On Adversarial Risk Analysis in Official Statistics

Fabrizio Ruggeri

Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche
Via Alfonso Corti 12, I-20133, Milano, Italy, European Union

fabrizio@mi.imati.cnr.it
www.mi.imati.cnr.it/fabrizio

Joint work with Victor Gallego, Jorge Gonzalez-Ortega, Roi Naveiro, Alberto Redondo, David Rios Insua and Refik Soyer

OUTLINE OF THE TALK

- Adversarial Risk Analysis (ARA)
- Adversarial Hypothesis Testing
- Bayesian approaches to protecting classifiers from attacks
  - ACRA: generative classifier
  - AB-ACRA: discriminative classifier
- Some ideas of possible interest in Official Statistics

ADVERSARIAL RISK ANALYSIS

- Adversarial risk analysis (ARA) is a relatively new area of research that informs decision-making when facing intelligent opponents and uncertain outcomes
- It is a decision-theoretic alternative to classical game theory that uses Bayesian subjective distributions to model the goals, resources, beliefs, and reasoning of the opponent
- It enables an analyst to express his/her Bayesian beliefs about an opponent's utilities, capabilities, probabilities and the type of strategic calculation that the opponent is using
- Within that framework, the analyst then solves the problem from the perspective of the opponent while placing subjective probability distributions on all unknown quantities
- This produces a distribution over the actions of the opponent that permits the analyst to maximise his/her expected utility, accounting for the uncertainty he/she has about the opponent

ARA IN A NUTSHELL [1]

One of the advantages of ARA is its ability to partition the uncertainty into three separate components:

- **Aleatory uncertainty**: uncertainty in the outcome, conditional on the choices of each the opponents, to be handled by conventional statistical risk analysis
- **Epistemic uncertainty**: uncertainty in the opponent's utility function and assessment of the probability of outcomes conditional on the decisions that are made (by the analyst and the opponent), to be handled by in a Bayesian framework, making subjective probability assessments about each of these quantities
- **Concept uncertainty**: uncertainty about how the opponent is making his decision, since he might be a game theorist and seeks an equilibrium solution, or, perhaps, he randomizes, or follows some other protocol

[1] Based on Banks, Gallego, Naveiro, Rios Insua, 2020

ARA IN A NUTSHELL [2]

To make this a little more concrete, consider a sealed bid auction between Daphne and Apollo, each of whom wants to own a first edition of the *Theory of Games and Economic Behaviour*. Daphne's **aleatory uncertainty** is the value she receives conditional on her bid and Apollo's. If she has not been allowed to examine the book prior to the auction, then its condition is a random variable–perhaps it is old and torn, or perhaps it has marginalia written by John Nash, and both circumstances affect its value. **Epistemic uncertainty** arises because Daphne does not know the value of the book to Apollo, nor what he thinks is the probability that he will win with a bid of x dollars, nor how much money Apollo has. The **concept uncertainty** reflects the fact that Daphne does not know whether Apollo is determining his bid using classical game theory, or whether he is simply bidding some unknown fraction of his true top-dollar value, or using some other principle.

[2] Cited from Banks, Gallego, Naveiro, Rios Insua, 2020

CLASSIFICATION UNDER ATTACKS

- Sentiment analysis problem: assess if a movie had positive or negative reviews
- 2400 IMDb reviews (1200 positive, 1200 negative) extracted from Kotzias et al (2015)
- 150 binary features indicating the presence or absence of the most common words in the dataset
- Label indicating positive ($y = 0$) or negative ($y = 1$) review
- Attacker A: manipulate positive reviews $\Rightarrow$ classified as negative ones
- Attacker B: manipulate negative reviews $\Rightarrow$ classified as positive ones
- Attacks based on change of two words at most
- Accuracy of 4 classifiers over clean and attacked test data

CLASSIFICATION UNDER ATTACKS

Classifier	**Clean data**	**Attacked data** Attacker A	**Attacked data** Attacker B
Logistic Regression	0.728 ± 0.005	0.322 ± 0.011	0.418 ± 0.010
Naive Bayes	0.722 ± 0.004	0.333 ± 0.009	0.405 ± 0.009
Neural Network	0.691 ± 0.019	0.338 ± 0.021	0.417 ± 0.015
Random Forest	0.720 ± 0.005	0.327 ± 0.011	0.397 ± 0.013

- Accuracy comparison (with precision) of four classifiers on clean and attacked data
- All models trained under same conditions, randomly splitting the dataset into train and test subsets with sizes 90% and 10%, respectively
- Accuracy means and standard deviations are estimated via hold-out validation over 10 repetitions
- MAP (Maximum a posterior) under Gaussian prior for Logistic Regression
- Two hidden layers for Neural Network

WHY ADVERSARIAL CLASSIFICATION?

- In multiple domains such as malware detection, automated driving systems, or fraud detection, classification algorithms are susceptible to being attacked by malicious agents willing to perturb the value of instance covariates to pursue certain goals
- Such problems pertain to the field of adversarial machine learning and have been mainly dealt with, perhaps implicitly, through game-theoretic ideas with strong underlying common knowledge assumptions
- These are not realistic in numerous application domains in relation to security and business competition
- We present an alternative Bayesian decision theoretic framework that accounts for the uncertainty about the attacker's behavior using adversarial risk analysis concepts

BAYESIAN DECISION ANALYSIS WITH ONE AGENT

- (Bayesian) Decision Analysis supports a Decision Maker (DM) in making decisions under uncertainty:
  - Set of alternatives (actions) $a \in \mathcal{A}$
  - Unknown parameter θ depending on *state of nature*
  - Consequence $c(a, \theta)$ of action a when θ occurs
  - Utility function $u(c(a, \theta))$
  - Posterior distribution $\pi(\theta|x)$ on parameter θ, after observing x
  - Optimal action satisfies the Maximum (Subjective) Expected Utility Principle:

$$a^* = \arg\max_{a \in \mathcal{A}} \int u(c(a, \theta))\pi(\theta|x)d\theta$$

- Just one agent playing against *Nature*

BAYESIAN DECISION ANALYSIS WITH TWO AGENTS

- Unlike the "usual" case of an agent against *Nature*, here we suppose there is an opponent, like in a two-person simultaneous game
- Suppose an Attacker (A) selects an action from the set $\mathcal{A} = \{a_1, \ldots, a_n\}$, e.g. bombing a train or an airport, and a Defender (D) chooses an action from the set $\mathcal{D} = \{d_1, \ldots, d_m\}$, e.g. more police at train station or airport
- For each pair of actions (a, d) there is a consequence $s \in \mathcal{S}$, e.g. casualties or not
- We see the problem from the viewpoint of D
- $\pi_D(a)$: D's belief about A's probability of choosing action $a \in \mathcal{A}$
- $p_D(s|a, d)$: D's subjective probability for each possible outcome $s \in \mathcal{S}$ given every choice $(a, d) \in \mathcal{A} \times \mathcal{D}$
- $u_D(d, a, s)$: D's utility for each combination of outcome and pair of choices
- D's expected utility maximised by choosing $d^* \in \mathcal{D}$ s.t.

$$d^* = \arg\max_{d \in \mathcal{D}} \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} u_D(d, a, s) p_D(s|a, d) \pi_D(a) da ds$$

BAYESIAN DECISION ANALYSIS WITH TWO AGENTS

- $d^* = \arg\max_{d \in \mathcal{D}} \int_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} u_D(d, a, s) p_D(s|a, d) \pi_D(a) da ds$
- How do we choose $u_D(d, a, s)$, $p_D(s|a, d)$ and $\pi_D(a)$?
- Utility based on D's preferences and priors from expert's opinions, possibly combined with past data
- Knowledge of A's behaviour, like in Game Theory $\Rightarrow$ unrealistic
- In ARA $\pi_D(a)$ requires
  - Strategic thinking about A: Utility maximiser? Level-k thinker?
  - Guesses about A's probabilities and utilities
- Our goal: help D to assess A's optimal decisions to mitigate their effects

FEW CONCEPTS IN CYBERSECURITY

- Attackers' goals classified along two dimensions:
  - Violation type
    - integrity violation (aimed at moving the prediction about particular instances toward the attacker's target, for example, to have malicious samples misclassified as legitimate)
    - availability violation (aimed at increasing the predictive error to make the system unusable)
    - privacy violation (exploratory attacks to gather information about the ML system)
  - Attack specificity
    - targeted attacks, addressing a few specific defenders
    - indiscriminate attacks, affecting many defenders in a random manner

FEW CONCEPTS IN CYBERSECURITY

- Adversary's knowledge about ML system
  - White box or perfect knowledge attacks: the adversary knows every aspect of the system but this is almost never the case in security scenarios, except perhaps for insider attacks or in sequential settings where the ML system moves first, training an algorithm to fit its parameters, and the adversary, who moves afterwards, has time to observe the behaviour of the system and learn about it
  - Black box or zero knowledge attacks: the adversary has capabilities to query the system but does not have any information about the data, feature space, or specific algorithms used and this is the most reasonable assumption when attacking and defending decisions are made simultaneously
  - Grey box or limited knowledge: the most common type of attacks in security settings, especially when attacking and defending decisions are made sequentially but there is private information that agents are not willing to share

FEW CONCEPTS IN CYBERSECURITY

- Adversary's capability to influence on data and other features
  - Poisoning attacks obfuscating the training data to later induce errors during operations
  - Evasion attacks not influencing on training data, but performing modifications during operations, e.g. when trying not to be detected by an antivirus
  - Structural attacks changing the underlying structure of the problem affecting process parameters
  - Parallel attacks performing multiple attack paths toward a single objective or the simultaneous execution of attacks across interconnected systems

