Lightweight Privacy-Preserving Task Assignment in Skill-Aware Crowdsourcing

Louis Béziaud[1,3] Tristan Allard[1,2] David Gross-Amblard[1,2]

[1] Univ. Rennes 1, Rennes, France

[2] IRISA, Rennes, France first.last@irisa.fr

[3] ENS Rennes, Rennes, France first.last@ens-rennes.fr

DEXA'17 — August 30, 2017


UMR IRISA
UNIVERSITÉ DE RENNES 1
rennes

Supported by the ANR grant ANR-16-CE23-0004. *CROWDGUARD* crowdguard.irisa.fr

Crowdsourcing

"Crowdsourcing represents the act of a company or institution taking a function once performed by employees and **outsourcing** it to an **undefined** (and generally **large**) network of people in the form of an open call" [How06]

- Amazon Mechanical Turk: crowdsourcing marketplace
- Wikipedia: writing encyclopedia articles
- Galaxy Zoo: classification of galaxies

Knowledge-Intensive Crowdsourcing [Bas+15]

Some tasks **require** particular **skills** to be completed

- testing an application on a specific device
- a translation from English to French
- programming in C++

Privacy in Crowdsourcing?

Knowledge $\Longrightarrow$ Privacy Issues

Skills can be **quasi-identifiers** / **sensitive data** :

- unique combination of skill, location, availability, wage, specific device, ...

The platform is **not a trusted third party** (negligence, illegitimate use, external attack)

Example of Data Breach

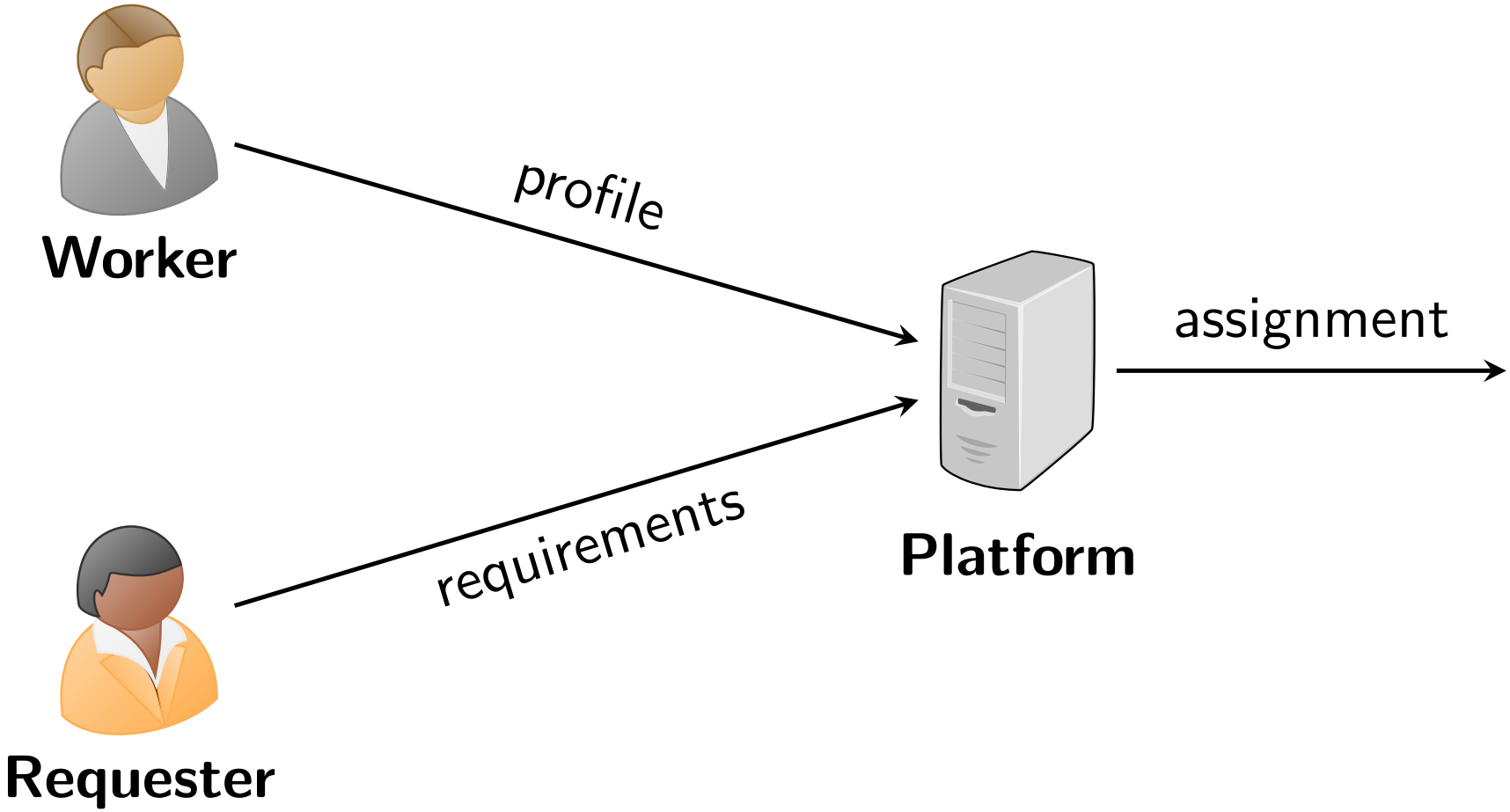
- A worker ID on mTurk gives access to the Amazon profile: real name, wish lists, book reviews, tagged products, ... [Lea+13]
- An Uber's executive illegitimately tracked a journalist's location [BW14]
- Ashley Madison dating service's user base was stolen [Man15]

