RAIS

Real-time Analytics for Internet of Sports

Marie Curie European Training Network

PRIVACY IN FITNESS DATA SHARING

Thomas Marchioro

RAIS Seminar September 24, 2020

Outline

- IoT fitness data and how they are currently managed
- Possible directions towards privacy preservation for fitness data
- Data anonymization and k-anonymity
- Distributed k-anonymity
- Anonymization and time series

IoT fitness data and how they are currently managed

What are IoT fitness data?

We call IoT fitness data all the information collected by **fitness apps** and associated **IoT wearable devices**.

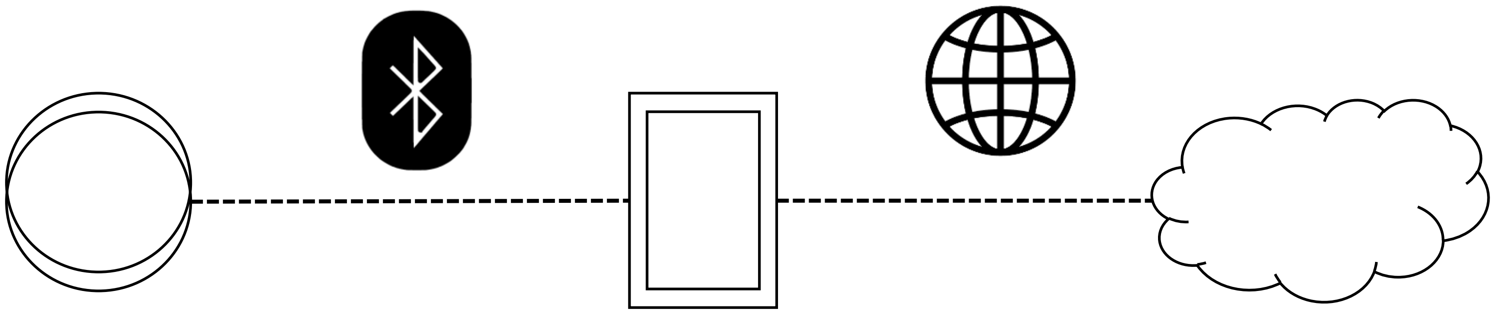
The information collected by these apps is partly *inserted by the user manually*

- Date of birth
- Height and weight
- Calories intake

and partly *automatically collected by the devices*

- Steps and calories consumption
- Heartrate
- Sleep hours

How IoT fitness data are currently managed



Wearable device

Smartphone app

Proprietary cloud

Collect raw data from sensor	Format and process data (e.g., oscillometers → # steps)	Analyse, store and share data

How IoT fitness data are currently managed

Most fitness apps' privacy policy:

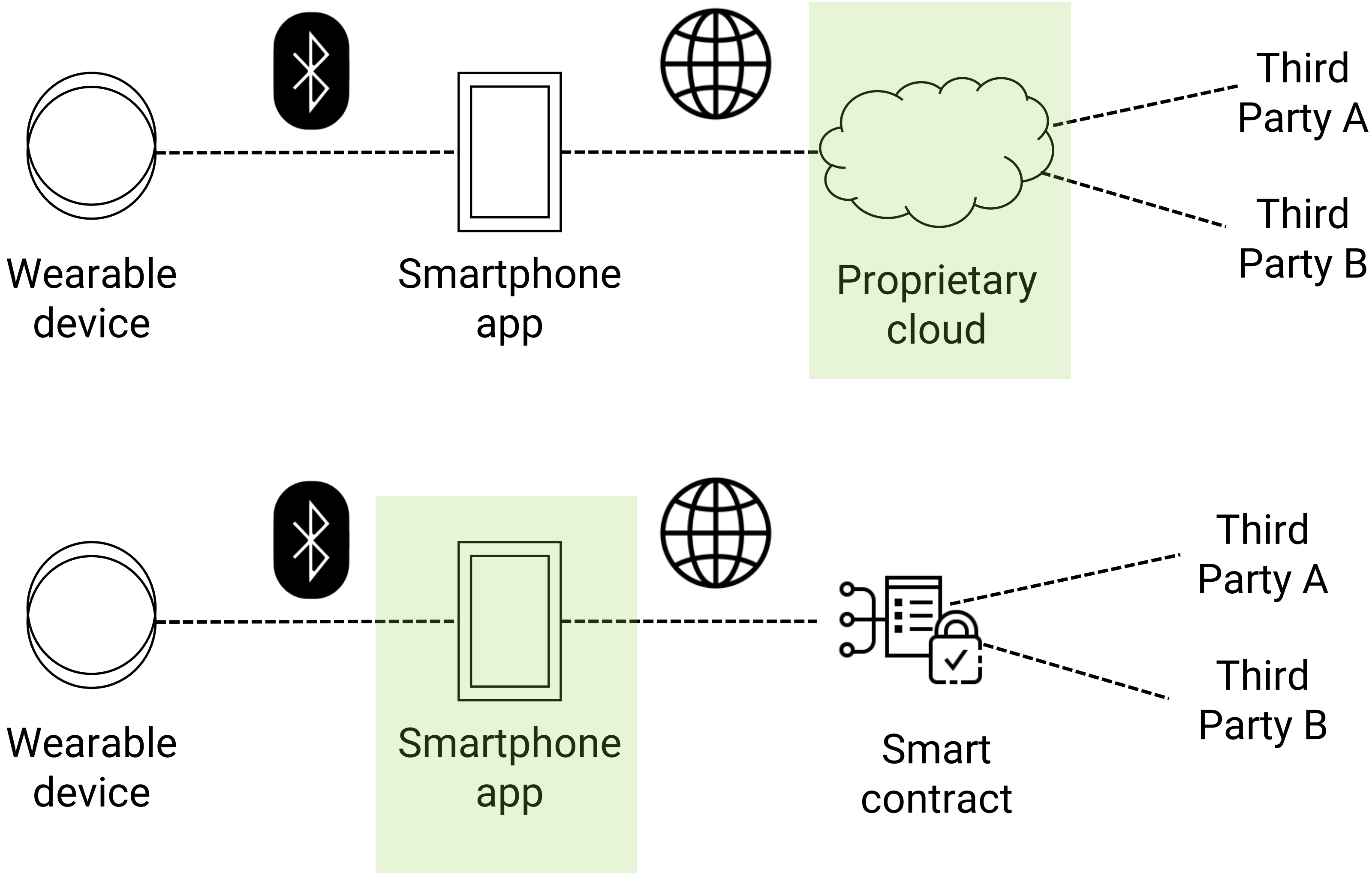
***.*.* Sharing with third party service providers and business partners**

To help us provide you with products and services described in this Privacy Policy, we may, where necessary, share your personal information with our third-party service providers and business partners.

This includes our delivery service providers, data centers, data storage facilities, customer service providers, advertising and marketing service providers and other business partners. These third parties may process your personal information on <COMPANY>'s behalf or for one or more of the purposes of this Privacy Policy. **We guarantee that the sharing of personal information necessary for providing services to you is solely for legitimate, legal, necessary, specific, and explicit purposes.** <COMPANY> will conduct due diligence and have contracts in place to ensure that third-party service providers comply with the applicable privacy laws in your jurisdiction. There may be occasions that third-party service providers have their sub-processors.

Possible directions towards privacy preservation for fitness data

Moving data management at user's side



Solutions for preserving user's privacy

- Process sensitive data locally or in a trusted environment

PROS	CONS
- Utility is fully preserved	- Hard to create trusted environments
- Guarantees almost complete privacy	- Third parties must disclose their code

- Disclosed anonymized data

PROS	CONS
- Various known methods	- Utility is not fully preserved
- Third parties don't have to disclose their code	- Privacy leaks are possible under continuous observation

Anonymization and k-anonymity

How fitness data look like?

- Most companies allow users to retrieve their personal data.

USER DATA

ID	Sex	Height	Weight	Nickname	Propic	Birth date
702*****	M	191.0	74.2	marchiorot	[URL]	1995-09-30

ACTIVITY DATA

Date	Steps	Distance	Calories
2020-02-30	12578	10238	2747
2020-02-31	8352	6632	2502
2020-02-32	13299	11014	2849
2020-02-33	6344	5536	2297

What makes a user anonymous?

- Consider a table of records where each user is assigned with a unique ID
- Is removing this ID sufficient to make users anonymous?

ID	Age	Sex	Steps
1758	26	M	[Time series]
1416	23	M	[Time series]
1932	26	F	[Time series]
2099	23	F	[Time series]
1896	21	F	[Time series]
1661	28	M	[Time series]
1522	28	F	[Time series]
2087	21	F	[Time series]

What makes a user anonymous?