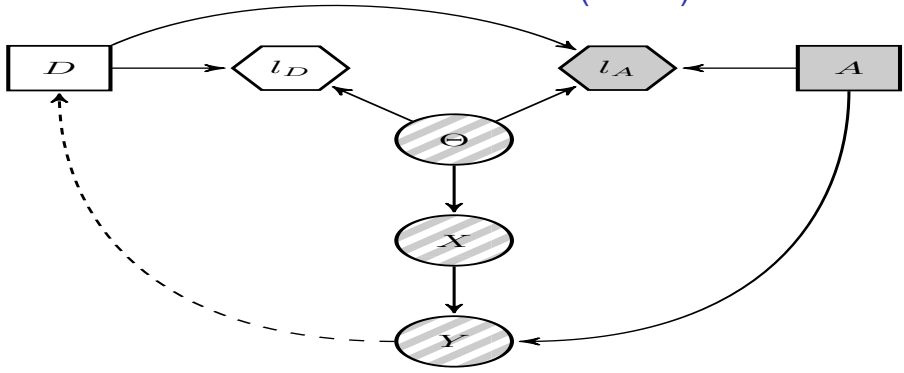
FEW CONCEPTS IN CYBERSECURITY

- Protecting (ML) learning systems against the modelled attacks
  - Reactive defences aimed to mitigate, even eliminate, the effects of an eventual attack, e.g., timely detection of attacks, frequent retraining of learning algorithms, or verification of algorithmic decisions by experts
  - Proactive defences aimed to prevent attack execution, e.g., producing provably secure algorithms against perturbations
- Protecting learning systems at different stages
  - Training time: defences aimed train learning systems robustly, anticipating future adversarial attacks
  - Operation time: defences, when receiving a potentially attacked instance, aimed to undertake inference about possible originating instances to make the corresponding decision

ADVERSARIAL HYPOTHESIS TESTING

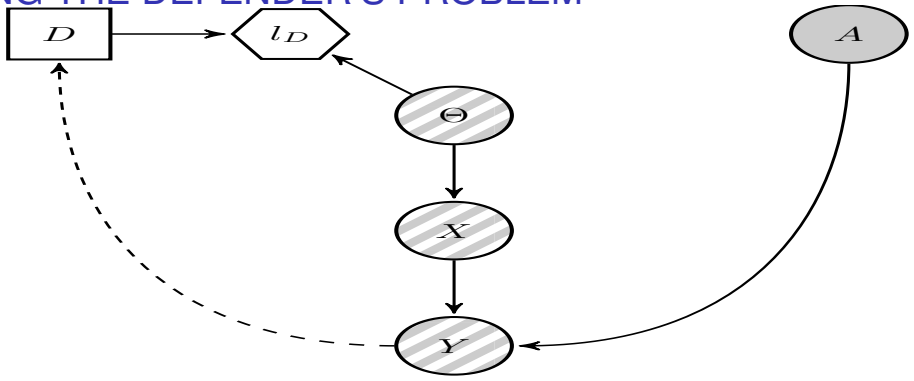
- Test of two simple hypotheses: $\Theta = \{\theta_0, \theta_1\}$
- Observation x generated according to a model depending on θ
- x altered to y by A's action a
- y observed by D $\Rightarrow$ D's decision d on θ based on y, without observing x
- Depending on d and actual $\theta \Rightarrow$ losses (utilities) for both agents
- Efforts by A in minimising the loss
- Support for D in choosing θ to minimise the loss
- Example: $X \sim \mathcal{E}(\theta)$ failure time s.t. $\theta_0 \ll \theta_1 \Rightarrow$ better reliability for θ_0

AHT: BI-AGENT INFLUENCE DIAGRAM (BAID)



- Decisions: D (depending on Y) and A
- Random: $\Theta \to X \to Y$ (Y influenced also by the decision A)
- Losses: l_D and l_A depending on Θ and related decisions (l_A also on decision D)
- Node: decision (square), uncertainty (circle), deterministic (double), utility (hex.)
- Arrow: conditional relation (solid), information available at decision time (dashed)

SOLVING THE DEFENDER'S PROBLEM



- Diagram from D's viewpoint ($\Rightarrow$ A random)
- Belief $\pi_D(\theta)$ on hypotheses: $p_D(\theta = \theta_i) = \pi_i^D, \quad i = 0, 1$
- Belief $\pi_D(x|\theta)$ on how data depend on hypotheses: $X|\theta_i \sim \pi_D(x|\theta_i), \quad i = 0, 1$
- Belief $\pi_D(y|x, a)$ on how action $a \in \mathcal{A}$ by A modifies actual x into observed y
- Belief $\pi_D(a)$ on the attack a performed by A

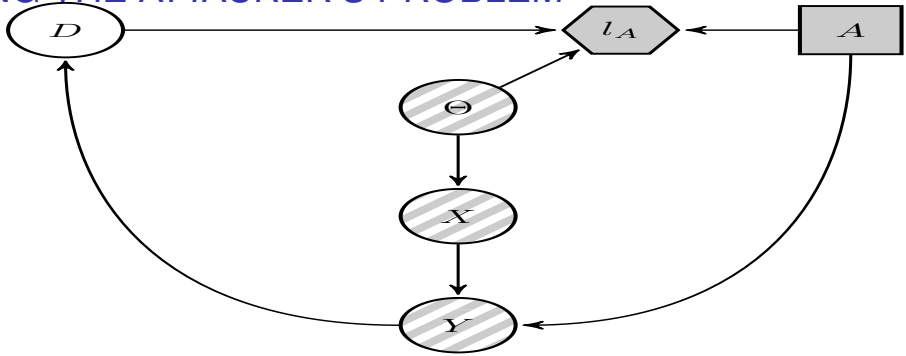
SOLVING THE DEFENDER'S PROBLEM

- Decision space $\mathcal{D} = \{d_0, d_1\}$ s.t. $d_j = \{\text{Defender supports } \theta_j\}$, $j = 0, 1$
- Loss function: $l_D(d_0, \theta_0) = l_D(d_1, \theta_1) = 0$, $l_D(d_0, \theta_1) = 1 \geq c_D = l_D(d_1, \theta_0)$
- Solve: $\arg\min_{d \in \mathcal{D}} \sum_{i=0}^{1} l_D(d, \theta_i)\, \pi_D(\theta_i | y)$
- $\Rightarrow d_0$, i.e. support for θ_0, optimal decision for D if and only if

$$\pi_1^D \iint \pi_D(y|x,a)\, \pi_D(x|\theta_1)\, \pi_D(a)\, dx\, da \leq c_D\, \pi_0^D \iint \pi_D(y|x,a)\, \pi_D(x|\theta_0)\, \pi_D(a)\, dx\, da$$

- Here we use Bayes Theorem $\pi_D(\theta_i|y) = \dfrac{\pi_D(y|\theta_i)\pi_i^D}{\pi_D(y)}$ and neglect the denominator in the minimisation

SOLVING THE ATTACKER'S PROBLEM



- Diagram from A's viewpoint ($\Rightarrow$ D random)
- Belief $\pi_A(\theta)$ on hypotheses: $p_A(\theta = \theta_i) = \pi_i^A, \quad i = 0, 1$
- Belief $\pi_A(x|\theta)$ on how data depend on hypotheses: $X|\theta_i \sim \pi_A(x|\theta_i), \quad i = 0, 1$
- Belief $\pi_A(y|x, a)$ on how action $a \in \mathcal{A}$ modifies actual x into observed y
- Belief $\pi_A(d|y)$ on the decision d taken by D upon observing y
- Loss function $l_A(d, \theta, a)$ for (d_0, d_1) and (θ_0, θ_1)

SOLVING THE ATTACKER'S PROBLEM

- Optimal decision for A given by a^* s.t.

  $$a^* = \arg\min_{a \in \mathcal{A}} \sum_{j=0}^{1} \sum_{i=0}^{1} \int\int l_A(d_j, \theta_i, a)\, \pi_A(d_j|y)\, \pi_A(\theta_i)\, \pi_A(y|x, a)\, \pi_A(x|\theta_i)\, dy\, dx$$

- D does not know $\pi_A(\theta)$, $\pi_A(x|\theta)$, $\pi_A(y|x, a)$, $\pi_A(d|y)$ and $l_A(d, \theta, a)$
- $\Rightarrow$ model uncertainty around them through random probabilities and losses
  $F = (\Pi_A(\theta), \Pi_A(x|\theta), \Pi_A(y|x, a), \Pi_A(d|y), L_A(d, \theta, a))$
- $\Rightarrow$ find optimal random attack

  $$A^* = \arg\min_{a \in \mathcal{A}} \sum_{j=0}^{1} \sum_{i=0}^{1} \int\int L_A(d_j, \theta_i, a)\, \Pi_A(d_j|y)\, \Pi_A(\theta_i)\, \Pi_A(y|x, a)\, \Pi_A(x|\theta_i)\, dy\, dx$$

- Required $\pi_D(a)$ approximated through simulation, sampling from F
- Samples $\left(\Pi_A^k(\theta_i), \Pi_A^k(x|\theta_i), \Pi_A^k(y|x, a), \Pi_A^k(d_j|y), L_A^k(d_j, \theta_i, a)\right)$, $k = 1, ..., K$
- $\Rightarrow a_k^* = \arg\min_{a \in \mathcal{A}} \sum_{j=0}^{1} \sum_{i=0}^{1} \int\int L_A^k(d_j, \theta_i, a)\, \Pi_A^k(d_j|y)\, \Pi_A^k(\theta_i)\, \Pi_A^k(y|x, a)\, \Pi_A^k(x|\theta_i)\, dy\, dx$
- $\Rightarrow \hat{\pi}_D(a) \approx \#\{a_k^* = a\}/K$

