EPFL

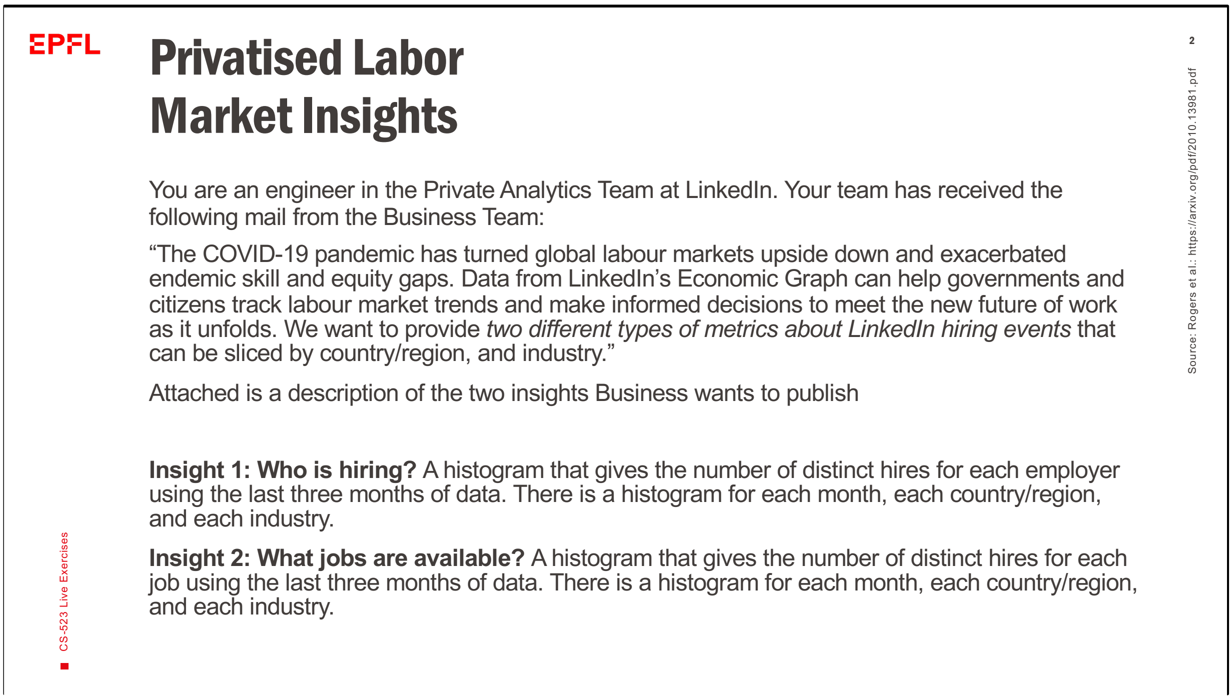
SPRIN
SECURITY AND PRIVACY ENGINEERING LABOR

CS-523 Advanced topics on Privacy Enhancing Technologies

Privacy-preserving Data Publishing (Part II)

Interactive exercises

Mathilde Raynal
SPRING Lab
mathilde.raynal@epfl.ch



Privatised Labor Market Insights

You are an engineer in the Private Analytics Team at LinkedIn. Your team has received the following mail from the Business Team:

“The COVID-19 pandemic has turned global labour markets upside down and exacerbated endemic skill and equity gaps. Data from LinkedIn’s Economic Graph can help governments and citizens track labour market trends and make informed decisions to meet the new future of work as it unfolds. We want to provide *two different types of metrics about LinkedIn hiring events* that can be sliced by country/region, and industry.”

Attached is a description of the two insights Business wants to publish

Insight 1: Who is hiring? A histogram that gives the number of distinct hires for each employer using the last three months of data. There is a histogram for each month, each country/region, and each industry.

Insight 2: What jobs are available? A histogram that gives the number of distinct hires for each job using the last three months of data. There is a histogram for each month, each country/region, and each industry.