- Say we have records of a group of users from two different months
- In this case, it is easy to link most users from first month with users from second month (even if the IDs are removed)
- This means that there is information that identifies these users

Month 1

Age	Sex	Steps
26	M	[Time series]
23	M	[Time series]
26	F	[Time series]
23	F	[Time series]
21	F	[Time series]
28	M	[Time series]
28	F	[Time series]
21	F	[Time series]

Month 2

Age	Sex	Steps
21	F	[Time series]
23	F	[Time series]
21	F	[Time series]
23	M	[Time series]
26	F	[Time series]
26	M	[Time series]
28	F	[Time series]
28	M	[Time series]

k-Anonymity

- The attributes that help distinguishing a certain user from the others are called *quasi identifiers*
- Notice that two people with identical quasi identifiers are in principle undistinguishable

Month 1

Age	Sex	Steps
26	M	[Time series]
23	M	[Time series]
26	F	[Time series]
23	F	[Time series]
21	F	[Time series]
28	M	[Time series]
28	F	[Time series]
21	F	[Time series]

Month 2

Age	Sex	Steps
21	F	[Time series]
23	F	[Time series]
21	F	[Time series]
23	M	[Time series]
26	F	[Time series]
26	M	[Time series]
28	F	[Time series]
28	M	[Time series]

k-Anonymity

- Idea of k-anonymity: generalize and/or suppress quasi identifiers until you form groups of undistinguishable users with at least k members (each group is called *anonymous class*)
- Applying one degree of generalization we obtain the following tables:

Month 1

Age	Sex	Steps
25-30	M	[Time series]
20-24	M	[Time series]
25-30	F	[Time series]
20-24	F	[Time series]
20-24	F	[Time series]
25-30	M	[Time series]
25-30	F	[Time series]
20-24	F	[Time series]

Month 2

Age	Sex	Steps
20-24	F	[Time series]
20-24	F	[Time series]
20-24	F	[Time series]
20-24	M	[Time series]
25-30	F	[Time series]
25-30	M	[Time series]
25-30	F	[Time series]
25-30	M	[Time series]

k-Anonymity

- We can suppress the "Sex" attribute and make the table 4-anonymous

Month 1

Age	Sex	Steps
25-30	*	[Time series]
20-24	*	[Time series]
25-30	*	[Time series]
20-24	*	[Time series]
20-24	*	[Time series]
25-30	*	[Time series]
25-30	*	[Time series]
20-24	*	[Time series]

Month 2

Age	Sex	Steps
20-24	*	[Time series]
20-24	*	[Time series]
20-24	*	[Time series]
20-24	*	[Time series]
25-30	*	[Time series]
25-30	*	[Time series]
25-30	*	[Time series]
25-30	*	[Time series]

Datafly algorithm

Input: table T with quasi identifiers (QIs) $A_1, \ldots, A_d$, generalization hierarchies

1. Build frequency list f of distinct QIs tuples in T containing pairs of the type (tuple, occurrences)
2. While (there are frequencies occurring less than k times that account for more than k tuples) do
   a. Set $A_i \leftarrow$ attribute with highest number of distinct values
   b. Update $f \leftarrow$ generalize the values of A_i in f
3. Update $f \leftarrow$ suppress sequencies in f occurring less than k times
4. Update $f \leftarrow$ enforce k requirement on suppressed tuples in f
5. Construct T_k from f
6. Return T_k

Execution of Datafly

Say generalization hierarchies are:

Age: $N \rightarrow [N - N\text{mod}5, N - N\text{mod}5 + 4] \rightarrow [N - N\text{mod}10, N - N\text{mod}10 + 9]$

Sex: {M, F, ?} → *

And we want to achieve 3-anonymity

	Age	Sex	Steps
	26	M	[Time series]
	23	M	[Time series]
	26	F	[Time series]
	23	F	[Time series]
	21	F	[Time series]
	28	M	[Time series]
	29	F	[Time series]
	21	F	[Time series]
Distinct vals:	5	2	-

Execution of Datafly

Say generalization hierarchies are:

Age: $N \rightarrow [N - N\text{mod}5, N - N\text{mod}5 + 4] \rightarrow [N - N\text{mod}10, N - N\text{mod}10 + 9] \rightarrow *$