Related Work

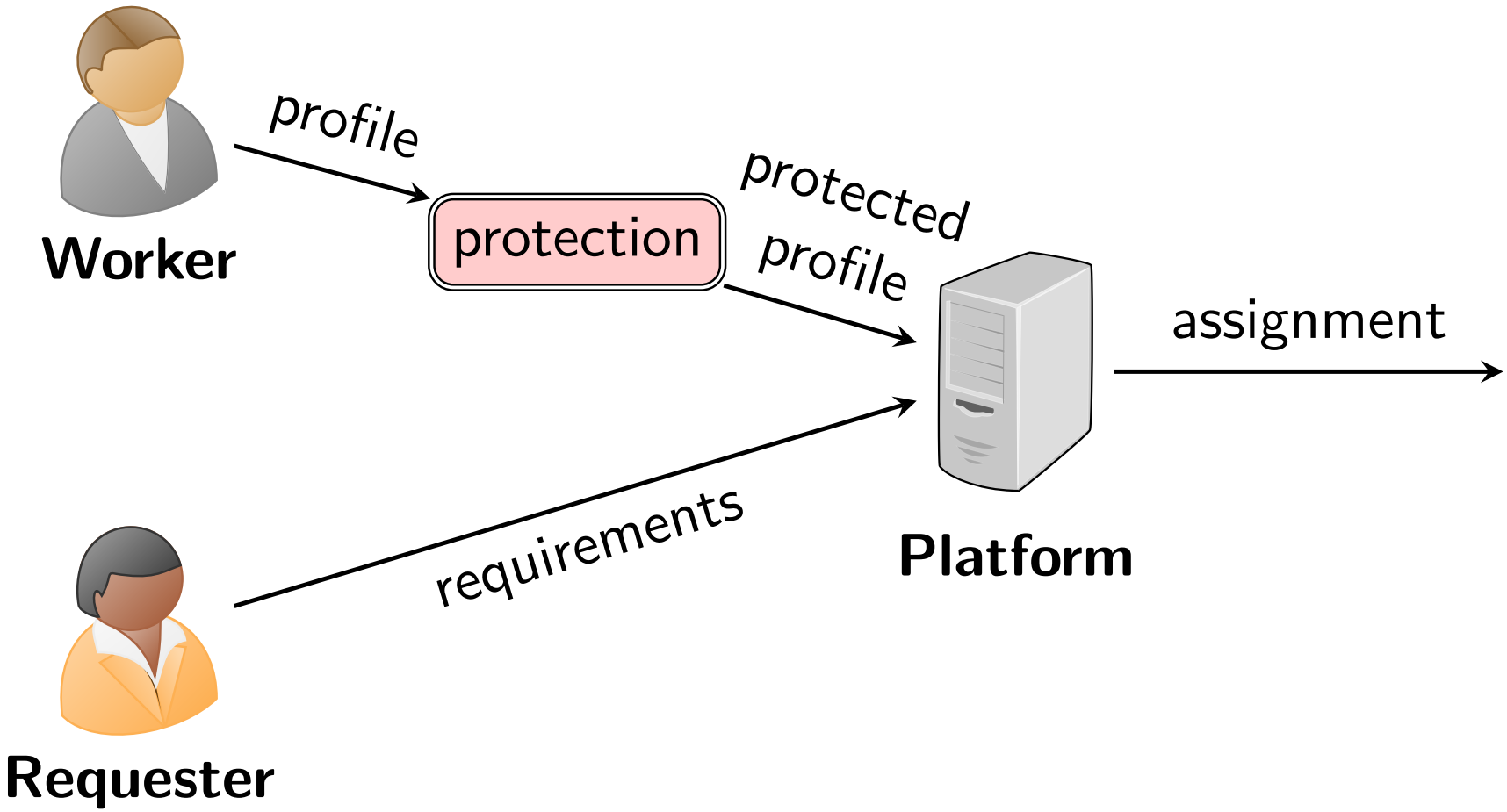
- Optimal task assignment using a **decentralized** maximum flow algorithm inside a Paillier **homomorphic cryptosystem** [Kaj16]
  $\Rightarrow$ need **more than a century** to assign 100 workers and tasks, according to the author
- Privacy-preserving crowdsourced surveys [Kan+14] or spatial crowdsourcing [TGS14]
  $\Rightarrow$ **no skills**, **no assignment**

How can we do **privacy-preserving** task assignment with **lightweight techniques** (i.e. no encryption, centralized)?

