Privacy Research at SAP

Volkmar Lotz, Francesco di Cerbo, SAP
April 27, 2023
Agenda

Security and privacy research at SAP – an overview

Generating Realistic Synthetic Curricula Vitae for Machine Learning Applications under Differential Privacy
Security and Privacy Research at SAP
About SAP Security Research

37 researchers (incl. 6 PhD students)
+ 24 students (Feb ‘23)

2 Locations: France and Germany

Focus on H3 organic innovation: „invent the future of security“

Scientific background: regular top4 conference papers, PhD program, …

8 collaborative research projects (EU or national funding)

22 years of applied research

- IoT security, Encrypted DB, AI for Threat Intelligence, Privacy, Software and OS security
- Training, Security Consulting, Certification, M&A
Why security research at SAP?

Security risks are evolving and emerging

- Increased risk exposure and impact through extended attack surface, stronger dependencies, complex supply chains and increased asset value / criticality
- Evolving threat landscape, AI empowered attackers, tension in world economy and politics
- Compliance risks, trend to regulate security and privacy
- New technologies, inherent technical complexity, risk of overlooking critical threats (e.g., AI and ML)
- Combinatorial complexity, finding the needle in the haystack (assessing false positives)

Grand challenges in security and privacy are unsolved

- Secure and privacy-friendly data business
- Trustworthy large-scale and distributed systems, including AI
- Elimination of software vulnerabilities
- Self-healing systems and applications
- Usable security and human factors
- Impact of quantum technologies
Applied Research: Bridging Academia with SAP Business

Business-driven
- Our research is driven by the needs of SAP and our customers, and is aligned with SAP’s strategy and business. We apply new principles in security research through business value analysis and by refining and adapting methodologies and technology concepts.

Ready to be adopted
- Our research aims at a level of maturity that makes its results suitable for being transferred into and adopted by SAP’s development and business units. Issues of scale and technology integration are explicitly in scope.
The 5 pillars of the Security Research portfolio

- Privacy & Trust
- Applied Cryptography
- Zero Vulnerabilities
- Cyber Defense
- Future Technology

Data focus

Software focus
The 5 pillars of the Security Research portfolio

Data focus

- Privacy & Trust
- Applied Cryptography
- Zero Vulnerabilities
- Cyber Defense

Software focus

Anonymization
AI for Privacy
Trustworthy AI
Generating Realistic Synthetic Curricula Vitae for Machine Learning Applications under Differential Privacy

Francesco Di Cerbo, Francesco Aldà, Andrea Bruera
Is Data the New Oil?

Most unstructured data contain Personal Data

GDPR compliance risk over data processing
Can we Anonymize Unstructured Text?

<table>
<thead>
<tr>
<th>Anonymization</th>
<th>Anonymization is hard!</th>
<th>Re-identification attacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “making re-identification of an individual impossible”, paraphrasing GDPR</td>
<td>• If we consider arbitrary text (social media, medical records, résumés/CVs...)</td>
<td>• can exploit side-knowledge coming from external sources</td>
</tr>
</tbody>
</table>
ML models for HR?

As a data scientist, I need to develop a ML model that:
• Extracts information out of a submitted CV
• Computes a score for a CV against a job posting
• ...

Where do I get my training dataset of CVs? How can I be compliant with DPP requirements?

Can I anonymize CVs?
Re-Identification Risk Assessment Example: CVs/ Résumés (1/2)

Is this sufficient?
Re-Identification Risk Assessment Example: CVs/ Résumés (2/2)

Is this sufficient?

NO!

Uniqueness of education & work experience

We need another idea!

### Experience

<table>
<thead>
<tr>
<th>Role</th>
<th>Organization</th>
<th>Dates</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher</td>
<td>SAP Labs</td>
<td>Oct 2011 - Present</td>
<td>10 yrs 2 mos</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sophia Antipolis, Nice Area, France</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td>Free University of Bolzano-Bozen</td>
<td>Feb 2007 - Sep 2011</td>
<td>6 yrs 8 mos</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td>Center for Applied Software Engineering, Free University of Bolzano/Bozen</td>
<td>2007 - Sep 2011</td>
<td>4 yrs 9 mos</td>
</tr>
<tr>
<td>Assistant Professor</td>
<td>Free University of Bolzen-Bolzano</td>
<td>2007 - Sep 2011</td>
<td>4 yrs 9 mos</td>
</tr>
<tr>
<td>PhD Student</td>
<td>DIST - University of Genova</td>
<td>2004 - 2007</td>
<td>3 yrs 1 mo</td>
</tr>
</tbody>
</table>

See all 6 experiences

### Education

<table>
<thead>
<tr>
<th>University</th>
<th>Field</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Università degli Studi di Genova</td>
<td>Msc + PhD, Engineering</td>
<td>1997 - 2008</td>
</tr>
</tbody>
</table>
“Realistic Synthetic Data”

From a dataset we learn its statistical properties under differential privacy and we use them to generate synthetic data
We used a public CV dataset
Real data can be used as well