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

Privatised Labor Market Insights

The email continues:

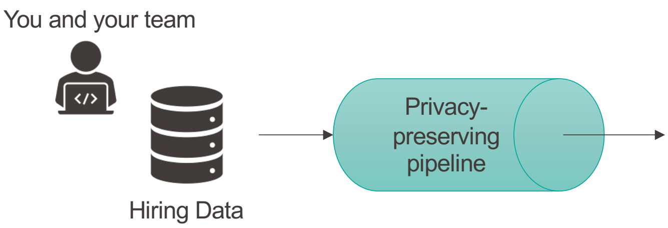
“Although we find that this data can be incredibly helpful to governments, policy makers, and individuals in the global workforce during these challenging times, we want to ensure member trust is preserved and *no individual can be identified* based on the reports we provide.”

In this exercise, we will build a private data pipeline to publish the two insights.

We will look at 4 things (so 4 steps): data, privacy, implementation, and utility.

LinkedIn

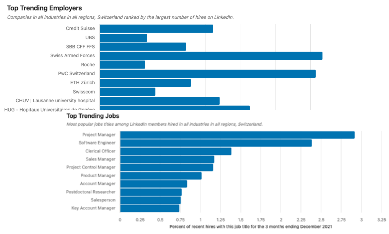
Public



You and your team

Hiring Data

Privacy-preserving pipeline



Top Trending Employers

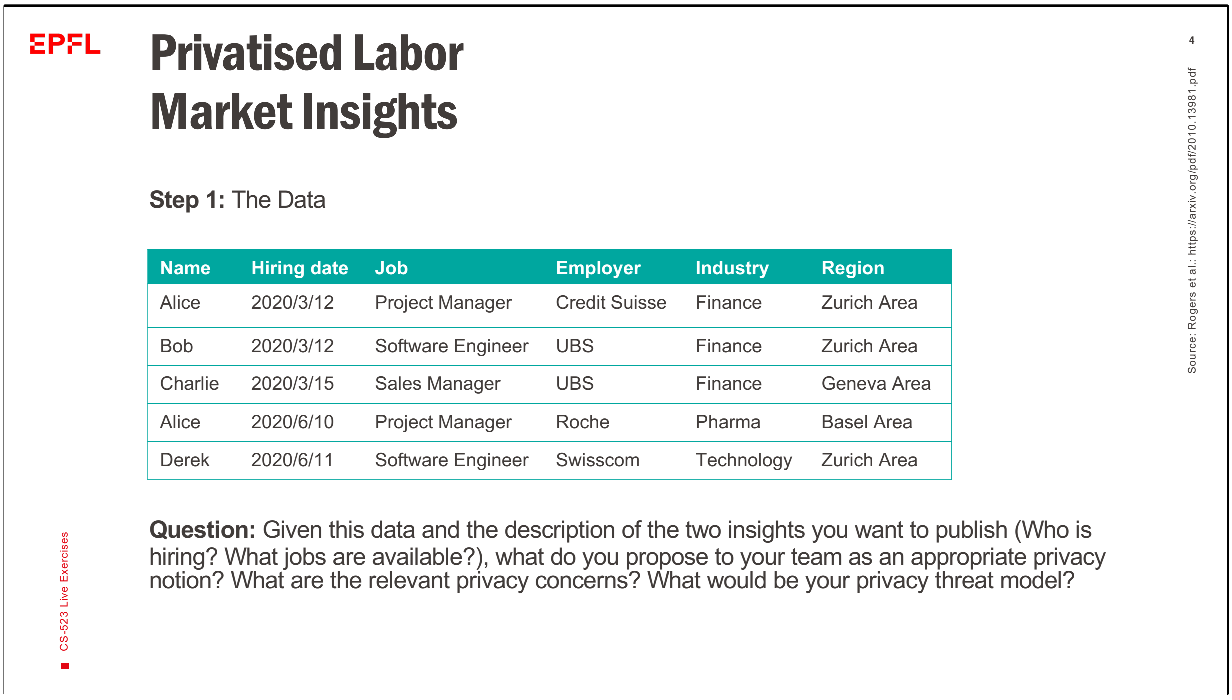
Companies in all industries in all regions, Switzerland ranked by the largest number of hires on LinkedIn.