Approach


Worker
profile
assignment
requirements
Platform
Requester

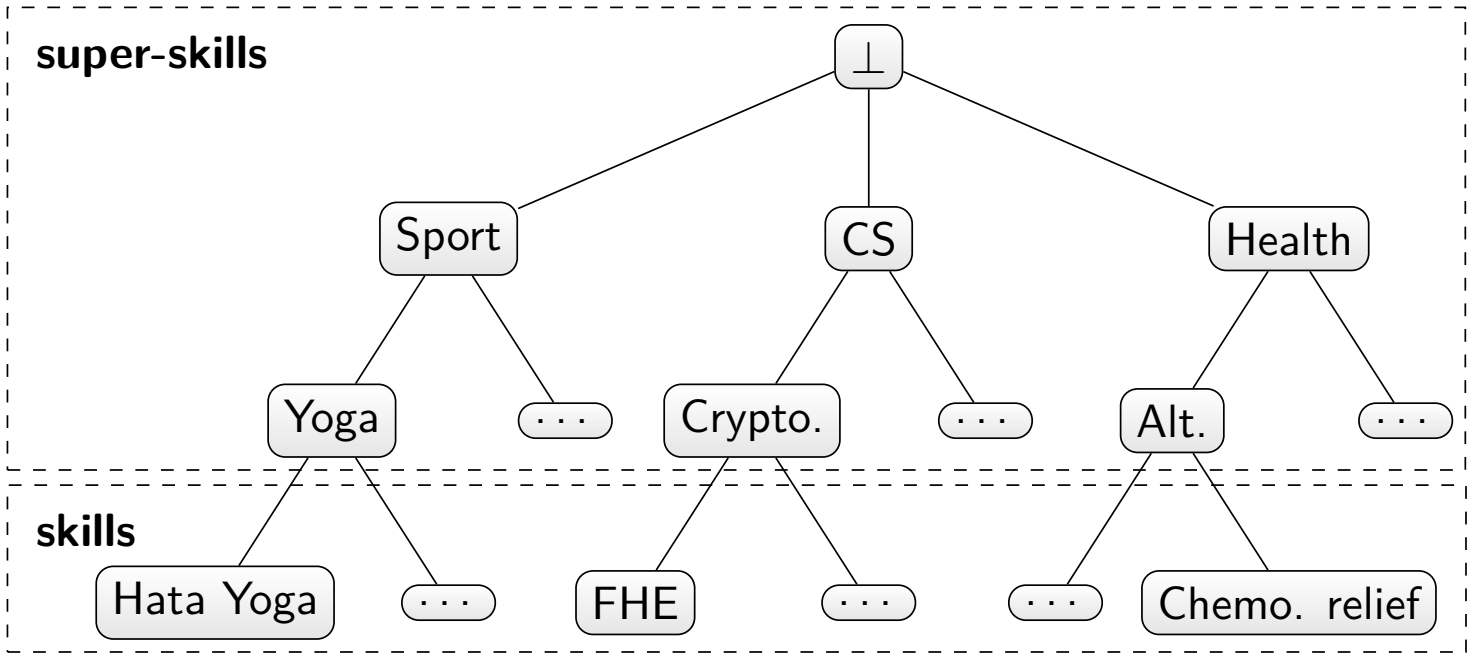
Approach


profile
protection
protected
profile
assignment
Worker
Platform
requirements
Requester

Skills Profile

Skills Taxonomy

Skills are organized in a tree-structure with a *is-a* relationship.



Profile

- **skill** $s_i \in Boolean$
- **super-skill** $s_i =$ proportion of descendant skills possessed

$\Rightarrow$ **profile** $=$ array of **skills**

Differential Privacy [Dwo06]

Idea

The **outcome of any analysis is essentially equally likely**, independent of whether an individual joins, or refrains from joining, the data set

Differential Privacy [Dwo06]

M gives ϵ-differential privacy if for all pairs of data-sets x, y differing in one element, and all subsets S of possible outputs,

$$\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(y) \in S]$$

Properties

- Sequential & parallel **composition** [McS09] $\implies$ budget ϵ sharing
- Post-processing [DR14] $\implies$ can re-use data and add information

Task Assignment Problem

Assignment problem

- **workers** $\mathcal{P} = \{p_0, p_1, ..., p_n\}$
- **tasks** $\mathcal{T} = \{t_0, t_1, ..., t_n\}$
- **cost function** $C : \mathcal{P} \times \mathcal{T} \to \mathbb{R}$

Find a bijection $f : \mathcal{P} \to \mathcal{T}$ such that $\sum_{p \in \mathcal{P}} C(p, f(p))$ is minimized.
w.l.o.g. assumes $|\mathcal{P}| = |\mathcal{T}|$

simplex algorithm, Hungarian algorithm, minimum-cost flow, ...