SOLVING THE ATTACKER'S PROBLEM

Choice of random probabilities and loss F

- $\pi_A(\theta_0) = \alpha\,[\pi_D(\theta_0)]^\beta$ with $\alpha, \beta \sim \mathcal{U}(0,1)$
- $\pi_A(x|\theta)$ based on distortion of $\pi_D(x|\theta)$, i.e. c.d.f. $F_A(x|\theta) = [F_D(x|\theta)]^\gamma$ with random nonnegative γ
- $\pi_A(y|x,a)$ based on $\pi_D(y|x,a)$ with $y \to \delta y$, random nonnegative δ
- Parametric form for $L_A(d,\theta,a)$ with distribution over such parameters
- $\Pi_A(d|y)$ could be either specified or obtained through (k-level) strategic thinking leading to a hierarchy of decision making problems
- Approach applied to more complex tests
- Batch acceptance where D should decide whether to accept (d_0) or reject (d_1) a batch with acceptable and/or faulty items

DATA DISCLOSURE[3]

- The agency must decide which set of alternative forms of data is best to release, taking into account extent of disclosure and data utility associated with each form of data release
- In this scenario there are two set of users for released data: those interested in research or statistical studies (society), and those willing to disclose confidential information about data providers (intruder)
- The society is interested in an unknown quantity θ_S while the intruder is interested in an unknown quantity θ_I and they try to infer their target values based on the information contained in the released data
- The extent of disclosure measures how the intruder's inference about θ_I, based on released data, can harm data providers and/or the statistical agency
- This is balanced by the data utility that measures the extent to which the released data is useful to society
- The suggested optimal form of data release maximises the data utility among those releases whose extent of disclosure falls below a fixed threshold

[3] Trottini (2001) Research in Official Statistics

DATA DISCLOSURE

- Suppose that two data sets D_1 and D_2 are considered for release
- Suppose that the intruder's posterior distributions for θ_I given D_1 and D_2 are $\mathcal{N}(-5, 1)$ and $\mathcal{N}(5, 1)$, respectively, while the society's posterior distributions for θ_S given D_1 and D_2 are $\mathcal{N}(-7, 2)$ and $\mathcal{N}(7, 2)$
- Suppose also that the agency's posterior distributions for θ_I and θ_S given the original data are $\mathcal{N}(4.9, 0.001)$ and $\mathcal{N}(-6.9, 0.001)$, respectively, and that intruder, society and agency use a quadratic loss function
- Under quadratic error loss, the optimal action is the posterior mean and the uncertainty (i.e. the associated expected loss) is the posterior variance
- Thus the intruder's optimal estimates of θ_I when D_1 and D_2 are released are -5 and 5, respectively, while the uncertainty in both cases is 1
- Similarly, society's optimal estimates of θ_S when D_1 and D_2 are released are -7 and 7, respectively, and the uncertainty in both cases is 2
- Based on the measures of disclosure in Duncan and Lambert (1986) as adapted by Trottini (2001), the agency should flip a coin to decide which data set is best to release, since in either case the intruder's uncertainty is 1 and the society's uncertainty is 2, and D_1 and D_2 are perfectly equivalent

DATA DISCLOSURE

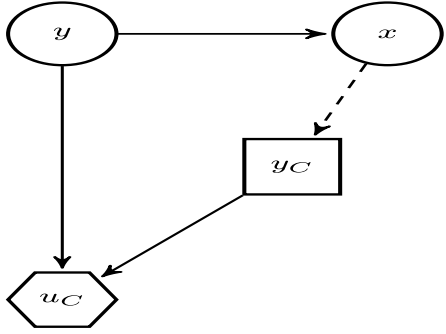
- However the agency posterior distribution for θ_I and θ_S are $\mathcal{N}(4.9, 0.001)$ and $\mathcal{N}(-6.9, 0.001)$
- The agency is very confident that the value of θ_I is approximately 4.9 (with 95% posterior credibility interval $[4.84, 4.96]$) and the value of θ_S is -6.9 (with 95% posterior credibility interval $[-6.96, -6.84]$)
- If D_2 is released the intruder's optimal estimate of θ_I is 5, which is very close to what the agency's believes to be the value of θ_I, while the society's optimal estimate of θ_S, 7, is very poor from the agency's point of view
- If D_1 is released, instead, the intruder's optimal estimate of θ_I, -5, is very inaccurate from the agency's point of view while the society's optimal estimate of θ_S, -7, is very close to what the agency believes to be the value of θ_S
- Thus if we take into account not only the intruder's and society's perspectives, but also the agency's knowledge of the intruder's and the society's targets θ_I and θ_S, the two forms of data release, D_1 and D_2, are no longer equivalent and D_1 should be intuitively released

DATA DISCLOSURE AND ARA

- In this approach, it was assumed that the agency had a perfect knowledge about the opinions of the society and the intruder on the parameters θ_I and θ_S, while it is more realistic to suppose that it has only guesses
- We can think of a decision by the agency (our Defender) about two alternatives, D_1 and D_2, when receiving some information, correct or misleading, from the intruder and/or the society
- Some links with adversarial hypothesis testing can be thought, although careful thinking might need to specify what is the true value generated by ”Nature” and how it is modified on purpose by intruder and, possibly, society
- Another difference is the presence of a third party, the society, which is supported by the Defender, i.e., agency

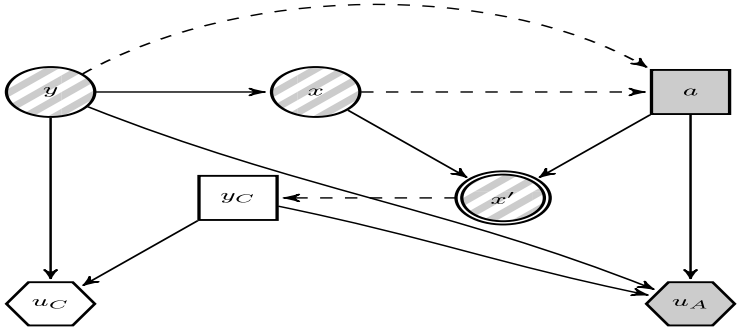
BINARY CLASSIFICATION

- Classifier C receives two types of objects: malicious ($y = +$) or innocent ($y = -$)
- Objects have features x whose distribution depends on their type y
- Classification problems broken down into two separate stages:
  - inference about $p_C(y|x)$, C's beliefs about type given features (avoided here)
  - decision about class assignment y_C, based on $p_C(y|x)$ and utility $u_C(y_C, y)$
- Node: decision (square), uncertainty (circle), deterministic (double), utility (hex.)
- Arrow: conditional relation (solid), information available at decision time (dashed)



ADVERSARIAL CLASSIFICATION

- Adversary A chooses attack a s.t. actual $x \to x' = a(x)$ observed by C
- A attacks only for malicious instances ($y = +$)
- Nodes in bi-agent influence diagram: grey (A), white (C), striped (both A and C)
- Decisions: attack a by A and classification y_C by C
- Utilities: $u_C(y_C, y)$ for C and $u_A(y_C, y, a)$ for A



CLASSIFIER PROBLEM

$$\begin{aligned}
\text{Find class } c(x') &= \arg\max_{y_C} \sum_{y \in \{+,-\}} u_C(y_C, y) p_C(y|x') \\
(\text{ignore } p(x')) &= \arg\max_{y_C} \Big[u_C(y_C, -) p_C(x'|-) p_C(-) \\
&+ u_C(y_C, +) p_C(+) \sum_{x \in \mathcal{X}'} p_C(a_{x \to x'}|x, +) p_C(x|+) \Big]
\end{aligned}$$

- Expected utility maximisation
- $\mathcal{A}(x)$: set of possible attacks for actual x
- $\mathcal{X}' = \{x : a(x) = x' \text{ for some } a \in \mathcal{A}(x)\}$: x's potentially leading to observed x'
- $p_C(y)$: beliefs about the class distribution
- $p_C(x|y)$: beliefs about feature distribution given the class (under no attacks)
- $u_C(y_C, y)$: utility in classifying y_C with actual y
- $p_C(a|x, y)$: beliefs about A's action, given x and y (Think of A's behaviour!)

ATTACKER PROBLEM

- Find A's optimal attack

$$
\begin{aligned}
a^*(x,y) &= \arg\max_a \int \Big[u_A(+,+,a)\ p + u_A(-,+,a)\ (1-p) \Big] f_A(p|a(x))dp \\
&= \arg\max_a \left[u_A(+,+,a) - u_A(-,+,a) \right] p^A_{a(x)} + u_A(-,+,a)
\end{aligned}
$$

- A: modify x so that C classifies malicious instances as innocent (A's maximum expected utility)
- A: modify only malicious instances, i.e. $y = +$, and not innocent, i.e. $y = -$
- C's decision: uncertain for A
- $u_A(y_C, y, a)$: utility for A when C says y_C, actual label is y and the attack is a
- $p_A(c(x')|x')$: A's beliefs about the classification result when C observes x'
- $p = p_A(c(a(x)) = +|a(x))$: A's beliefs about C classifying as malicious after observing $x' = a(x)$
- Uncertainty on p modelled via density $f_A(p|a(x))$ with expectation $p^A_{a(x)}$

CLASSIFIER PROBLEM

- Find $a^*(x, y) = \arg\max_a [u_A(+, +, a) - u_A(-, +, a)]\, p^A_{a(x)} + u_A(-, +, a)$
- C does not know A's utilities u_A and expectations $p^A_{a(x)}$
- C's uncertainty modelled through random utility U_A and random expectation $P^A_{a(x)}$
- Solve for the random optimal attack, optimising the random expected utility
  $A^*(x, +) = \arg\max_a \left([U_A(+, +, a) - U_A(-, +, a)]\, P^A_{a(x)} + U_A(-, +, a) \right)$
- $\Rightarrow p_C(a_{x \to x'} | x, +) \approx Pr(A^*(x, +) = a_{x \to x'})$, assuming a discrete set of attacks
- Approximation through simulation of K samples $\left(U^k_A(y_C, +, a), P^{A,k}_{a(x)}\right)$ from random utilities and probabilities
  $\Rightarrow A^*_k(x, +) = \arg\max_a \left(\left[U^k_A(+, +, a) - U^k_A(-, +, a)\right] P^{A,k}_{a(x)} + U^k_A(-, +, a) \right)$
- Estimation: $\widehat{p_C}(a_{x \to x'} \mid x, +) = \#\{A^*_k(x, +) = a_{x \to x'}\}/K$