Top Trending Jobs

Most popular jobs titles among LinkedIn members hired in all industries in all regions, Switzerland.

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

CS-523 Live Exercises



EPFL

Privatised Labor Market Insights

Step 1: The Data

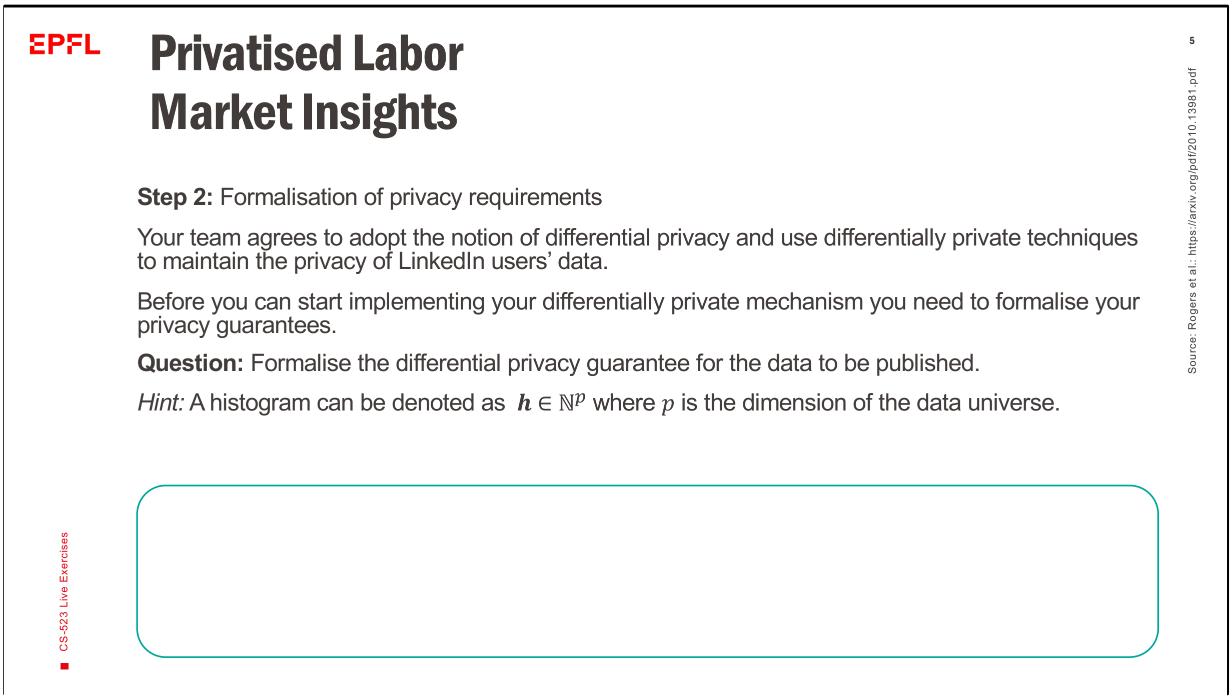
Name	Hiring date	Job	Employer	Industry	Region
Alice	2020/3/12	Project Manager	Credit Suisse	Finance	Zurich Area
Bob	2020/3/12	Software Engineer	UBS	Finance	Zurich Area
Charlie	2020/3/15	Sales Manager	UBS	Finance	Geneva Area
Alice	2020/6/10	Project Manager	Roche	Pharma	Basel Area
Derek	2020/6/11	Software Engineer	Swisscom	Technology	Zurich Area

Question: Given this data and the description of the two insights you want to publish (Who is hiring? What jobs are available?), what do you propose to your team as an appropriate privacy notion? What are the relevant privacy concerns? What would be your privacy threat model?

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

We want to publish aggregate statistics. As the statistics might be quite fine-grained (slice by industry and region) and updated over time, the formal notion of differential privacy seems a good fit.

Our main privacy concerns should be that somebody learns about a hiring event



Privatised Labor Market Insights

Step 2: Formalisation of privacy requirements

Your team agrees to adopt the notion of differential privacy and use differentially private techniques to maintain the privacy of LinkedIn users' data.