Assignment Quality

Assignment's total cost depends on the weight-function

Relative Quality

$$\mathtt{q_{rel}} = \frac{\sum_{(t,p)\in\mathcal{A}} \mathtt{C}(t,p)}{\sum_{(t,\tilde{p})\in\tilde{\mathcal{A}}} \mathtt{C}(t,\tilde{p})}$$

Fraction of Perfect Assignments

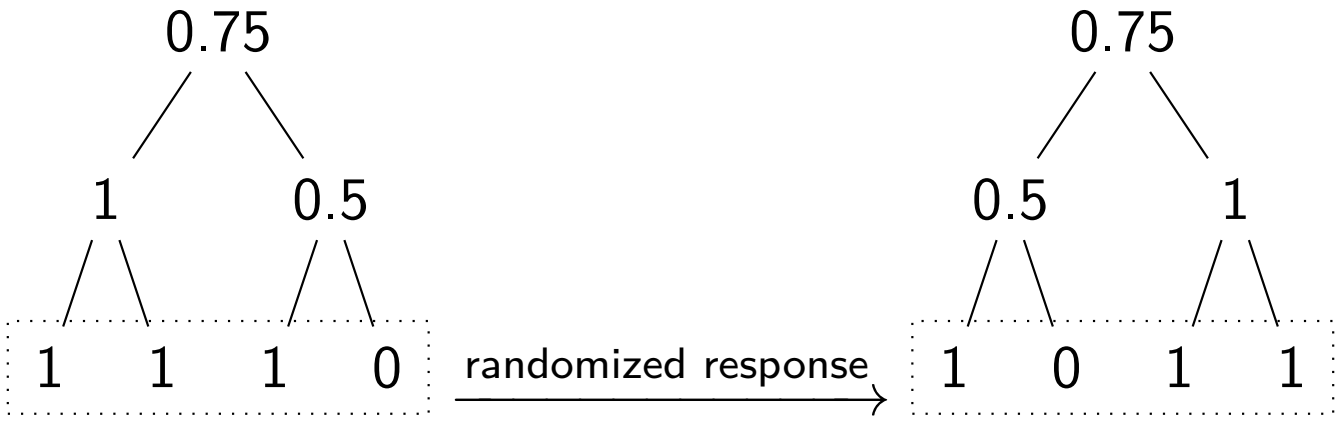
$$\mathtt{fpa} = \frac{\left|\left\{\tilde{p} \geq t \mid (t,\tilde{p}) \in \tilde{\mathcal{A}}\right\}\right|}{|\tilde{\mathcal{A}}|}$$

Randomized Response [War65]

Proposed by S. L. Warner in 1965 for **survey interviews**. It allows respondents to respond to sensitive issues while maintaining confidentiality.

$$\widetilde{x} = \begin{cases} x & \text{with probability} \quad 1 - \Pr_{flip} \\ 1 & \qquad\qquad\quad 0.5 \times \Pr_{flip} \\ 0 & \qquad\qquad\quad 0.5 \times \Pr_{flip} \end{cases}$$

Th. randomized response satisfies ϵ-differential-privacy if $\Pr_{flip} = \frac{2}{1+e^{\epsilon}}$



Existing Weight Functions

Missing Weight Function ($\sim$ Hamming)

number of skills required by the task and missing in the worker's profile

$$\texttt{MWF}(t, \tilde{p}) = \sum_i t[i] \wedge \neg \tilde{p}[i]$$

Ancestors Weight Function (from [MGM16])

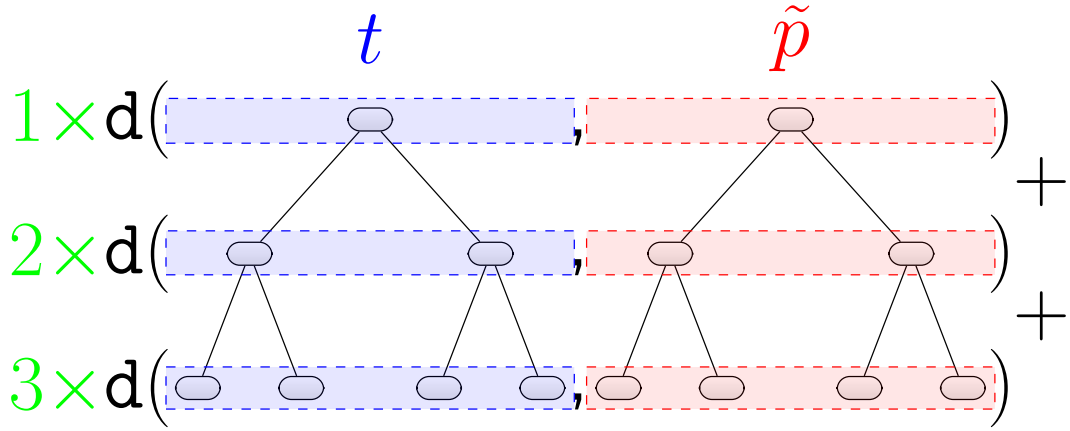
distance between the depth of each worker's skill and its closest required skill

