Designing formally private mechanisms for the p% rule

James Bailie

Australian Bureau of Statistics – Methodology Division
james.bailie@abs.gov.au

University of Wollongong – 5 February 2020

The $p\%$ rule

An aggregate statistic X is a disclosure risk if one contributor can determine another to within $p\%$:

$$X - x_1 - x_2 < x_1 \frac{p}{100},$$

where x_1, x_2 are the values of the largest and second-largest contributors (Wolf and Hundepool 2012).

The $p\%$ rule

An aggregate statistic X is a disclosure risk if one contributor can determine another to within $p\%$:

$$X - x_1 - x_2 < x_1 \frac{p}{100},$$

where x_1, x_2 are the values of the largest and second-largest contributors (Wolf and Hundepool 2012).

The $p\%$ rule can provide assessment of disclosure risk and highlight when treatment is required. But it does not provide that treatment.

Designing a $p\%$ mechanism

- We want to develop some method that can provide $p\%$ protection.
- How can we add noise to ensure that 'an attacker can't determine a contributor's value to within $p\%$'?
- We can use Pufferfish – a customisable, flexible formal privacy framework – to encode this idea.

Pufferfish (Kifer and Machanavajjhala 2014)

A Pufferfish *instantiation* is a tuple $(\mathbb{S}, \mathbb{S}_{\text{pairs}}, \mathbb{D}, \epsilon)$ where

1. $\mathbb{S}$ are the (potential) secrets.
2. $\mathbb{S}_{\text{pairs}} \subset \mathbb{S} \times \mathbb{S}$ are the discriminative pairs.
3. $\mathbb{D}$ are the data evolution scenarios. Each $\theta \in \mathbb{D}$ is a probability distribution over all possible datasets.
4. $\epsilon > 0$ is the acceptable level of privacy leakage.

Pufferfish (Kifer and Machanavajjhala 2014)

A mechanism $\mathcal{M}$ satisfies $(\mathbb{S}, \mathbb{S}_{\text{pairs}}, \mathbb{D}, \epsilon)$-Pufferfish if

1. for all pairs $(s_1, s_2) \in \mathbb{S}_{\text{pairs}}$,
2. for all data evolution scenarios $\theta \in \mathbb{D}$ (with $P(s_i|\theta) \neq 0$),
3. for all outputs ω (with $P(\mathcal{M}(D) = \omega|\theta) \neq 0$),

the prior-to-posterior odds ratio is bounded:

$$e^{-\epsilon} \leq \frac{P(s_1|\mathcal{M}(D) = \omega, \theta)}{P(s_2|\mathcal{M}(D) = \omega, \theta)} \Big/ \frac{P(s_1|\theta)}{P(s_2|\theta)} \leq e^{\epsilon}.$$

(Probability is over the randomness of $\mathcal{M}$ and D.)

Australian Bureau of Statistics

Requirements:

- Passive confidentiality
- Protect a certain sensitive variable x^S for a given record
- Produce sanitised microdata from which aggregates can be safely published
- Linear relationships between variables. (Let $\mathcal{R}_i$ be the set of variables linearly related to the sensitive variable x^S of passive claimant i.)

Australian Bureau of Statistics

A Pufferfish mechanism $\mathcal{M}$:

1. Secrets $\mathbb{S}$ are the statements: "The passive claimant's sensitive variable x^S is in the interval $[(1-p)x, (1+p)x)$" for all $x \in \mathbb{R}$.
2. Discriminative pairs $(s_1, s_2) \in \mathbb{S}_{\text{pairs}}$ are secrets on the neighbouring intervals:
   - s_1 is the statement "$x^S \in [(1-p)x, (1+p)x)$",
   - s_2 is the statement "$x^S \in [(1+p)x, \frac{(1+p)^2}{1-p}x)$".
3. $\theta \in \mathbb{D}$ if θ encodes the linear relationships between the sensitive variable x^S and other variables. That is: for all related variables $x^r \in \mathcal{R}_i$, there is a constant α^r such that $P(x^S = \alpha^r x^r | \theta) = 1$.
4. The mechanism $\mathcal{M}$ multiplies each variable in $\mathcal{R}_i$ by e^ν where $\nu \sim \text{Laplace}(0, b)$ with $b = \frac{-4}{\epsilon} \ln(1-p)$. ($\nu$ is sampled once per passive claimant.)

Proof sketch

1. Protecting the interval $[(1-p)x, (1+p)x)$ is the same as protecting the interval $[\ln x + \ln(1-p), \ln x + \ln(1+p))$ in the logarithm.
2. Since $1 + p < (1-p)^{-1}$, we can protect $[\ln x + \ln(1-p), \ln x + \ln(1+p))$ by protecting $[\ln x + \ln(1-p), \ln x - \ln(1-p))$.
3. So we have reduced the problem to ensuring indistinguishability of neighbouring intervals $[x-c, x+c)$ and $[x+c, x+3c)$.
4. We can do this by adding noise $\mu \sim \mathrm{Laplace}(0, b)$ with $b = \frac{4c}{\epsilon}$ (Lemma 8.1, Kifer and Machanavajjhala 2014).

Where to now?

- Protecting contributors in aggregates (not microdata). Need to encode the attacker's knowledge about another contributor value into $\mathbb{D}$.
- Designing a mechanism for this type of scenario.
  - Adapting the Wasserstein mechanism in (Song et al. 2017)?

Where to now?

- Protecting contributors in aggregates (not microdata). Need to encode the attacker's knowledge about another contributor value into $\mathbb{D}$.
- Designing a mechanism for this type of scenario.
  - Adapting the Wasserstein mechanism in (Song et al. 2017)?

Questions?

References

Kifer, Daniel and Ashwin Machanavajjhala (Jan. 2014). "Pufferfish: A framework for mathematical privacy definitions". In: *ACM Transactions on Database Systems* 39.1, pp. 1–36. ISSN: 03625915. DOI: 10.1145/2514689.

Song, Shuang, Yizhen Wang, and Kamalika Chaudhuri (2017). "Pufferfish Privacy Mechanisms for Correlated Data". In: *Proceedings of the 2017 ACM International Conference on Management of Data - SIGMOD '17*. Chicago, Illinois, USA: ACM Press, pp. 1291–1306. DOI: 10.1145/3035918.3064025.

Wolf, Peter-Paul de and Anco Hundepool (2012). "p% Should Dominate". In: *Privacy in Statistical Databases*. Ed. by David Hutchison et al. Vol. 7556. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–10. DOI: 10.1007/978-3-642-33627-0_1.