MAGNET contributions in HYAIAI

Jan Ramon

Paris, June 2023

Contents

Administration and funding

- Timing: quite new line of work $\Rightarrow$ slow start, many branches of future work.
- Funding:
  - Hyaiai: in the form of 'joint students or post-docs' $\rightarrow$ many constraints, not many candidates
    - e.g., Carlos Cotrini (Magnet-Lacodam) stayed only 6 months.
    - e.g., Magnet-Tau developed post-doc proposal with FlandersMake on verification of self-driving vehicles but got no (good) candidate
  - Other funding: ANR, MEL, region, I-Site, HE, PEPR ...

Administration and funding

- Work overview:

Topic	Funding
Interpretable privacy requirements	Region + ANR
Tailored noise	FRM
Privacy-preserving negotiators	HYAIAI
Interpretable privacy metrics in medicine	HE + HYAIAI
Declarative approach to decentralized algs	MEL/ANR/I-SITE + PEPR
Verification of decentralized algorithms	ULille/HE

- HYAIAI funded a bit & greatly helped coordination (meetings, discussions, ideas, pointers)

Introduction: interpretable trust

- We want to make AI explainable, reliable, resilient, accurate, secure, transparent . . . but is such **large / complete** system still understandable?
- We need to explain not only the algorithm or learned model, **but also why algorithms are trustworthy**
- Let's consider **interpretable privacy**: understand how privacy is / can be guaranteed without in-depth knowledge of (ϵ, δ)-differential privacy, (ϵ, α)-Renyi privacy, pufferfish privacy, indistinguishability, composition rules, etc.

Interpretable privacy requirements

- Objectives:
  - Interpretability **for end-users**: Why can I trust this system will protect my privacy?
  - Interpretability **for developers**: Building privacy-preserving system without a PhD in privacy & cryptography.
- How?
  - Specify **privacy requirements**
  - Let system work out details of the privacy defenses to be implemented (developer)
  - Let system generate a proof that the implementation preserves privacy.

– Basu and Ramon, Interpretable privacy with optimizable utility,XKDD-2021, LNCS
– Journal paper and PhD thesis in preparation (Basu, Cotrini & Ramon)

Interpretable privacy requirements

- A declarative approach:
  - Specify input and output
  - Specify privacy requirements
  - The system adds the optimal (minimal / sufficient) amount of noise.
- An example:
  - Input $x \in \mathbb{R}^n$, output $y \in \mathbb{R}^m$ with $y = Ax$ with $A \in \mathbb{R}^{m \times n}$.
  - Privacy requirements: each x_i should remain private.
  - Model: compute $\hat{y} = A(x + \eta) + \xi$
  - Constraint program:

$$b_i^\top (A \mathrm{diag}(\sigma_\eta) A^\top + \mathrm{diag}(\sigma_\xi))^{-1} b_i \geq \frac{2 \log(1.25/\delta)}{\epsilon^2}$$

Tailored noise

- Same principle, other question
- Shall we use Gaussian, Laplacian, binomial, Poisson or another mechanism?
- Just specify the application, automate the selection with a numerical method.
- Especially important when noise values are post-processed, e.g., if functions with steep derivatives are applied, e.g., $1/x$, $\tan(x)$, $\log(x)$

– Pleska, A. PhD thesis, Chapter 4, May 2023
– article in preparation (Basu, Pleska & Ramon)

Tailored noise

- Input: data set $X \in \mathbb{R}^{n \times d}$
- Output $y = f(X) = h(\sum_{i=1}^{n} g(X_{i,:})) \in \mathbb{R}^m$ where g and h are non-linear functions.
- We want private (noisy) y with minimal error $\Rightarrow$ avoid regions in domain of f with steep gradient.
- Set $P(\hat{X}_{i,:} = \hat{v} | X_{i,:} = v) = p_{v,\hat{v}}$ with

$$p_{v,\hat{v}_1} \leq \epsilon p_{v,\hat{v}_2} + \delta \qquad \text{(private)}$$

and minimize $\mathbb{E}\left[\|X - \hat{X}\|_2^2\right]$

A declarative approach to decentralized algorithms

- Decentralized algorithms: more resilient than centralized.
- Gossip algorithm: propagate information through the network
  - Can be made very resilient against attacks [Sabater et al. MLJ 2022] [Sabater et al; PETS 2023]
- Can we do better than local differential privacy without encryption against an adversary who knows the complete communication network (in honest but curious setting)?
  - Yes: start with lots of noise, and then cancel the noise until (central) DP levels.
- How much noise is needed?
  - Specify privacy requirements
  - Solve constraint program (SDP)

– Sabater, Ben Mokthar & Ramon: ongoing work

Privacy-preserving negotiators

- Idea: negotiating without disclosing more than needed
- Applications:
  - matching websites (ride sharing, crowdsourcing . . .)
  - logistics (e.g., Nalian)
- Postdoc Carlos Cotrini, 6 months
- Can be modeled as bandit problem

– Cotrini, Fromont, Gaudel, Ramon. Ongoing work

Interpretable privacy metrics in medical applications

- HE TRUMPET project: Federated learning between hospitals
- WP3: develop new, interpretable privacy metric integrating both **statistical privacy** and **regulatory** (e.g., GDPR) requirements.
- Medical use cases
- Step 1: combine
  - Predictive performance in crticial (medical) applications
  - Privacy (of personal patient data)

– Taibi & Ramon, ongoing work

Interpretable privacy metrics in medical applications

- Classic differential privacy assumes very strong adversary (who knows all but one patients)
  - Inconsistent: Federated learning uses multi-party computation (MPC), which often only assumes a honest majority
  - Overkill: Hospitals are reasonably well controlled entities, which can be kept liable.
- Idea:
  - Better: Honest fraction privacy (similar to MPC security assumptions)
  - Interpretable: We understand better what we are really protecting

Interpretable privacy metrics in medical applications

- Make $f(x)$ private.
- DP Gaussian mechanism: $\sigma^2 \geq 2\log(1.25/\delta)(\Delta f)^2/\epsilon^2$
  - Noise: $\mathcal{N}\left(0, \sigma^2\right)$.
- Minimal honest fraction:
  $\sigma^2 + \psi\sigma^2_{pop} \geq 2\log(1.25/\delta)(\Delta f)^2/\epsilon^2$
  - Assume at least ψ instances in data sets of honest parties
  - $\sigma^2_{pop} = \mathrm{var}_x(f(x))$ is the population variance
- Further reduction to σ to protect only against attribute attack rather than membership attack.

Interpretable privacy metrics in medical applications

- Next steps:
  - Further extend statistical privacy metric
  - add regulatory dimension
    - with TimeLex partner
    - GDPR, AI act ...
    - Can we integrate legal concepts legal concepts (e.g., 'minimization') with technical concepts (e.g., quantity of and risk for information) ?
    - Can we make a partial order of 'appropriateness' of methods?

Verification of decentralized algorithms

- Position 3: Legal transparency and verifiability (MAGNET, TAU): *"understandable, symbolic explanation is needed which convincingly shows why an algorithm is fair or privacy-friendly."*
- A. Korneev (12/2022–11/2025) : verification of decentralized algorithms (ULille, HE)
- Verify correctness of decentralized algorithm on private data:
  - zero knowledge proofs to verify computation
  - randomized strategies to ensure all proofs are verified, verifiers are trusted, everybody can access a summary of the verification

Conclusions

- Hyaiai helped Magnet to form a network on interpretable AI and get valuable ideas.
- The start needed time, due to covid, the novel direction of the work, the search for qualified researchers . . .
- We obtained interesting results and got funding for further work in this direction.