Learning of conditional probabilities through a
differentially-private algorithm

Structured dataset with extracted information

To ML App
(structured data input)

Bayesian Network Learning

synthetic, structured dataset of relevant features

CV Generation with LLMs (GPT-2)

#3

corpus of CVs generated using the synthetic structured dataset

Information Extraction

corpus of CVs (PDF, plain text)

#1

To ML App
(unstructured data input)

Unstructured Data
Anonymization &
Synthetic CV Generation

Anonymized dataset

Unstructured Data Anonymization &
Synthetic CV Generation

Anonymized dataset

Anonymized dataset

Anonymized dataset
Step #2: Bayesian Network Learning
Bayesian Network for CV attributes: Toy Example

Original Structured Dataset from Step #1

<table>
<thead>
<tr>
<th>age</th>
<th>field</th>
<th>edu</th>
<th>exp</th>
<th>role</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>IT</td>
<td>B.Sc.</td>
<td>&lt;1 year</td>
<td>Intern</td>
</tr>
<tr>
<td>1</td>
<td>Science</td>
<td>Ph.D.</td>
<td>&lt;1 year</td>
<td>Data Scientist</td>
</tr>
<tr>
<td>2</td>
<td>IT</td>
<td>B.Sc.</td>
<td>3-6 years</td>
<td>Developer</td>
</tr>
<tr>
<td>3</td>
<td>IT</td>
<td>M.Sc.</td>
<td>1-3 years</td>
<td>Junior Developer</td>
</tr>
<tr>
<td>4</td>
<td>Economics</td>
<td>B.Sc.</td>
<td>1-3 years</td>
<td>Junior Sales Accountant</td>
</tr>
<tr>
<td>5</td>
<td>Economics</td>
<td>M.Sc.</td>
<td>3-6 years</td>
<td>Sales Accountant</td>
</tr>
<tr>
<td>6</td>
<td>Science</td>
<td>Ph.D.</td>
<td>3-6 years</td>
<td>Senior Data Scientist</td>
</tr>
<tr>
<td>7</td>
<td>Science</td>
<td>M.Sc.</td>
<td>6-10 years</td>
<td>Expert Data Scientist</td>
</tr>
<tr>
<td>8</td>
<td>IT</td>
<td>M.Sc.</td>
<td>10+ years</td>
<td>Expert Developer</td>
</tr>
<tr>
<td>9</td>
<td>Economics</td>
<td>M.Sc.</td>
<td>10+ years</td>
<td>Sales Manager</td>
</tr>
<tr>
<td>10</td>
<td>Economics</td>
<td>B.Sc.</td>
<td>10+ years</td>
<td>Expert Sales Accountant</td>
</tr>
<tr>
<td>11</td>
<td>Science</td>
<td>M.Sc.</td>
<td>10+ years</td>
<td>Data Science Manager</td>
</tr>
</tbody>
</table>

| F    | A         | P(Exp is <1|F,A) | ... | P(Exp is 10+|F,A) |
|------|-----------|--------|-----|--------|
| IT   | 20-30     | 1/3    | ... | 0      |
| IT   | 30-40     | 0      | ... | 1      |
| ...  | ...       | ...    | ... | ...    |
| Economics | 40-50 | 0      | ... | 1      |
Bayesian Network for CV attributes: Toy Example

- The structure is defined (no privacy loss)
- Learn conditional probabilities (CPDs) from data
- Provide **differential privacy** by perturbing CPDs to prevent information leakage
- Sample new values
Bayesian Network for CV attributes: Toy Example

- The structure is defined (no privacy loss)
- Learn conditional probabilities (CPDs) from data
- Provide **differential privacy** by perturbing CPDs to prevent information leakage
- Sample new values

<table>
<thead>
<tr>
<th>age</th>
<th>field</th>
<th>edu</th>
<th>exp</th>
<th>role</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20-30</td>
<td>IT</td>
<td>M.Sc.</td>
<td>1-3 years</td>
</tr>
<tr>
<td>1</td>
<td>20-30</td>
<td>Economics</td>
<td>B.Sc.</td>
<td>1-3 years</td>
</tr>
<tr>
<td>2</td>
<td>30-40</td>
<td>Economics</td>
<td>B.Sc.</td>
<td>10+ years</td>
</tr>
<tr>
<td>3</td>
<td>30-40</td>
<td>IT</td>
<td>M.Sc.</td>
<td>10+ years</td>
</tr>
<tr>
<td>4</td>
<td>40-50</td>
<td>Science</td>
<td>B.Sc.</td>
<td>1-3 years</td>
</tr>
</tbody>
</table>
Differential Privacy

- De facto standard for privacy-preserving statistical data analysis: why?
  - Strong mathematical guarantees
  - Independent of adversary side knowledge and computational power
  - Composability
  - Robustness to post-processing
  - ...