RANDOM UTILITY

- Random utility $U_A(y_C, +, a)$ includes two components
  - A's gain from C's decision
  - random cost B of implementing an attack
- $Y_{y_C y}$: gain when C decides y_C with y actual label
- $-Y_{++} \sim Ga(\alpha_1, \beta_1)$ with expected gain $-\alpha_1/\beta_1$ for A and variance α_1/β_1^2
- $Y_{-+} \sim Ga(\alpha_2, \beta_2)$ with expected gain $\alpha_2/\beta_2 = e$ for A, and variance α_2/β_2^2
- $Y_{+-} = Y_{--} = \delta_0$, Dirac at 0: no gain for A from innocent instances
- $\Rightarrow$ A's gain $(Y_{y_C y} - B)$
- If A risk prone $\Rightarrow U_A(y_C, y, a) = \exp(\rho\,(Y_{y_C y} - B))$ with random risk proneness coefficient $\rho \sim U[a_1, a_2]$, $a_1 > 0$

RANDOM PROBABILITY

- $P^A_{a(x)}$, A's (random) expected probability that C classifies as malicious for $x' = a(x)$
- C guesses A's beliefs about C's classification when observing $x' \Rightarrow$ delicate
- Hierarchy of decisions: A should know what C does when knowing what A does ...
- Probabilities to be specified at each stage until no more available information $\Rightarrow$ non-informative distribution at that stage
- Heuristic at first stage based on $Pr_C(c(x') = +|x') = r$ (C classifies as malicious observing x'), with some uncertainty around it $\Rightarrow P^A_{a(x)} \sim \beta e(\delta_1, \delta_2)$, with mean $\delta_1/(\delta_1 + \delta_2) = r$ and adequate variance
- In general, given observed x', consider all instances leading to it
  - p_1: proportion of instances originally malicious
  - p_2: proportion of instance originally innocent
  - $\Rightarrow r = p_1/(p_1 + p_2)$

SPAM DETECTION

- m emails as *bag-of-words*: binary features about presence (1) or not (0) of n words
- Label indicates whether the message is spam $(+)$ or not $(-)$
- Email as n-dimensional vector $x = (x_1, x_2, .., x_n)$ of 0's or 1's, with label y
- Only word insertion attacks $\Rightarrow$ 0's replaced by 1's
- Interest in insertion of one word at most
- $I(x)$: set of indices s.t. $x_i = 0$ in $x \Rightarrow \mathcal{A}(x) = \{a_0, a_i; \forall i \in I(x)\}$ set of possible attacks with identity a_0 and a_i transforming i-th 0 into 1
- $J(x')$: set of indices with value 1 in x' received by $C \Rightarrow \mathcal{X}' = \{x', x'_j; \forall j \in J(x')\}$ and x'_j message potentially leading to x', with j-th 1 in x' replaced with 0

SPAM DETECTION

- $u_C(y_C, y)$ standard
- $p_C(y)$ and $p_C(x|y)$ standard if considering only exploratory attacks and using a generative classifier (i.e. based on a generative model $p_C(x, y)$) to estimate them
- Strategic component for $p_C(a_{x \to x'}|x, y)$ and use of ARA to approximate it
- Adversary's random utilities obtained as before
- Beta distribution for $P^A_{a(x)}$ with adequate variance and mean r_a
  - $q_0 = p_C(x'|-)p_C(-)$: original label $-$ left unchanged by A
  - $q_j = p_C(x'_j|+)p_C(+), \forall j \in J(x')$: original label $+$ changed by A
  - $q_{n+1} = p_C(x'|+)p_C(+)$: original label $+$ left unchanged by A
  - $r_a = \dfrac{\sum_{i \in J[a(x)]} q_i + q_{n+1}}{q_0 + \sum_{i \in J[a(x)]} q_i + q_{n+1}}$, i.e. mean set equal to C's own probability of classifying malicious for $x' = a(x)$
  - $P^A_{a(x)}$: A's (random) expected probability of C classifying malicious for $x' = a(x)$

SPAM DETECTION

- Spambase Data Set from UCI Machine Learning repository
  - 4601 emails, out of which 1813 are spam
  - 54 relevant words for each email $\Rightarrow$ 54 dimensional vector x of 0's and 1's
  - data randomly split into training (75%) and test (25%) sets, with 100 repetitions
- Training not affected by attacks $\Rightarrow \widehat{p}_C(y)$ and $\widehat{p}_C(x|y)$ from Naive Bayes classifier
- Simulations (sample size 1000) with 4 utilities for C and different variances for random expected probability $P^A_{a(x)}$ (increasing percentages k of maximum value)
- Comparison between ACRA and Naive Bayes: accuracy, utility, false positive (FPR) and false negative rates
- ACRA more robust w.r.t. attacks, identifying more attacked spam emails, even for larger k, i.e. variance, worsening the performance
- ACRA $\Rightarrow$ lower FPR, i.e. less non-spam are rejected as spam (more important than accepting spam)

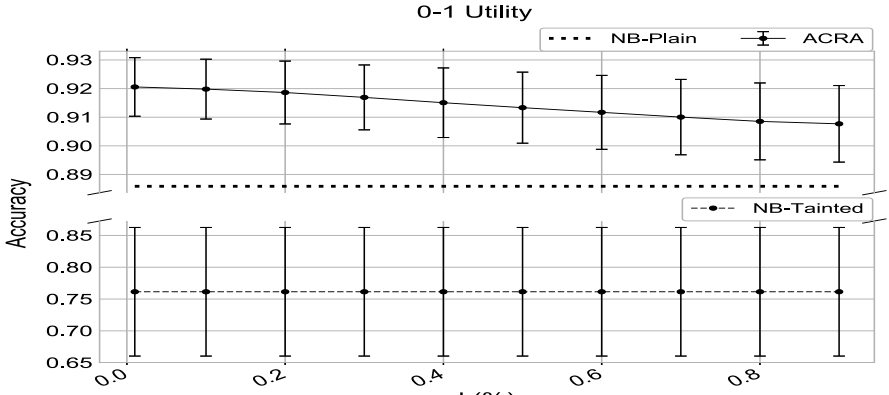
SPAM DETECTION

- Checking utility robustness through 4 utilities for C:
  - $0/1$ Utility $\Rightarrow$ 1 if correctly classified and 0 o.w.
  - Three utilities taking values
    - 1 if correctly classified
    - -1 for spam classified as legit
    - $-2/-5/-10$ for legit classified as spam
- Random utilities for A (m=mean, v=variance)
  - $-U_A(+,+,a) \sim Ga(2500, 0.002) \Rightarrow m = 5, v = 0.01$
  - $U_A(-,+,a) \sim Ga(2500, 0.002) \Rightarrow m = 5, v = 0.01$
  - $U_A(-,-,a) = U_A(+,-,a) = \delta_0$
- Random cost $B = d(a) \cdot \alpha$, with $d(a) = \#$ word changes and $\alpha \sim U[0.4, 0.6]$
- Random risk proneness coefficient $\rho \sim U[0.4, 0.6]$

SPAM DETECTION

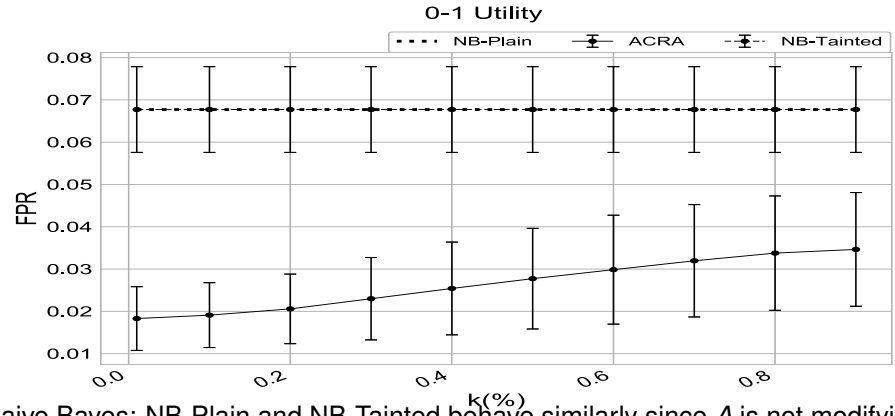
- Beta distribution for $P^A_{a(x)}$ with mean $r = Pr_C(c(a(x)) = +|a(x))$
  - Concave to avoid malicious $a(x)$ concentrated around 0 or 1
  - $\Rightarrow$ variance $\leq \Delta = \min\left\{[r^2(1-r)]/(1+r), [r(1-r)^2]/(2-r)\right\}$
  - Adjustable variance at $k\Delta$ with $k \in \{0.01, 0.1, 0.2, \cdots, 0.9\}$
- $K = 1000$ Monte Carlo sample size

SPAM DETECTION



- Starting problem for C: find $c(x') = \arg\max_{y_C} \sum_{y \in \{+,-\}} u_C(y_C, y) p_C(y|x')$
- $0/1$ utility function, i.e. 1 for correctly classified instance and 0 otherwise
- Naive Bayes: NB-Plain for original data and NB-Tainted for attacked data
- k: percentage of maximum variance for $P^A_{a(x)}$