$$\texttt{AWF}(t, \tilde{p}) = \sum_{s_i \in \tilde{p}} \min_{s_j \in t} (\text{depth}(\text{lca}(s_i, s_j)))$$

d_{max} = depth of the taxonomy

lca = lowest common ancestor

Climbing Weight Function ↕



Idea 1 leveraging the *is-a* relationship by computing a distance for each level

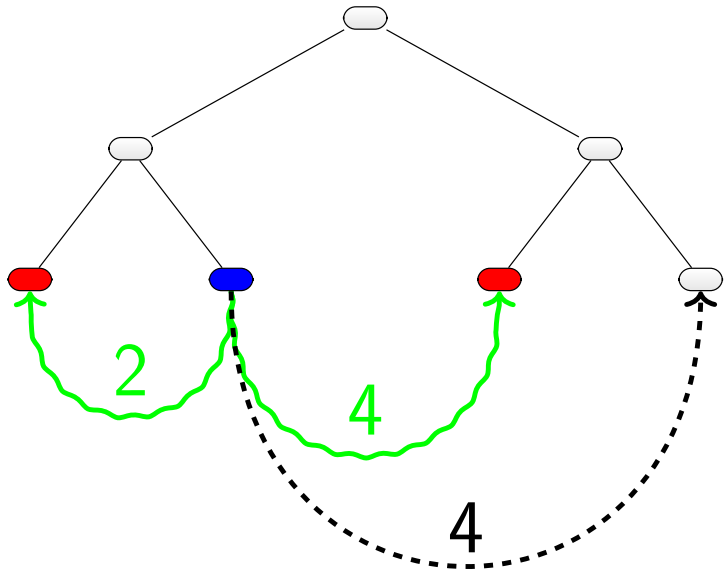
top = more precision, less relevance

bottom = less precision, more relevance

$$\mathtt{CWF}(t,\tilde{p}) = \sum_{\text{level } i} i\times\mathtt{d}(u_i, v_i)$$

$\mathtt{d}$ = **any weight function**

Touring Weight Function $\leftrightarrow$



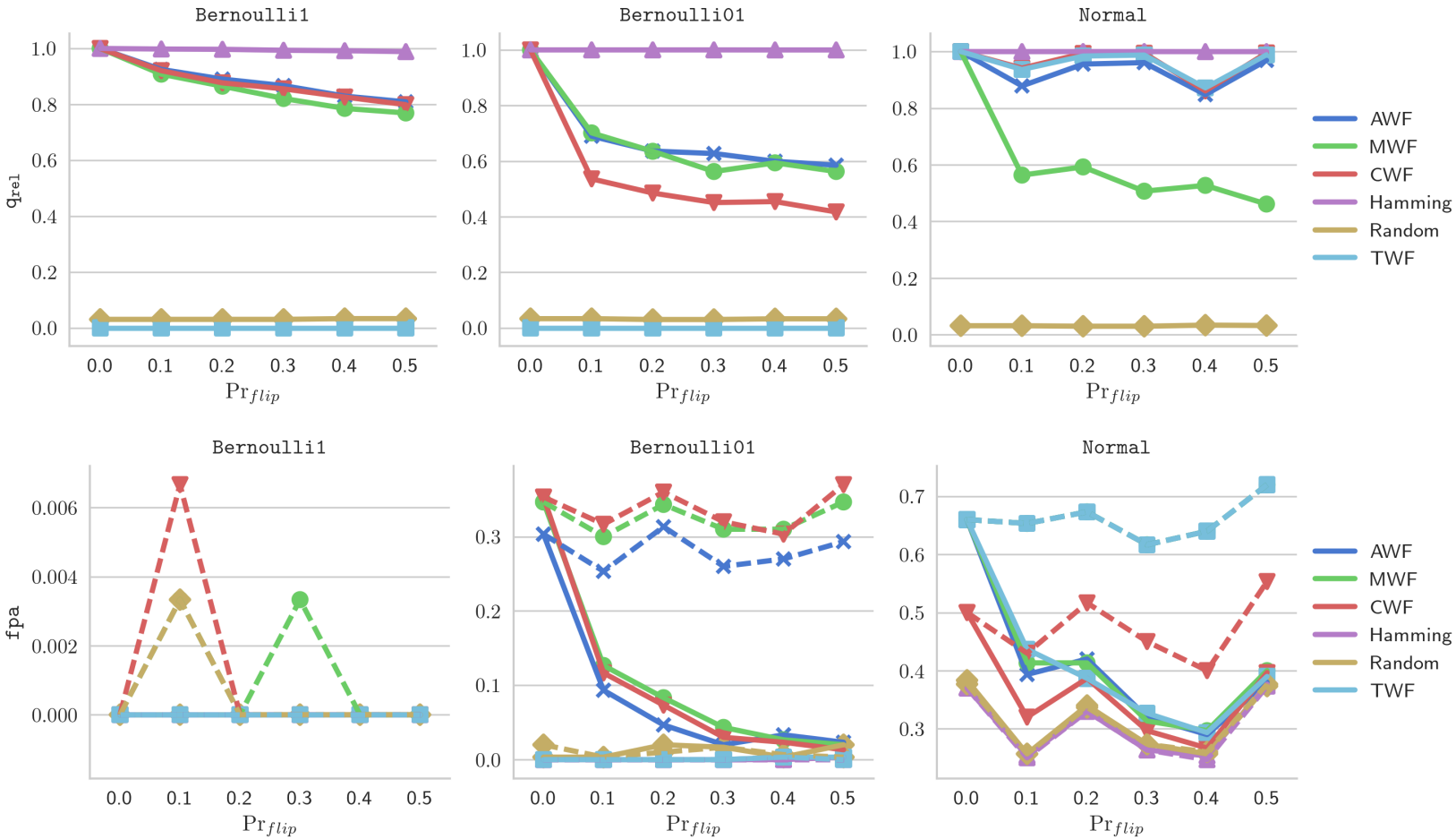
Idea 2 leveraging the "neighborhood" property: two **close skills are more likely to have the same value** than two distant skills