Sex: {M, F, ?} → *

"Age" can be furtherly generalized

	Age	Sex	Steps
	25-30	M	[Time series]
	20-24	M	[Time series]
	25-30	F	[Time series]
	20-24	F	[Time series]
	20-24	F	[Time series]
	25-30	M	[Time series]
	25-30	F	[Time series]
	20-24	F	[Time series]
Distinct vals:	2	2	-

Execution of Datafly

Say generalization hierarchies are:

Age: $N \rightarrow [N - N\text{mod}5, N - N\text{mod}5 + 4] \rightarrow [N - N\text{mod}10, N - N\text{mod}10 + 9] \rightarrow *$

Sex: {M, F, ?} → *

3-anonymity is satisfied, Datafly stops

	Age	Sex	Steps
	20-30	M	[Time series]
	20-30	M	[Time series]
	20-30	F	[Time series]
	20-30	F	[Time series]
	20-30	F	[Time series]
	20-30	M	[Time series]
	20-30	F	[Time series]
	20-30	F	[Time series]
Distinct vals:	1	2	-

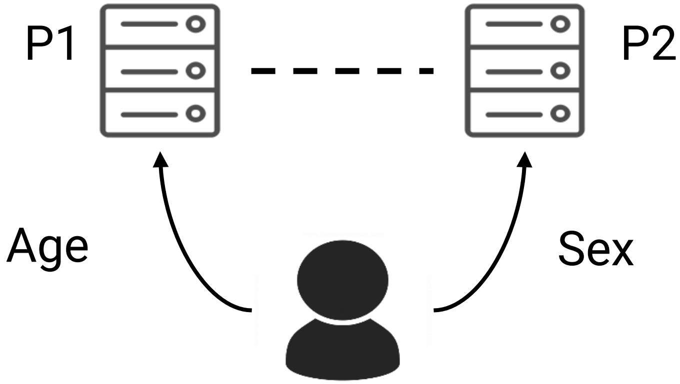
Distributed k-anonymity

Distributed k-anonymity

Problem: k-anonymity algorithms require access of a "trusted" entity to the whole data.

Solution: Distribute the k-anonymization process.

Core idea: vertically partition the data (i.e. split the columns) and distribute them among L "semi-trusted" parties $P_1, \ldots, P_L$, who apply generalization and suppression locally following a distributed version of the Datafly algorithm.



Distributed k-anonymity

Exploits the fact that Datafly acts on singular QI columns and requires to know only:

1) Whether k-anonymity has been reached

2) Which QI column has the highest number of distinct values

Each user is assigned with a temporary ID and must use the same for all the parties to whom the data are shared.

This way, parties can easily compare the common/distinct value of their columns denoting them as a partition of the set of users, e.g.

$$\gamma_1 = \big\{\{1,2,6\},\{3,4,5,7,8\}\big\}$$
$$\gamma_2 = \big\{\{1,3\},\{2,4\},\{5,6\},\{7\},\{8\}\big\}$$

Distributed k-anonymity

However sharing γ_1 and γ_2 would leak information about the users.

So how can P1 and P2 know when to stop without disclosing them?

Solution: **Secure Set Intersection**

If D_1 and D_2 are two subsets of the users' set, it is possible to compute two bits b_1 and b_2 such that

$$b_1 \oplus b_2 = 1 \iff |D_1 \cap D_2| \leq k$$

without disclosing D_1 and D_2.

Distributed k-Anonymity

P1 has access to the "Sex" QI, P2 has access to the "Age" QI.

Suppose we aim to achieve 3-anonymity. Initial partitions are:

For P1: $\{\{1,2,6\},\{3,4,5,7,8\}\}$

For P2: $\{\{1,3\},\{2,4\},\{5,6\},\{7\},\{8\}\}$

P1

ID	Sex	Steps
1	M	[Time series]
2	M	[Time series]
3	F	[Time series]
4	F	[Time series]
5	F	[Time series]
6	M	[Time series]
7	F	[Time series]
8	F	[Time series]

P2

ID	Age	Calories
1	21	[Time series]
2	23	[Time series]
3	21	[Time series]
4	23	[Time series]
5	26	[Time series]
6	26	[Time series]
7	28	[Time series]
8	28	[Time series]