SPAM DETECTION



- Naive Bayes: NB-Plain and NB-Tainted behave similarly since A is not modifying innocent instances
- Increasing k (and variance for $P^A_{a(x)}$) $\Rightarrow$ increases FPR
- Reducing FPR crucial in spam detection, as filtering out a non-spam is worse than letting spam reach the user

DISCUSSION ABOUT ACRA

- So far ACRA tested with A's distributions centered around the expected values of C's, but it proves quite robust even when moving away
- Changing all words in the spam detection problem $\Rightarrow 2^n$ possible attacks
  - Ad hoc procedure, e.g., changing only one word and from 0 to 1
  - Smaller sample size
  - Approximations, parallelisation
- Further extensions
  - From binary classification to multi-label (e.g. malware: trojan, adware, virus)
  - From exploratory to poisoning attacks, i.e. attacks also during training
  - Attacks not only on malicious instances but also on innocent ones
  - From generative classifiers ($P(X, Y)$) to discriminative ones ($P(Y|X = x)$)

DISCRIMINATIVE CLASSIFIERS

- In the earlier approach (generative classifier) we supposed to know $p(y)$ and $p(x|y)$, e.g. from a classifier applied to the training set
- Here we suppose to know only $p(y|x)$, $y \in \{1, \ldots, k\}$, and address the problem of classifying an instance when x' is observed $\Rightarrow$ solve $\arg\max_{y_C} \psi(y_C)$ where

$$\begin{aligned} \psi(y_C) &= \int_{\mathcal{X}_{x'}} \left(\sum_{y=1}^{k} u(y_C, y) p(y|x = a^{-1}(x')) \right) p(x|x')dx \\ &= \sum_{y=1}^{k} u(y_C, y) \left[\int_{\mathcal{X}_{x'}} p(y|x = a^{-1}(x')) p(x|x')dx \right] \end{aligned}$$

- $p(y|x)$ is based on untainted x
- $\mathcal{X}_{x'}$, the set of reasonable instances x leading to x' if attacked
- Optimisation solved via Monte Carlo using sample $\{x_n\}_{n=1}^{N}$ from $p(x|x')$ but ...
- ... there is a problem: we do not know $p(x|x')$ and we have to estimate it

AB-ACRA

- Suppose $p(x'|x)$ known as result of strategic thinking, as before, about the possible attacks
- Efficient approach to sample from $p(x|x')$ making use of samples from $p(x'|x)$
- Sample from $p(x|x') \propto p(x'|x)p(x)$ for x and x' discrete
  - Proposal $\tilde{x}$ from transition distribution $q(x \to \tilde{x})$
  - Sampled $\tilde{x}' \sim p(X'|X = \tilde{x})$
  - $\Rightarrow$ accept $\tilde{x}$ if $\tilde{x}' = x'$ with probability $\alpha = \min\left\{1, \frac{p(\tilde{x})q(\tilde{x} \to x_i)}{p(x_i)q(x_i \to \tilde{x})}\right\}$
  - Very slow convergence
- Sample from $p(x|x')$ for x and x' continuous
  - $\tilde{x}$ and $\tilde{x}'$ generated as above
  - Based on Approximate Bayesian Computation (ABC) techniques, accept $\tilde{x}$ if $\phi(\tilde{x}', x') < \epsilon$ for a given distance ϕ and tolerance ϵ
  - For high dimensions, use summary statistics s to accept $\tilde{x}$ if $\phi(s(\tilde{x}'), s(x')) < \epsilon$

ROBUSTIFYING DURING TRAINING

- $p(y|x, \beta)$: model for instances y given covariates x and parameter β
- Training data $\mathcal{D} = \{(x_i, y_i)_{i=1}^N\}$ and prior $p(\beta)$
  $\Rightarrow$ posterior $p(\beta|\mathcal{D})$ and predictive $p(y|x, \mathcal{D})$
- Using $0-1$ utilities $\Rightarrow \arg\max_{y_C} \int\int p(y_C|x, \beta)p(x|x')p(\beta|\mathcal{D})dx\, d\beta$
- Given a sample x, adversarial perturbation $x' \sim p(x'|x)$ and adequate objective function $\mathcal{L}(\beta, x, y)$ (mathematical details and conditions skipped)
  $\Rightarrow$
  - $p(y, x|\beta) \approx p(y, x'|\beta)$, i.e. x and x' affect y similarly
  - $p(y|x, \beta) \approx p(y|x', \beta) \Rightarrow \arg\max_{y_C} \int\int p(y_C|x', \beta)p(x'|x)p(\beta|\mathcal{D})dx'\, d\beta$
  - $p(x|\beta) \approx p(x'|\beta) \Rightarrow \arg\max_{y_C} \int\int p(y_C|x', \beta)p(\beta|\mathcal{D})d\beta$

ROBUSTIFYING DURING TRAINING

Classifier	**Attacked (Raw)** Attacker A	**Attacked (CK)** Attacker A	**Attacked (AB-ACRA)** Attacker A
Logistic Regression	0.315 ± 0.007	0.499 ± 0.008	**0.513** ± 0.008
Naive Bayes	0.325 ± 0.007	0.645 ± 0.025	**0.665** ± 0.024
Neural Network	0.389 ± 0.024	0.592 ± 0.032	**0.638** ± 0.030
Random Forest	0.313 ± 0.009	**0.720** ± 0.013	0.710 ± 0.017

Classifier	**Attacked (Raw)** Attacker B	**Attacked (CK)** Attacker B	**Attacked (AB-ACRA)** Attacker B
Logistic Regression	0.412 ± 0.004	0.713 ± 0.008	**0.760** ± 0.011
Naive Bayes	0.406 ± 0.008	0.783 ± 0.039	**0.800** ± 0.035
Neural Network	0.437 ± 0.014	**0.727** ± 0.050	0.725 ± 0.052
Random Forest	0.402 ± 0.005	0.779 ± 0.011	**0.782** ± 0.008

- Accuracy comparison (with precision) of four classifiers on attacked data with no defense, CK (Common Knowledge) defense and AB-ACRA defense

CONCLUSIONS ABOUT ACRA

- Here more emphasis on modelling and conceptual aspects while the paper contains many details about algorithmic ones and comparisons with classical classifiers
- Like in ABC, the choice of summary statistics in AB-ACRA might be critical
- AB-ACRA and ACRA become computationally expensive for large scale problems $\Rightarrow$ differentiable classifiers as an alternative (they are based on gradients of classifiers)
- Adaptive attackers can be dealt with changing random probability and random utility accordingly
- Here we have considered attacks to i.i.d. sequences but data could come, say, from an autoregressive model

ADVERSARIAL CLASSIFICATION AND OFFICIAL STATISTICS

- An important problem is represented by attacks on synthetic data aimed at determining if an item comes from the original dataset T or the synthetic dataset S
- Some scenarios could be addressed with different ARA approaches, e.g. when the attacker knows only S, only S and a subset R of T, or both S and R and the generative model G
- An important aspect of ARA concerns robustification during the training phase of algorithms, considering algorithms for de-identification of data sets or the creation of synthetic data, making them more robust against possible attacks
- One class of complex problems of large interest, both conceptually and from an applied point of view, refers to differential games in which the states of agents and their payoffs evolve continuously according to differential equations (and depend on each other strategies)

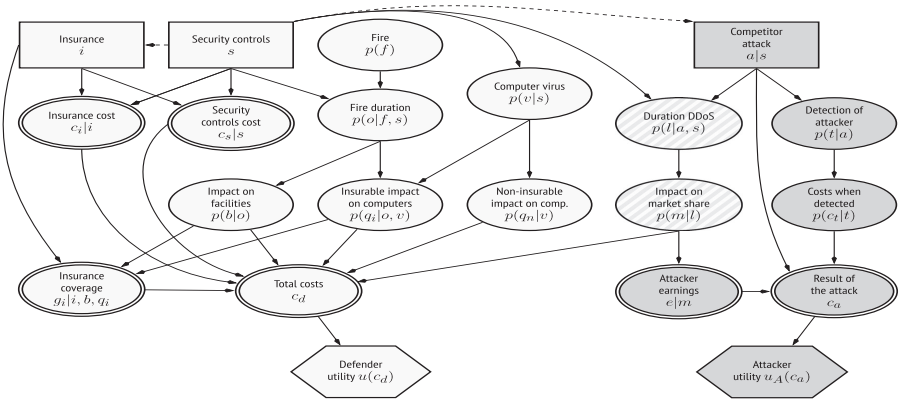
CYBERSECURITY

From an ISTAT report on BUSINESSES AND ICT — 2022, dated 3/1/2023:

In 2022, 69.9% of small and medium-sized enterprises (SMEs) will adopt at least four out of 12 digital activities (69.1% in the EU27). The increased use of remote working, which involved more than seven out of ten companies in 2022, has influenced the increase in the share of companies with at least 10 employees that have documents on cybersecurity measures, practices, or procedures (48.3%, up from 34.4% in 2019; 37% in the EU27). Cybersecurity is a concern for 45.1% of larger companies, which have taken out cybersecurity insurance to protect themselves. Among smaller companies, the share is 14.4% (44.6% and 22.6%, respectively, in the EU27). The adoption of robotics and the employment of ICT specialists remain stable compared to 2020.

For each survey, corporate behaviour is assessed against 12 specific characteristics that contribute to the definition of the composite digitalisation indicator called the Digital Intensity Index (DII), used to identify the areas in which Italian and European companies encounter the greatest difficulties. With regard to the 12 indicators by employee size, the most significant gaps are found, to the detriment of SMEs (companies with 10-249 employees), in the presence of ICT specialists, the decision to invest in ICT training in the previous year, the use of online meetings, and specialized documentation on cybersecurity standards and measures.