$$\texttt{TWF}(t, \tilde{p}) = \sum_{s_i \in t} \sum_{s_j \in \tilde{p}} (s_i \rightsquigarrow s_j)$$

$\rightsquigarrow$ = distance in the taxonomy tree

Results (subset)

- taxonomy: perfect tree, height = 3, branching = 4, 121 nodes, 81 skills
- `Bernoulli1`: user has skill s_i with probability 0.1 (10%)
- `Bernoulli01`: user has skill s_i with probability 0.01 (1%)
- `Normal`: skills are drawn from a normal distribution (around 30%)
- 100 workers, 100 tasks



Conclusion

- **privacy-preserving** approach to the problem of assigning tasks to workers with new **weight functions**
- **lightweight**
- **client-side**
- **pluggable** into existing platforms

Future Work

- large-scale real-life skill dataset
- complete skills taxonomy (*e.g.* Skill-Project `skill-project.org`)
- performance vs quality trade-off $\leftrightarrow$ mix perturbation + encryption
- does client-side differential privacy make sens?

Conclusion

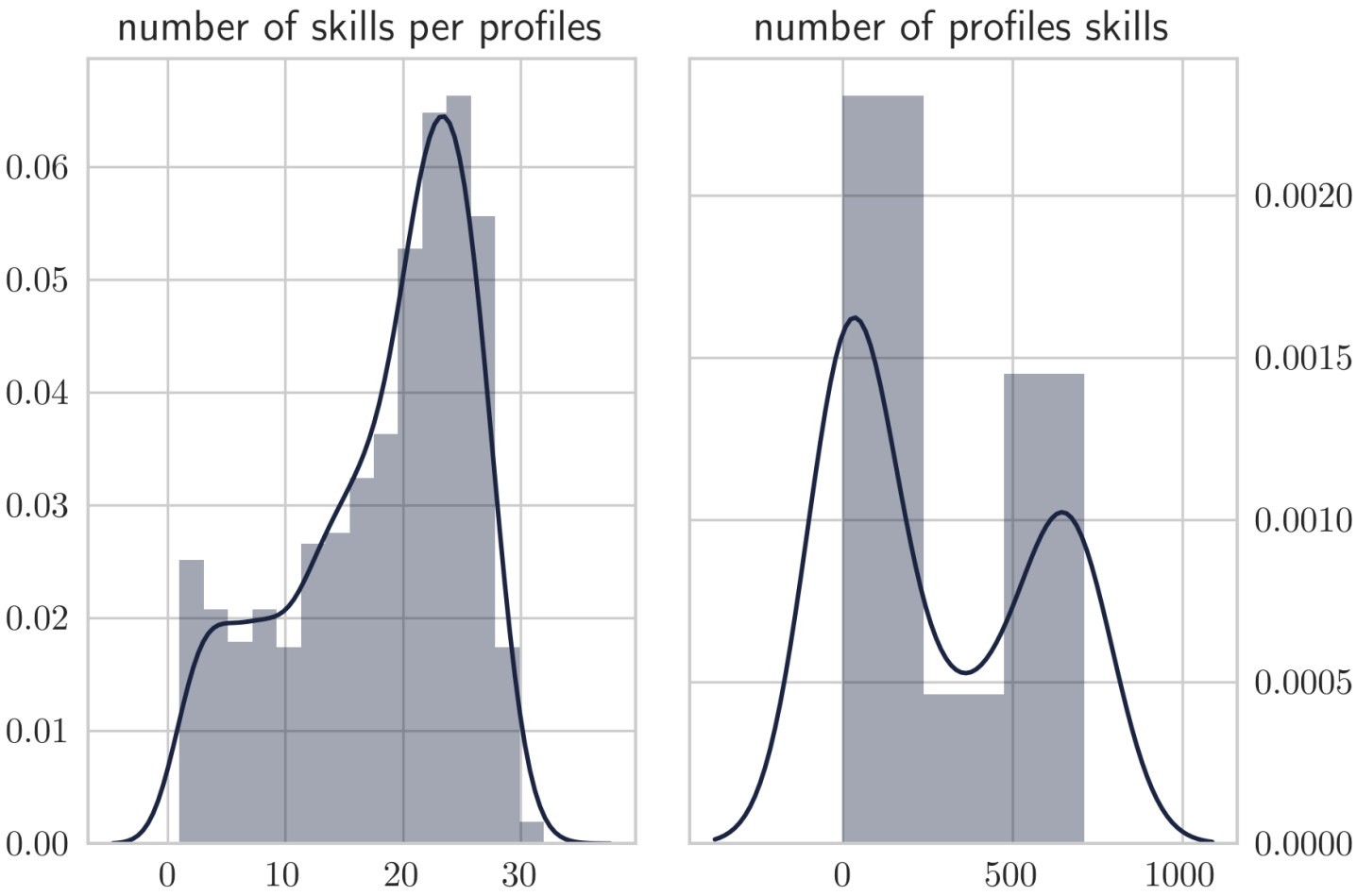
- **privacy-preserving** approach to the problem of assigning tasks to workers with new **weight functions**
- **lightweight**
- **client-side**
- **pluggable** into existing platforms