Before you can start implementing your differentially private mechanism you need to formalise your privacy guarantees.

Question: Formalise the differential privacy guarantee for the data to be published.

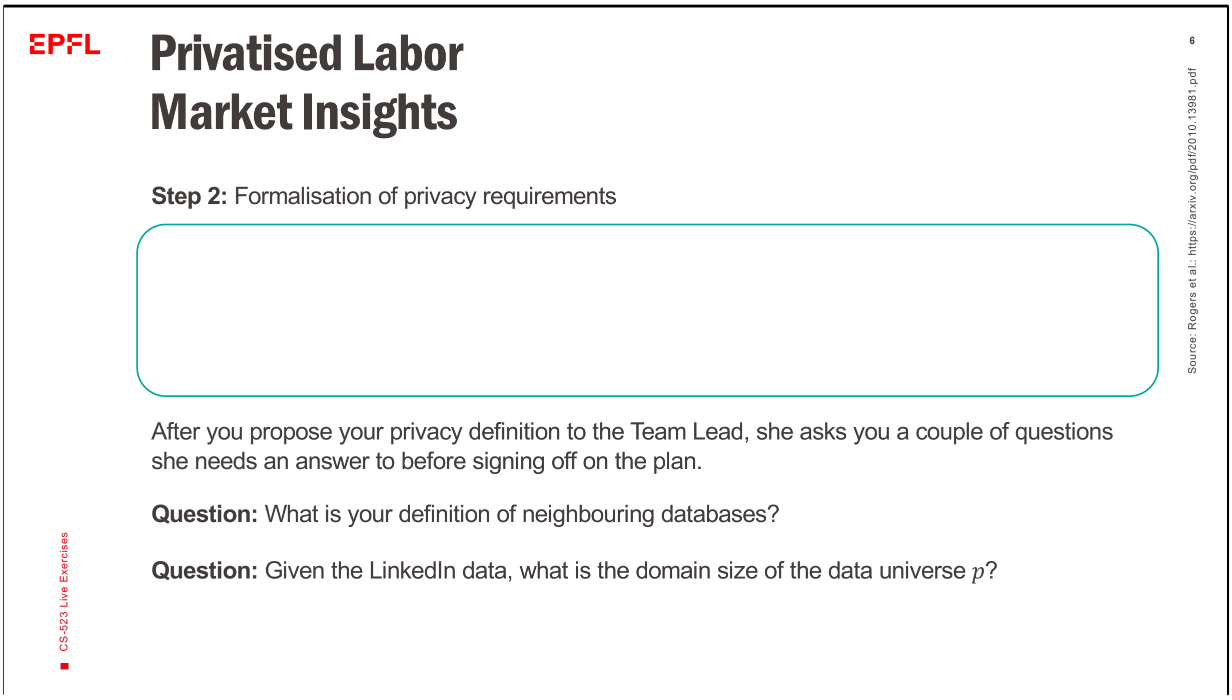
Hint: A histogram can be denoted as $\boldsymbol{h} \in \mathbb{N}^p$ where p is the dimension of the data universe.

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

CS-523 Live Exercises

5

Should be some form of P[M(D) = O] <= e^epsilon P[M(D-r) = O] for all O in output space



Privatised Labor Market Insights

Step 2: Formalisation of privacy requirements

After you propose your privacy definition to the Team Lead, she asks you a couple of questions she needs an answer to before signing off on the plan.

Question: What is your definition of neighbouring databases?