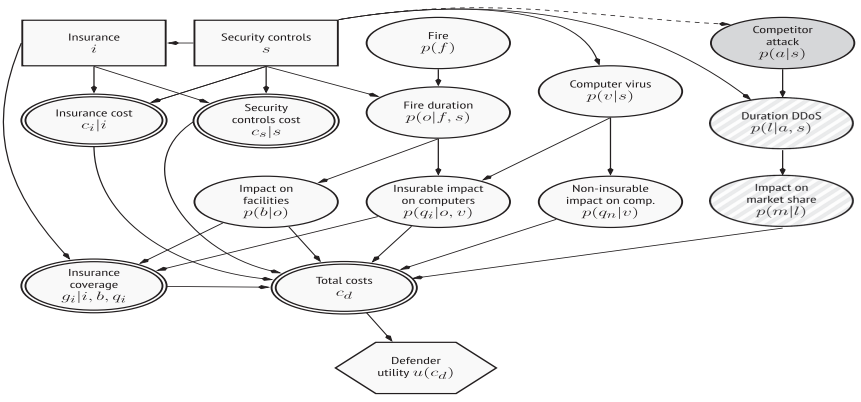
CYBERSECURITY[4]



BAID for cybersecurity risk management

[4] From Rios Insua, Couce-Vieira, Rubio, Pieters, Labunets, Rasines (2021), Risk Analysis

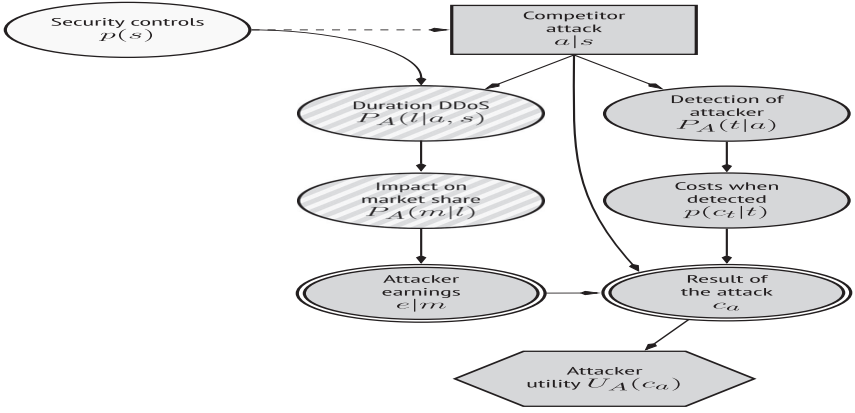
CYBERSECURITY[5]



Cybersecurity risk management: Defender viewpoint

[5] From Rios Insua, Couce-Vieira, Rubio, Pieters, Labunets, Rasines (2021), Risk Analysis

CYBERSECURITY[6]

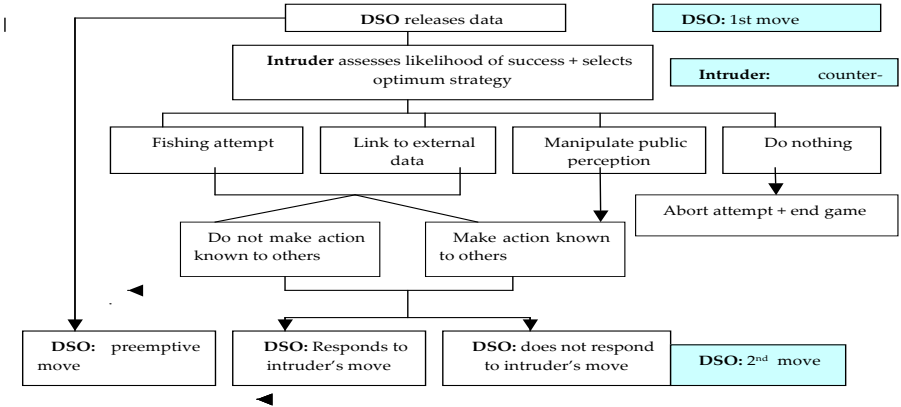


Cybersecurity risk management: Attacker viewpoint

[6] From Rios Insua, Couce-Vieira, Rubio, Pieters, Labunets, Rasines (2021), Risk Analysis

SEQUENTIAL GAME[7]

- Statistical Disclosure Control to ensure balance between confidentiality and data utility
- Disclosure event from interaction between agents: data stewardship organisation (DSO) and intruder

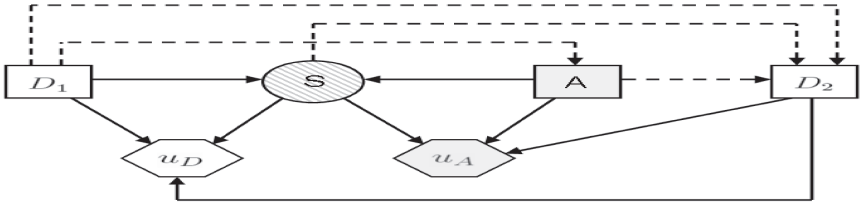


[7] From Mackey and Elliot (2011), Transactions in Data Privacy

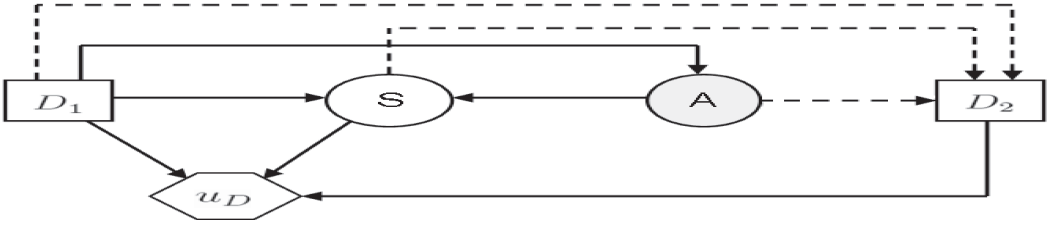
SEQUENTIAL GAMES

- We now present a Defend-Attack-Defend game
- In this game the Defender (she) first deploys her defensive resources, then the Attacker (he) observes the deployment and chooses his attack and, finally, the Defender attempts to recover from the attack as best she can
- We present both an influence diagram and a game tree
  - Nodes D_1 and D_2 correspond to the Defender's first and second decisions, i.e., $d_1 \in \mathcal{D}_1$ and $d_2 \in \mathcal{D}_2$, respectively
  - Node A represents the Attacker's decision $a \in \mathcal{A}$
  - S is a shared uncertainty node about the success of the attack

SEQUENTIAL GAMES



(a) Multi-agent influence diagram for the Defend-Attack-Defend game.



(b) The Defend-Attack-Defend game tree.

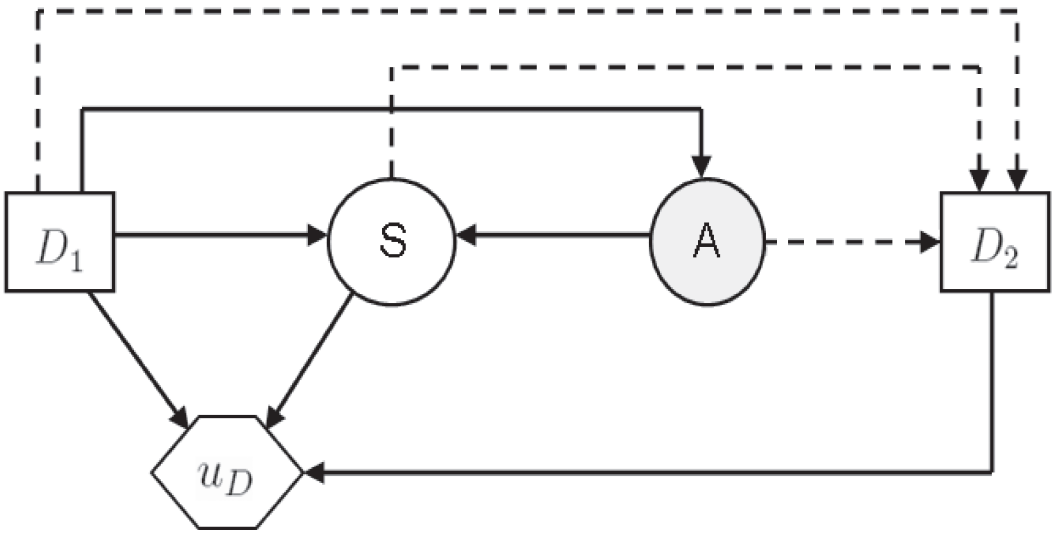
SEQUENTIAL GAMES

- In the previous slide the MAID shows the information available to each opponent at each stage of the game, and the relationships between utilities, decisions, and chance, while the game tree shows the sequence of play, which is helpful in backwards induction (the computational technique used to find the optimal decisions)
- As written, this model assumes that the only relevant uncertainty is the success level S of the attack, which depends probabilistically on $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$
- The payoff to the Defender depends on (d_1, s, d_2): the cost of her initial defence, the success of the attack, and the success of the recovery effort
- The payoff to the Attacker depends on (a, s, d_2): the effort in mounting his attack, its success, and the success of her recovery effort
- The model may be easily generalised so that the outcome of the recovery effort is also a random variable

SEQUENTIAL GAMES

- Game theory needs the Defender to know the Attacker's utilities and probabilities, and the Attacker to know the Defender's utilities and probabilities, and for both to know that these are common knowledge
- Often, the utility functions and probability assessments depend upon all of d_1, s, a, and d_2 (i.e., the utility and probability are affected not only by the chance outcome, but also by the choices of each opponent)
- However here we make the simplifying assumption that utilities and probabilities depend only upon the outcome of the opponent's decision, but not also on the decision itself
- In such cases, denote the Defender's and Attacker's utility functions by $u_D(d_1, s, d_2)$ and $u_A(a, s, d_2)$, respectively, and their probability assessments about the success of attack by $p_D(S = s|d_1, a)$ and $p_A(S = s|d_1, a)$, respectively

