# Privacy-Preserving Self-Preferences

M2R Internship 2017-2018

**Supervised by :** Tristan Allard (Univ. Rennes 1/Irisa) & Tassadit Bouadi (Univ. Rennes 1/Irisa)

**Collaborators :** David Gross-Amblard (Univ. Rennes 1/Irisa) & Virginie Sans (Univ. Rennes 1/Irisa)

**Contact :** tristan.allard@irisa.fr and tassadit.bouadi@irisa.fr

**Keywords:** preference modeling, privacy, self-data, personal information management systems.

An ever-increasing quantity and diversity of personal data feeds the database systems of various companies (*e.g.,* emails, shopping baskets, news, geolocations, physiological measures, electrical consumption, movies, social networks, posts on forums, professional resume). Although individuals often benefit indirectly from this large-scale systematic capture of their data (*e.g.,* free access to services), the use value they get from it remains strongly limited by its fragmentation in non-cooperative *data silos* and by the usage allowed and supported by each silo [1, 3]. Personal information management systems (*PIMS* for short) aim at giving to individuals technical means to re-collect, integrate, manage, and use their data (or at least a part of it) in a central location *under their control* (*e.g.,* a personal computer, a rented virtual machine). The expected uses of a PIMS are those of a full-fledged data-centric personal assistant, including for example panoramic integrated personal data visualizations (*e.g.,* health data from health centres and physiological measures from wearables), vendor relationship management and privacy-preserving recommandations (*e.g.,* comparing offers from electricity providers given one's detailed electrical consumption), automatic completion of online profiles (*e.g.,* privacy preferences on a social network)[1]. Whatever the PIMS execution environment, it is dedicated to serve a single individual, or at most a few closely related individuals (*e.g.,* the members of a family). Therefore, the local computing resources — CPU and RAM — are scarce. However the promesses that PIMS carry on will be strongly hampered if they fail in eliciting *personal preferences* from the wealth of personal data they store, simply because they crucially rely on accurately modeling, reasoning, and using personal preferences.

In order to benefit from the self-preferences of similar individuals for enriching the local self-preferences, three steps are necessary : (1) groups of similar individuals must be formed (see, *e.g.,* [5]), (2) global group preferences must be aggregated from local self-preferences, and (3) *some* group preferences must be selected for enriching local self-preferences. Performing these steps (possibly combined together) in a distributed and privacy-preserving manner is a difficult challenge. In this internship, we will investigate step 1 and step 2.

Several previous works have already tackled problems similar to step 1 (*e.g.,* privacy-preserving distributed clustering, *k*-nearest neighbors) — although usually the underlying data is not preferences or the underlying model is not as sophisticated as elaborate preference models.

Concerning step 2, the aggregation of a set of preferences (see, *e.g.,* [2]) may require to perform operations not supported efficiently by encryption schemes (*e.g.,* comparison operators for implementing `MIN` aggregate, or threshold computations). Note that step 1 may also require preference aggregation. Moreover, the low computing resources of PIMS may not be able to cope with encryption-intensive algorithms, calling for alternative protection-by-perturbation strategies (*e.g.,* differential privacy) and consequently the design of specific perturbation mechanisms for preferences and specific privacy/utility tradeoffs.

The objective of the internship is to :

- Conduct a literature review on the problem of: (i) clustering of individuals based on their preferences, while providing at the same time sound privacy guarantees

[1] See, *e.g.,* https://tinyurl.com/mesinfosValue for a large variety of use-cases.

- Propose a robust and privacy-respectful method for grouping individuals according to the similarity of their self-preferences. This requires : (i) designing a new similarity measure that supports preference orders (see, *e.g.,* [6, 4]) and privacy mechanisms, (ii) developing a self-preference aggregation method that accurately capture meaningful group preferences shared by a significant number of individuals, in a privacy-preserving manner and from commodity hardware devices

- Implement the proposed method using our Raspberry Pi cluster and integrate into (see, *e.g.,* COZY[2]

- Evaluate experimentally the proposed approach based on preference datasets acquired through the crowdsourcing platform *Crowdflower*[3], or obtained from the following source: `http://www.preference-learning.org/#Datasets`

## References

[1] S. Abiteboul, B. André, and D. Kaplan. Managing Your Digital Life. *Commun. ACM*, 58(5):32–35, 2015.

[2] S. Basu Roy, L. V. Lakshmanan, and R. Liu. From group recommendations to group formation. In *Proceedings of SIGMOD'15*, pages 1603–1616, 2015.

[3] Y. de Montjoye, S. S. Wang, and A. Pentland. On the trusted use of large-scale personal data. *IEEE Data Eng. Bull.*, 35(4):5–8, 2012.

[4] M. Fattore, R. Grassi, and A. Arcagni. Measuring structural dissimilarity between finite partial orders. In *Multi-indicator Systems and Modelling in Partial Order*, pages 69–84. Springer, 2014.

[5] T. Kamishima and S. Akaho. Efficient clustering for orders. In *Mining complex data*, pages 261–279. Springer, 2009.

[6] M. Studer and G. Ritschard. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511, 2016.

[2] See `https://cozy.io/fr/` for more information

[3] See `https://www.crowdflower.com/` for more information