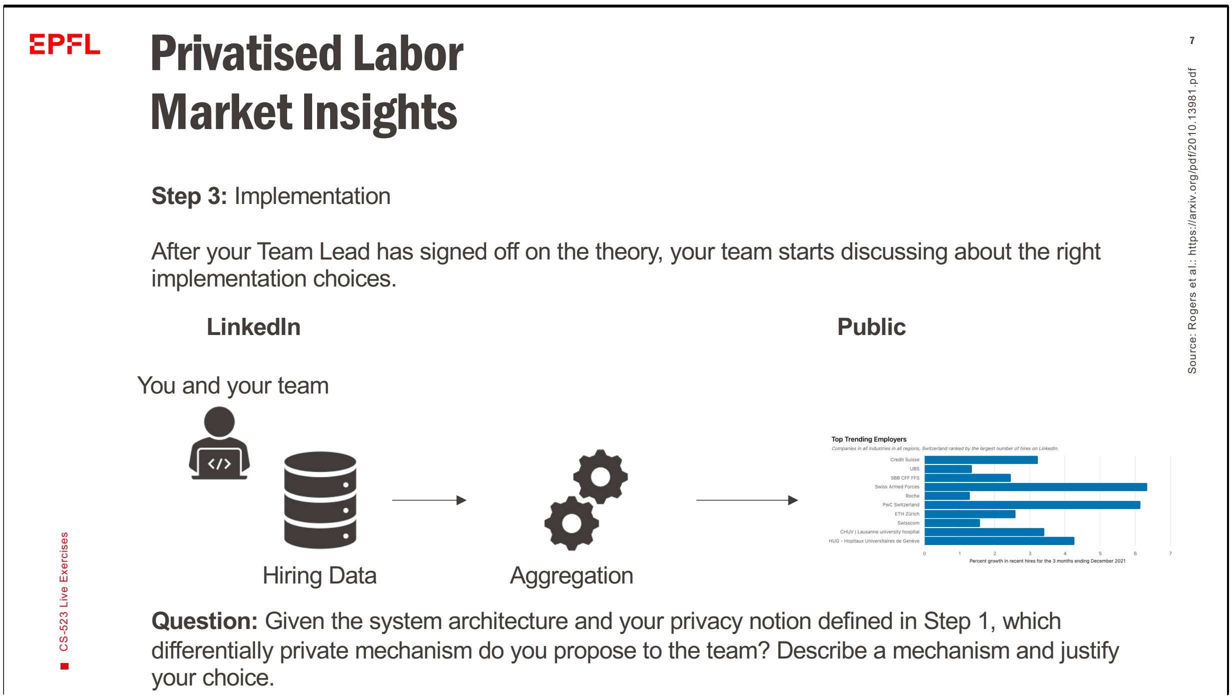
Question: Given the LinkedIn data, what is the domain size of the data universe p?

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

CS-523 Live Exercises

Your definition of neighbouring depends on the chosen unit of privacy.
As we care about user privacy, it would be best to give user-level privacy guarantees. Event-level privacy (each hiring event) however would be easier to work with in practice. With user-level privacy we need to consider the maximum number of hiring events a single user can contribute within any three month period. This is impractical

The domain size p might be known or unknown. If it is a histogram broken down by region only the domain size p is known. It corresponds to the number of cantons in Switzerland, for instance., If it is a histogram over the number of hires per employer the domain size is unknown as there might be new employers.