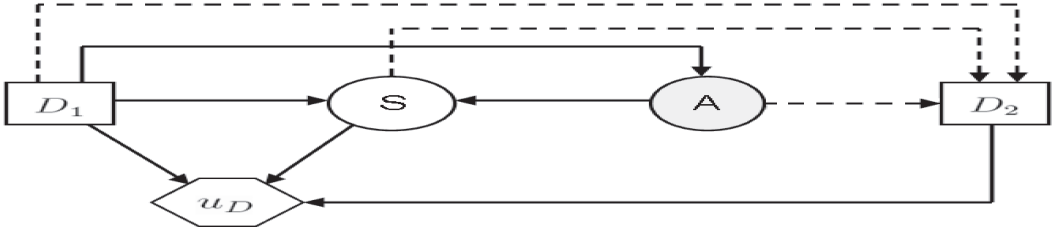
SEQUENTIAL GAMES



SEQUENTIAL GAMES

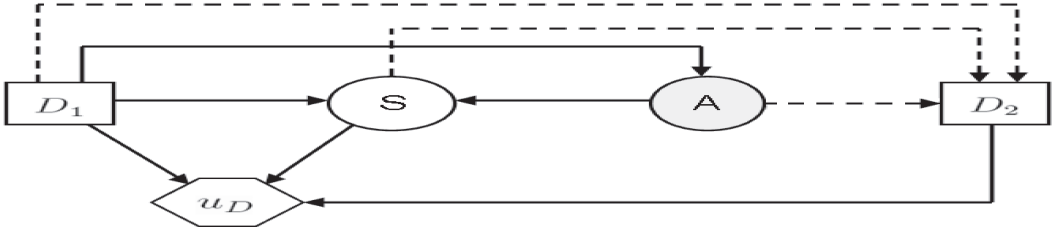
- Intuition would suggest to choose first the Defender's initial optimal defence d_1^* , then the optimal attack $a^*(d_1^*)$ after observing d_1^* and, finally, the optimal recovery action $d_2^*(d_1^*, s)$, after observing the consequence s of the previous actions
- But this is computationally unfeasible $\Rightarrow$ go backwards and use backwards induction
  - Find the optimal defence $d_2^*(d_1, s)$ for any pair (d_1, s)
  - For each pair (d_1, a), both Defender and Attacker can compute their expected utilities, since they have specified their own distributions of the consequences s for each (d_1, a) and their own utilities associated to their actions (d_1 for Defender and a for Attacker, consequence s and optimal decision d_2^*)
  - Both expected utilities are known to both opponents (common knowledge!)
  - For each initial decision d_1, the Attacker finds his optimal attack $a^*(d_1)$
  - Finally, the Defender chooses her optimal initial defence d_1^*, knowing which optimal attack will be chosen by the opponent and her optimal recovery action

SEQUENTIAL GAMES



- Using backwards induction, at node D_2 for the game tree above, the Defender's best response after observing $(d_1, s) \in \mathcal{D}_1 \times \mathcal{S}$ is $d_2^*(d_1, s) = \underset{d_2 \in \mathcal{D}_2}{\arg\max}\, u_D(d_1, s, d_2)$
- Under the common knowledge assumption, the Defender's behaviour at D_2 will be anticipated by the Attacker
- At node S, Defender's expected utility associated with each $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$: $\psi_D(d_1, a) = \int u_D(d_1, s, d_2^*(d_1, s)) p_D(S = s | d_1, a) ds$
- At node S, Attacker's expected utility associated with each $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$: $\psi_A(d_1, a) = \int u_A(a, s, d_2^*(d_1, s)) p_A(S = s | d_1, a) ds$
- Both expected utilities are known to both opponents

SEQUENTIAL GAMES

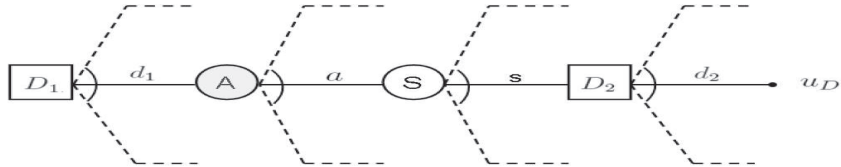


- The Attacker now finds his optimal attack at node A, after observing the Defender's move $d_1 \in \mathcal{D}_1$, by solving $a^*(d_1) = \arg\max_{a \in \mathcal{A}} \psi_A(d_1, a)$
- Knowing this, the Defender finds her maximum expected utility decision at node D_1 through $d_1^* = \arg\max_{d_1 \in \mathcal{D}_1} \psi_D(d_1, a^*(d_1))$, which gives the solution to the game
- Assuming common knowledge, rational players, and perfect information, Game Theory prescribes that
  - The Defender should choose $d_1^* \in \mathcal{D}_1$ at node D_1
  - The Attacker should choose attack $a^*(d_1^*) \in \mathcal{A}$ at node A after observing d_1^*
  - The Defender, after observing $s \in \mathcal{S}$, should choose $d_2^*(d_1^*, s) \in \mathcal{D}_2$ at node D_2

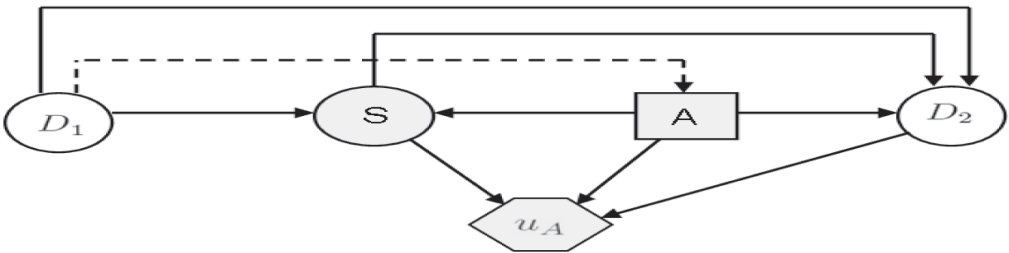
ARA FOR SEQUENTIAL GAMES

- ARA can be used in sequential games, and it allows one to drop the assumption that all utilities and probabilities are common knowledge
- From the ARA perspective, in the Defend-Attack-Defend model, the Attacker's decision at node A is uncertain to the Defender, and she must model her uncertainty through a random variable
- This is reflected in the influence diagram and the game tree (next slide), where the Attacker's decision node A has been converted to a chance node
- The Defender needs to assess $p_D(a|d_1)$, her predictive distribution about which attack a will be chosen at node A for each $d_1 \in \mathcal{D}_1$
- Additionally, she must make the (more standard) assessments about $u_D(d_1, s, d_2)$ and $p_D(s|d_1, a)$

ARA FOR SEQUENTIAL GAMES

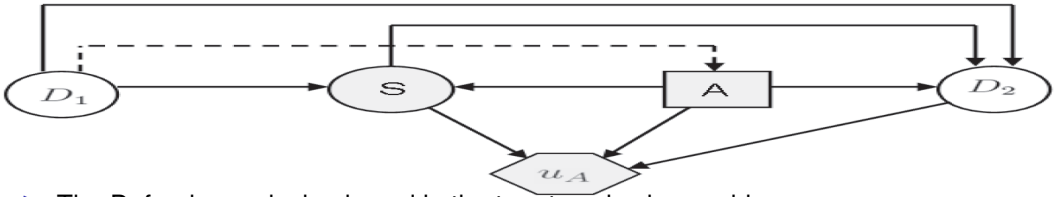


(a) The influence diagram for the ARA solution of the Defend-Attack-Defend game.



(b) The game tree for the ARA solution of the Defend-Attack-Defend game.

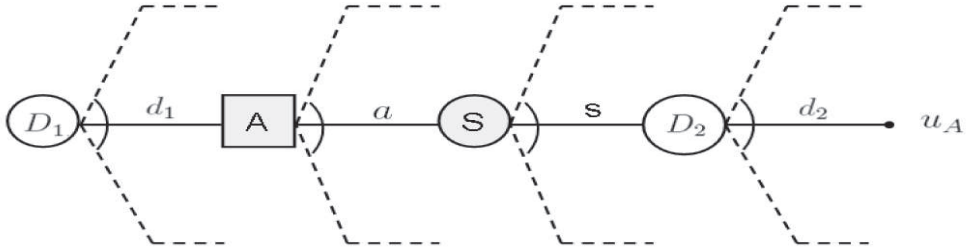
ARA FOR SEQUENTIAL GAMES



- The Defender works backward in the tree to solve her problem
- At node D_2 she computes her optimal action $d_2^*(d_1, s)$ for each $(d_1, s) \in \mathcal{D}_1 \times \mathcal{S}$, by $d_2^*(d_1, s) = \underset{d_2 \in \mathcal{D}_2}{\arg\max}\, u_D(d_1, s, d_2)$
- Then, at node S, for each $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$, she finds her expected utility: $\psi_D(d_1, a) = \int u_D(d_1, s, d_2^*(d_1, s)) p_D(S = s | d_1, a) ds$, for each $(d_1, a) \in \mathcal{D}_1 \times \mathcal{A}$
- Then, her subjective probability assessment of what the Attacker will do, $p_D(A|d_1)$, is used to compute her expected utility at node A for each $d_1 \in \mathcal{D}_1$: $\psi_D(d_1) = \int \psi_D(d_1, a) p_D(A = a | d_1) da$
- The Defender finds her maximum expected utility at D_1 as $d_1^* = \underset{d_1 \in \mathcal{D}_1}{\arg\max}\, \psi_D(d_1)$
- Thus, backwards induction shows that the Defender's best strategy is to choose first d_1^* at node D_1, and later, after observing $s \in \mathcal{S}$, choose $d_2^*(d_1^*, s)$ at node D_2