- The output distribution of a randomized algorithm should not be affected too much by small changes in its input

\[ f(\text{public}) \approx f(\text{private}) \]

- Can be achieved by adding calibrated Laplace noise
Step #3: CV Generation w/ GPT-2, GPT-3, GPT-4 (?)
Language models (GPT-X)

The cat

from: https://towardsdatascience.com/sentence-generation-with-n-gram-21a5eef36a1b

Natural Language Generation
Given a type of text (e.g., stories, summaries, fake news),
genenerate something which resembles what humans would write

Generation of synthetic text
A great advantage in terms of privacy and good as a data augmentation technique
Generating CV text

**Generated attributes**
- Personal details (name, address, residence, nationality, email, phone, ...)
- Education 1
- Education 2
- Education 3
- Years of working experience
- Work experience 1
- Work experience 2
- Work experience 3

**Fixed section titles with prompts**
1. PERSONAL INTRODUCTION
2. STUDIES AND EDUCATION
3. WORK EXPERIENCE AND CAREER
4. LANGUAGES
5. IT AND TECH SKILLS
6. HOBBIES AND INTERESTS
CURRICULUM VITAE:
Name: Ivana Losekann  
Nationality: DE  
Address: Pasquale-Dörschner-Ring 97  
70896 Apolda  
Country of residence: DE  
Phone number: +49(0)0251 55486  
Email: ilosekann@gehringer.de  
Field: IT - current position: Expert Developer

PERSONAL INFORMATION:  
Ivana Losekann is an IT specialist who specializes in IT application development.  
Field: IT - current position: Expert Developer  

EDUCATION:  
Humboldt University  
M.Sc. in Business Analytics  
B.Sc. in Computer Science  
Indian Institute of Technology Madras, India  
B.Sc. in Computer Science  

WORK EXPERIENCE AND CAREER - total years of work: 5-10  
In my latest job, I worked as an Expert Developer at Drubin GbR, Germany for 1-3 years. Currently, I’m based in India and working as a Senior Web Development Engineer with Vasto Ltd. Before that, I was working as a Developer at Drubin GbR, Germany for 3-5 years. My job consisted in In my previous job, I was the C++/C# Developer on the Software Engineering team and was the C#/.ASP. I worked as an Intern at Haeing, Germany for 1-3 years. As part of my duties I worked in the C/C++, Java, PHP and Web developer's area for more than 10 years.

LANGUAGE SKILLS:  
English,  
中文,  
English (native)

TECHNICAL SKILLS:  
Languages: PHP, Net, C

HOBBIES AND INTERESTS:  
My favourite movie of all time is "Wings of Desire" and the song "Love Theme" by Bollywood artist Aamir Khan.
CÁNDIDO NEVADO
Senior Data Scientist
Rao, Marne and Sharma, India
For 1-3 years
India

EXPERIENCE

CANDIDATE WITHOUT A LOCATION
Senior Data Scientist
Rao, Marne and Sharma, India
For 1-3 years
India

A DAY OF MY LIFE

My passions
In my free time I like - sports, music, movies, video games, watching movies, listening to music and playing games (e.g. NBA 2k, Hearthstone, etc.). In my free time I like watching my favorite movies and TV series like Breaking Bad, Game of Thrones, South Park, The Office.

HOBBIES AND INTERESTS

Public speaking issues with Yahoo investors
Public speaking issues with Yahoo investors

SOFTWARE DEVELOPER

TODD B. MORRIS
Software Developer
Multiplayer ICU, Inc. Subsidiary, Gomberg Medical, India
2018

EDUCATION

Ph.D. in Physics
Arzitu University, India
2019

Additional Synthetic CVs with Different Layouts

Work Experience
In my latest job, I worked as a Senior Data Scientist at Boschman, Germany for 1-3 years. I migrated to Hamburg, Germany after this, and worked for a couple of months as a Senior Data Scientist for a private company as an interim. Before, I was working as a Data Scientist at a OHR lab. Germany for 1-3 years. My job previous job, I worked as a Senior Data Scientist for 6 months. MYSEU-2015 Intern at the Bullo.

LANGUAGES

English
Chinese
French
Japanese

EDUCATION

Ph.D. in Physics
Arzitu University, India
2019

TODD B. MORRIS
Software Developer
Multiplayer ICU, Inc. Subsidiary, Gomberg Medical, India
2018

IT AND TECH SKILLS

Computer Science – MS
Evaluation

Comparing a 28k corpus of CVs with 60k generated CVs

Intrinsic Evaluation
• Comparison of a set of NLP metrics

Extrinsic Evaluation
• Performances of real and synthetic corpora in a common downstream task (classification of candidate role)
• Using FastText and BERT-based classifiers
• Comparison of raw and masked text (removal of explicit candidate role mentions)

More info? Check the paper at:
Conclusions & Call to Action

Data generation with privacy guarantees works!

• Synthetic CVs as training datasets or data augmentation

It can be applied to other domains

• Temporal sequences (Synthetic clickstream flow in CX)
• Bias evaluation
• …
Thank you.

Contact information:

Francesco di Cerbo
francesco.di.cerbo@sap.com

Volkmar Lotz
volkmar.lotz@sap.com