Distributed k-Anonymity

Remark: A table T is k-anonymous *only if* each QI column is at least k-anonymous

Therefore the first step for both parties is to anonymize their QIs locally:

For P1: $\gamma_1^{(0)} = \{\{1,2,6\},\{3,4,5,7,8\}\}$

For P2: $\gamma_2^{(0)} = \{\{1,2,3,4\},\{5,6,7,8\}\}$

P1

ID	Sex	Steps
1	M	[Time series]
2	M	[Time series]
3	F	[Time series]
4	F	[Time series]
5	F	[Time series]
6	M	[Time series]
7	F	[Time series]
8	F	[Time series]

P2

ID	Age	Calories
1	20-24	[Time series]
2	20-24	[Time series]
3	20-24	[Time series]
4	20-24	[Time series]
5	25-30	[Time series]
6	25-30	[Time series]
7	25-30	[Time series]
8	25-30	[Time series]

Distributed k-Anonymity

P1 and P2 use commutative encryption and exchange $\Gamma_1^{(0)}$ and $\Gamma_2^{(0)}$.

After assessing $\Gamma_1^{(0)} \neq \Gamma_2^{(0)}$, both P2 generalizes one step further.

On the next step $\Gamma_1^{(1)} = \Gamma_2^{(1)}$.

Therefore, the two tables are joined into $T = T_1 \bowtie T_2$.

P1

ID	Sex	Steps
1	M	[Time series]
2	M	[Time series]
3	F	[Time series]
4	F	[Time series]
5	F	[Time series]
6	M	[Time series]
7	F	[Time series]
8	F	[Time series]

P2

ID	Age	Calories
1	20-30	[Time series]
2	20-30	[Time series]
3	20-30	[Time series]
4	20-30	[Time series]
5	20-30	[Time series]
6	20-30	[Time series]
7	20-30	[Time series]
8	20-30	[Time series]

Distributed k-Anonymity

When the algorithm ends:

1) 3-anonymity is satisfied

2) Privacy is preserved

Age	Sex	Steps	Calories
20-30	M	[Time series]	[Time series]
20-30	M	[Time series]	[Time series]
20-30	F	[Time series]	[Time series]
20-30	F	[Time series]	[Time series]
20-30	F	[Time series]	[Time series]
20-30	M	[Time series]	[Time series]
20-30	F	[Time series]	[Time series]
20-30	F	[Time series]	[Time series]

Anonymization and time series

Anonymization and time series

Until now, we considered as quasi identifiers only the attributes that are *fixed* over time.

Steps and calories data are different every time they are posted.

Nonetheless, this doesn't mean that they don't provide *identifying information*.

Consider this steps series from 2 different weeks for a 3-anonymous class

6819, 7154, 6354, 6947, 6820, 11435, 16281
3430, 3427, 3492, 3597, 3765, 8486, 2775
4485, 4474, 4929, 4585, 4549, 3486, 1431

3337, 3398, 3563, 4119, 3664, 7846, 981
4790, 4377, 4697, 4209, 4167, 3209, 1790
7057, 7132, 6821, 7301, 6455, 12133, 14512

Anonymization and time series

In this case, it is easy to link the series just from steps data

6819, 7154, 6354, 6947, 6820, 11435, 16281
3430, 3427, 3492, 3597, 3765, 8486, 2775
4485, 4474, 4929, 4585, 4549, 3486, 1431

3337, 3398, 3563, 4119, 3664, 7846, 981
4790, 4377, 4697, 4209, 4167, 3209, 1790
7057, 7132, 6821, 7301, 6455, 12133, 14512

Of course it is not always this easy.

But what if we add calories? Heartrate? Sleep hours?

Time of the activities? Location?

Future work

- Use existing algorithms or develop new ones to link similar time series

- Assess to what extent time series can be exploited to violate user's privacy

- Propose anonymization techniques that take time series into account

References

1. Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002 Oct;10(05):557-70.
2. Jiang W, Clifton C. Privacy-preserving distributed k-anonymity. InIFIP Annual Conference on Data and Applications Security and Privacy 2005 Aug 7 (pp. 166-177). Springer, Berlin, Heidelberg.
3. Jiang W, Clifton C. A secure distributed framework for achieving k-anonymity. The VLDB journal. 2006 Nov 1;15(4):316-33.

THANK YOU!