Privatised Labor Market Insights

Step 3: Implementation

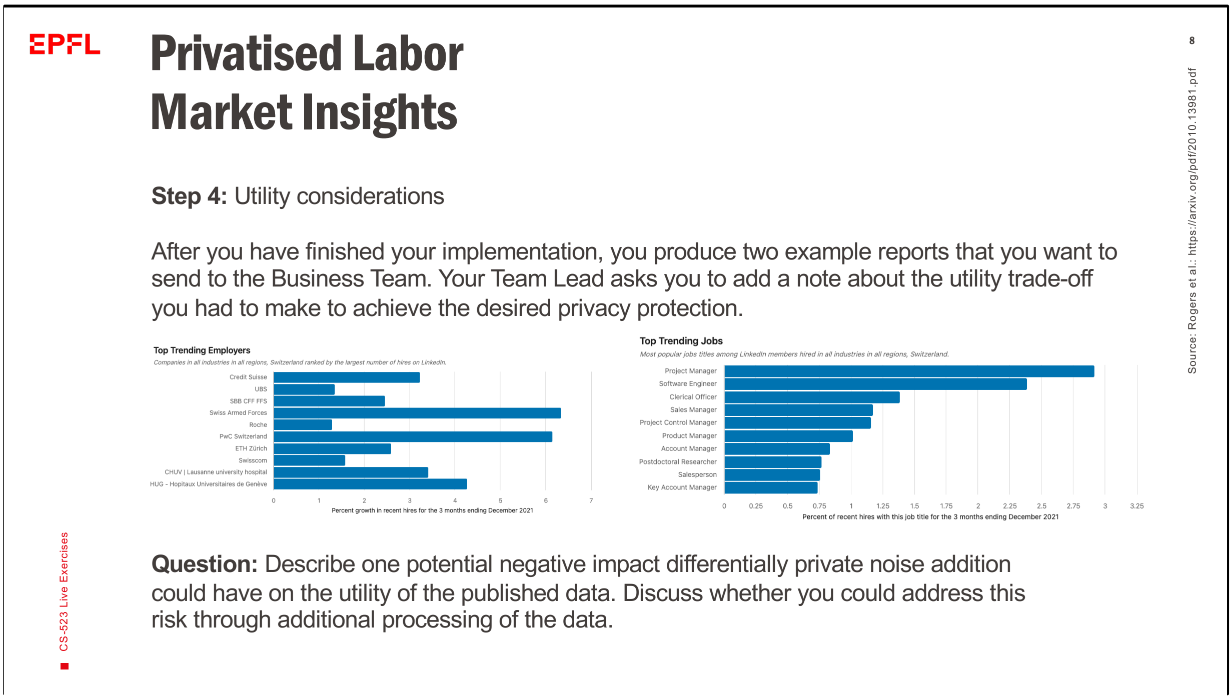
After your Team Lead has signed off on the theory, your team starts discussing about the right implementation choices.

LinkedIn
Public
You and your team
Hiring Data
Aggregation
Top Trending Employers
Companies in all industries in all regions, Switzerland ranked by the largest number of hires on LinkedIn.
Credit Suisse
UBS
SBB CFF FFS
Swiss Armed Forces
Roche
PwC Switzerland
ETH Zürich
Swisscom
CHUV | Lausanne university hospital
HUG - Hopitaux Universitaires de Genève
0
1
2
3
4
5
6
7
Percent growth in recent hires for the 3 months ending December 2021

Question: Given the system architecture and your privacy notion defined in Step 1, which differentially private mechanism do you propose to the team? Describe a mechanism and justify your choice.

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

The central mode of differential privacy (output perturbation) seems most appropriate here.
We could use a simple Laplace mechanism to publish noisy counts.
Details of implementation: How do you determine sensitivity?



EPFL

Privatised Labor Market Insights

Step 4: Utility considerations

After you have finished your implementation, you produce two example reports that you want to send to the Business Team. Your Team Lead asks you to add a note about the utility trade-off you had to make to achieve the desired privacy protection.

Top Trending Employers

Companies in all industries in all regions, Switzerland ranked by the largest number of hires on LinkedIn.

Credit Suisse
UBS
SBB CFF FFS
Swiss Armed Forces
Roche
PwC Switzerland
ETH Zürich
Swisscom
CHUV | Lausanne university hospital
HUG - Hopitaux Universitaires de Genève
0
1
2
3
4
5
6
7
Percent growth in recent hires for the 3 months ending December 2021

Top Trending Jobs

Most popular jobs titles among LinkedIn members hired in all industries in all regions, Switzerland.

Question: Describe one potential negative impact differentially private noise addition could have on the utility of the published data. Discuss whether you could address this risk through additional processing of the data.

Source: Rogers et al.: https://arxiv.org/pdf/2010.13981.pdf

Risk: There is a risk of false conclusions that you should protect your users from. Solution: Do not publish noisy results for small groups.

Risks: Statistics on which jobs small employers, like start-ups, are hiring for will get perturbed the most. This might negatively affect their growth if LinkedIn users use the statistics to decide what roles to specialise for.

If governments use the data to make policy decisions about distributing funding, smaller, non-urban areas, will be negatively affected.