Future Work

- large-scale real-life skill dataset
- complete skills taxonomy (*e.g.* Skill-Project `skill-project.org`)
- performance vs quality trade-off $\leftrightarrow$ mix perturbation + encryption
- does client-side differential privacy make sens?

Thank you! Questions? `louis.beziaud@ens-rennes.fr`

Backup: `Normal` skills profile



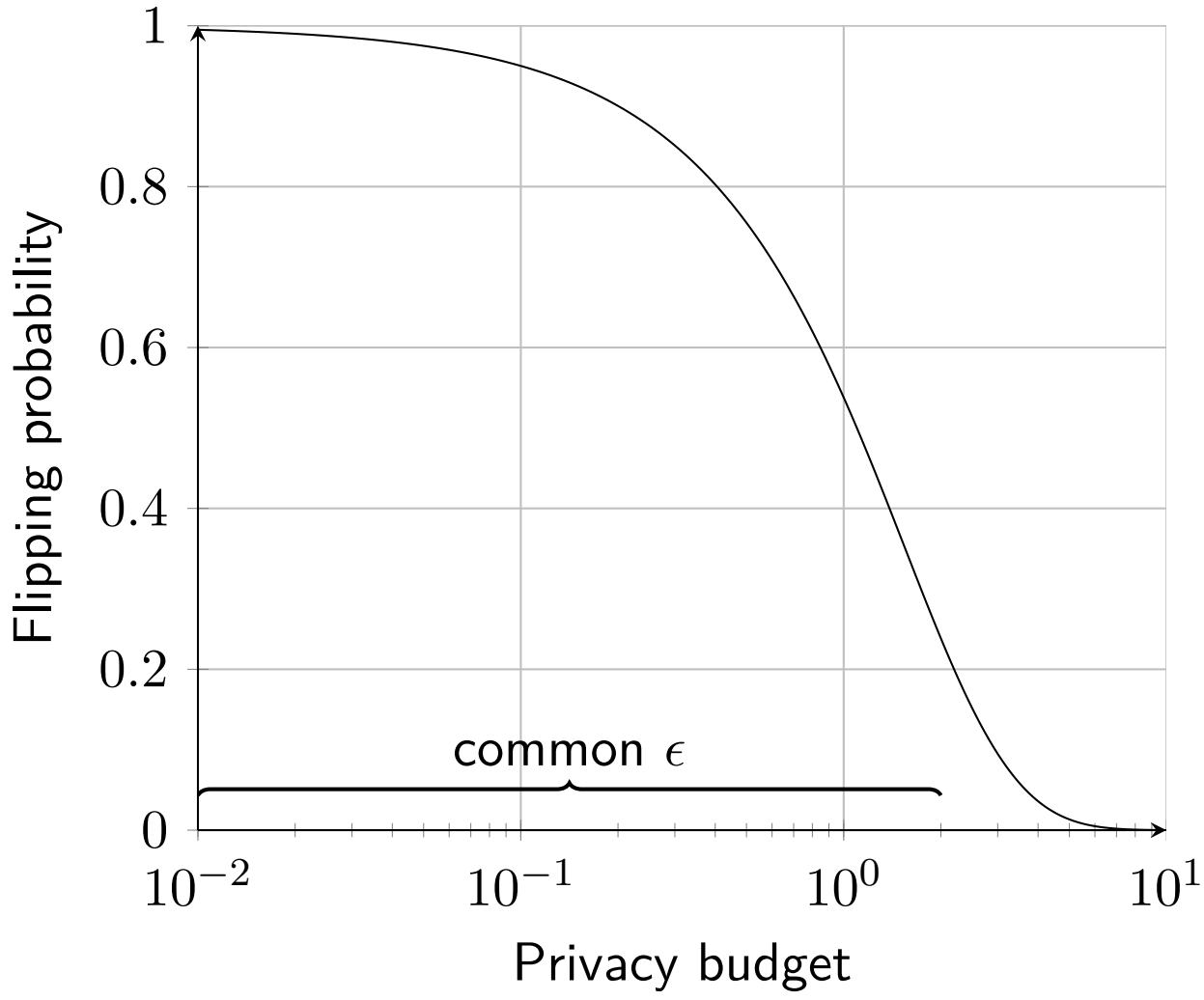
draw skills from $\mathcal{N}(\sqrt{r}\cos\theta, \sqrt{r}|\sin\theta|)$ with $\theta \in [0, 2\pi)$ for each profile, and $r = 0.06$. around 30% of skills

