be the strategy spaces for *APP* and *DM*, respectively. A strategy μ_{APP} for the *APP* and μ_{DM} for the *DM* can be either *pure* or *mixed*, as identified by parameters *m*, *n*, *y* and *x* in Figure 2.2. For pure strategies $m, n, y, x \in \{0, 1\}$, while for mixed strategies $0 < m, n, y, x < 1$. Moreover, let us represent each of the *DM*'s belief functions by conditional (posterior) probability distributions as $q = Pr(MA|\mathcal{S})$ and $p = Pr(MA|\mathcal{NS})$, which also imply that $1 - q = Pr(HA|\mathcal{S})$ and $1 - p = Pr(HA|\mathcal{NS})$.

Now, let's characterize the set of equilibrium strategies in $\mathbb{G}_D$, i.e., a set of strategy pairs that are mutual best responses to each other and no player has any incentive to move away from their strategy in that pair. In order to determine mutual best responses, we need to evaluate the actions (or strategies) of each player at each *information set* of the game. *APP*'s information set comprises of a single decision point (i.e., to select a signal $\lambda \in \{\mathcal{S}, \mathcal{NS}\}$) after Nature makes its selection of