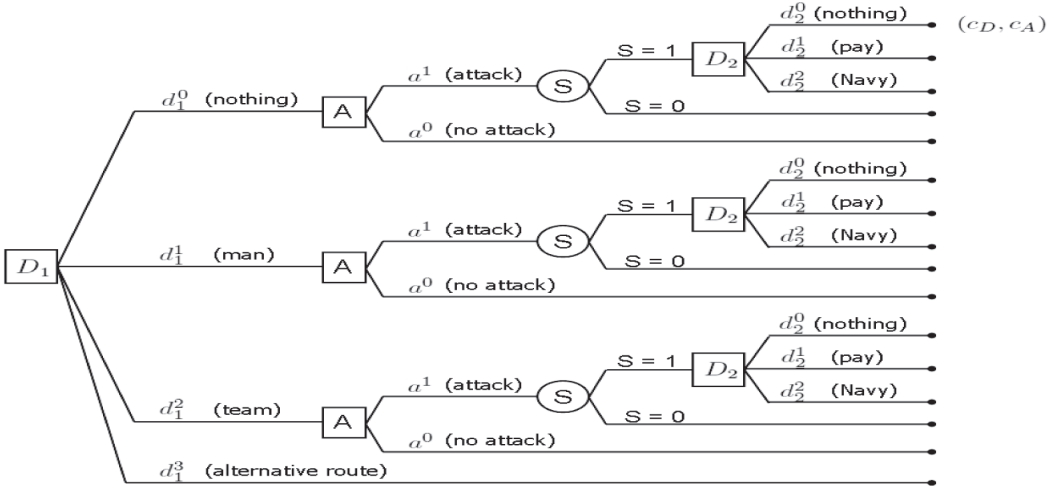
ARA FOR SEQUENTIAL GAMES

- As usual, ARA requires the decision maker to assess $p_D(a|d_1)$. This could be done through some form of risk analysis (non-strategic behaviour), or by developing a model for the strategic analysis of the opponent
- Given $p_D(a|d_1)$, the analysis of the Attacker's decision problem, as seen by the Defender, is shown in the next two figures, where the Attacker's probabilities and utilities are assessed by the Defender



(a) Influence diagram for the Defender's view of the Attacker's problem.

ARA FOR SEQUENTIAL GAMES



(b) Game tree for the Defender's view of the Attacker's problem.

ARA FOR SEQUENTIAL GAMES

- The assessment of the Attacker's probabilities and utilities may be based upon historical data and expert opinion, or, if those salient information is not available, then the Defender may choose to use a noninformative distribution $p_D(a|d_1)$
- In the former case, in order to elicit $p_D(a|d_1)$, the Defender must assess $u_A(a, s, d_2)$, $p_A(s|d_1, a)$, and $p_A(d_2|d_1, a, s)$, i.e., Attacker's probabilities and utilities
- In general, she does not know these quantities, and represents her uncertainty through a joint distribution F on the random quantities $U_A(a, s, d_2)$, $P_A(s|d_1, a)$, and $P_A(d_2|d_1, a, s)$
- These distributions might be elicited in various ways, e.g. Dirichlet process centered at Defender's probabilities
- The Defender then solves her perception of the Attacker's decision problem using backwards induction over the game tree, propagating her uncertainty, encoded by F, to get distributions over the random action $A^*(d_1)$ for each d_1

ARA FOR SEQUENTIAL GAMES

Specifically, if all choice sets are continuous, the Defender solves as follows

- At decision node D_2, compute the random expected utilities $(d_1, a, s) \to \Psi_A(d_1, a, s) = \int U_A(a, s, d_2) P_A(D_2 = d_2 | d_1, a, s) dd_2$
- At chance node S, compute the random expected utilities $(d_1, a) \to \Psi_A(d_1, a) = \int \Psi_A(d_1, a, s) P_A(S = s | d_1, a) ds$
- At decision node A , compute the random optimal initial decision $d_1 \to A^*(d_1) = \underset{a \in \mathcal{A}}{\arg\max}\, \Psi_A(d_1, a)$
- Thus, the Defender's predictive distribution over attacks, conditional on her first defensive decision d_1, is given by $\int_0^a p_D(A = x | d_1) dx = Pr[A^*(d_1) \leq a]$
- If the choice sets are not continuous, the Defender would reason similarly, but replacing integrals with sums and finding the predictive distribution over a set

ARA FOR SEQUENTIAL GAMES

- When the beliefs are complex, the Defender's predictive distribution can be approximated using Monte Carlo simulation
- Specifically, for each $d_1 \in \mathcal{D}_1$, the following algorithm finds the Defender's belief about the probability of each attack

Do, for $i = 1, \ldots, N$:

Draw $(U_A^i(a,s,d_2), P_A^i(S \mid d_1, a), P_A^i(D_2 \mid d_1, a, s)) \sim F$.

At chance node D_2, compute

$$(d_1, a, s) \rightarrow \psi_A^i(d_1, a, s) = \int U_A^i(a,s,d_2)\, P_A^i(D_2 = d_2 \mid d_1, a, s)\, dd_2.$$

At chance node S, compute

$$(d_1, a) \rightarrow \psi_A^i(d_1, a) = \int \psi_A^i(d_1, a, s)\, P_A^i(S = s \mid d_1, a)\, ds.$$

At decision node A, compute

$$d_1 \rightarrow a_i^*(d_1) = \text{argmax}_{a \in \mathcal{A}}\, \psi_A^i(d_1, a).$$

Approximate $\int_0^a p_D(A = x \mid d_1)dx$ **by** $\#\{a_i^*(d_1) \ : \ a_i^*(d_1) \leq a\}/N$
where $\#\{\cdot\}$ **is the cardinality of the set.**

ARA FOR SEQUENTIAL GAMES

- Given F, the Defender's assessment of $p_D(a|d_1)$ is straightforward
- In many situations, it is relatively simple to elicit the random utility $U_A(a, s, d_2)$ and the random probability $P_A(s|d_1, a)$ in F
- However, the assessment of $P_A(d_2|d_1, a, s)$ within F can be problematic; the Defender should exploit any information she has on how the Attacker analyses her decision problem
- She may think he seeks a Nash equilibrium, or uses level-k thinking; if she has little insight, she can represent that through some high-variance or non-informative distribution

FINAL THOUGHTS

- It might be worthwhile to reverse the ARA paradigm, choosing to side with the attacker and attempting actions aimed at disclosing personal data from the de-identified data. In this case, *A* could try various attacks and think about possible defences by *D* and see if she can achieve her goal. This approach could be useful for testing the level of protection of such data, hoping for a loss for *A*.
- National statistical offices, like ISTAT, are faced with the problem of people choosing to provide incorrect answers to questionnaires, either strategically (for example, to avoid disclosing undeclared property or illegal income) or randomly. Framing the problem within ARA could be one option, although it may not be simple and unlikely to improve upon existing techniques
- In literature it is mentioned that cryptography could be used for communications in federated machine learning, when information are coming from different sources which cannot share data. We are planning to consider adversarial cryptography with text encrypted with wavelet transform and then decrypted with inverse wavelet transform, with an attacker attempting to modify wavelet coefficients and/or decryption key during data transmission.
- Another area of possible interest is represented by attacks in federated learning

REFERENCES

- Banks, D., Rios, J., and Rios Insua, D. (2015). *Adversarial Risk Analysis* (Vol. 343). CRC Press.
- Rios Insua, D., Rios, J. and Banks, D. (2009). Adversarial Risk Analysis. *Journal of the American Statistical Association*, 104, 841-854.
- Gonzalez-Ortega, J., Soyer, R., Rios Insua, D. and Ruggeri, F. (2021), An Adversarial Risk Analysis Framework for Batch Acceptance Problem. *Decision Analysis*, 18, 25-40.
- Rios Insua, D., Ruggeri, F., Soyer, R. and Rasines, D.G. (2018), Adversarial Issues in Reliability. *European Journal of Operational Research*, 266, 1113-1119.
- Rios Insua, D., Ruggeri, F., Soyer, R. and Wilson S. (2020), Advances in Bayesian Decision Making in Reliability. *European Journal of Operational Research*, 282, 1-18.

REFERENCES

- Gonzalez-Ortega, J., Rios Insua, D., Ruggeri, F. and Soyer, R. (2021), Hypothesis Testing in Presence of Adversaries. *The American Statistician*, 75, 31-40.
- Naveiro, R., Redondo, A., Rios Insua, D. and Ruggeri, F. (2019), Adversarial Classification: An Adversarial Risk Analysis Approach. *International Journal of Approximate Reasoning*, 113, 133-148.
- Gallego, V., Naveiro, R., Redondo, A., Rios Insua, D. and Ruggeri, F. (2024), Protecting Classifiers From Attacks. *Statistical Science*, 39, 449-468.
- Soyer, R., Ruggeri, F., Rios Insua, D., Pierce, C. and Guevara, C. (2025), An Adversarial Risk Analysis Framework for Software Release. To appear in *Risk Analysis*.
- Rios Insua, D., Ruggeri, F., Alfaro, C. and Gomez, J. (2016), Robustness for Adversarial Risk Analysis. In *Robustness Analysis in Decision Aiding, Optimization and Analytics*, M. Doumpos, C. Zopounidis and E. Grigoroudis Eds., Springer, 19-58.
- Yang, J., Joshi, C. and Ruggeri, F. (2024), On Global Robustness of an Adversarial Risk Analysis Solution. *Statistica Neerlandica*, 78, 776-795.