Backup: Time

Weight function	Time (s)	Time complexity
`rand`	0.00099	$\mathcal{O}(1)$
`hamming`	0.02467	$\mathcal{O}(\lvert\mathcal{S}\rvert)$
`MWF`	0.01969	$\mathcal{O}(\lvert\mathcal{S}\rvert)$
`CWF`	0.30395	$\mathcal{O}(\lvert\mathcal{S}^T\rvert)$
`AWF`	1.99307	$\mathcal{O}(\lvert\mathcal{S}\rvert^2)$
`TWF`	2.96748	$\mathcal{O}(\lvert\mathcal{S}\rvert^2)$

Skill-project, 2208 nodes, 1562 skills

Backup: $\Pr_{flip}$ versus ϵ



References I

[Bas+15] Senjuti Basu Roy et al. "Task assignment optimization in knowledge-intensive crowdsourcing". In: *The VLDB Journal* 24.4 (2015), pp. 467–491. ISSN: 0949-877X. DOI: 10.1007/s00778-015-0385-2. URL: http://dx.doi.org/10.1007/s00778-015-0385-2.

[BW14] Johana Bhuiyan and Charlie Warzel. *God View": Uber Investigates Its Top New York Executive for Privacy Violations*. Nov. 2014. URL: https://www.buzzfeed.com/johanabhuiyan/uber-is-investigating-its-top-new-york-executive-for-privacy?utm_term=.em5Zl4KkG#.pqqYP9OEG.

[DR14] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042.

[Dwo06] Cynthia Dwork. "Differential Privacy". In: *Proc. of ICALP '06*. 2006, pp. 1–12.

[How06] Jeff Howe. "The rise of crowdsourcing". In: *Wired magazine* 14.6 (2006), pp. 1–4.

[Kaj16] Hiroshi Kajino. "Privacy-Preserving Crowdsourcing". PhD thesis. Univ. Tokyo, 2016. URL: https://sites.google.com/site/hiroshikajino1989/home/publications.

[Kan+14] Thivya Kandappu et al. "Loki: A privacy-conscious platform for crowdsourced surveys". In: *COMSNETS '14*. 2014, pp. 1–8.

[Lea+13] Matthew Lease et al. "Mechanical Turk is Not Anonymous". In: *SSRN Electronic Journal* (2013). DOI: 10.2139/ssrn.2228728.

[Man15] Steve Mansfield-Devine. "The Ashley Madison affair". In: *Network Security* 2015.9 (2015), pp. 8–16.

[McS09] Frank McSherry. "Privacy integrated queries: an extensible platform for privacy-preserving data analysis". In: *Proc. of ACM SIGMOD '09*. 2009, pp. 19–30.

[MGM16] Panagiotis Mavridis, David Gross-Amblard, and Zoltán Miklós. "Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing". In: *Proc. of WWW '16*. 2016, pp. 843–853.

[TGS14] Hien To, Gabriel Ghinita, and Cyrus Shahabi. "A framework for protecting worker location privacy in spatial crowdsourcing". In: *Proc. VLDB Endow.* 7.10 (June 2014), pp. 919–930. DOI: 10.14778/2732951.2732966.

[War65] Stanley L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *Journal of the American Stat. Assoc.* 60.309 (1965), pp. 63–69. DOI: 10.1080/01621459.1965.10480775.