# MIMiS: Minimally Intrusive Mining of Smartphone User Behaviors

Pravallika Devineni[1], Evangelos E. Papalexakis[1], Kalina Michalska[2] and Michalis Faloutsos[1]
[1]Department of Computer Science, University of California, Riverside, CA
[2]Department of Psychology, University of California, Riverside, CA
pdevi002@ucr.edu {epapalex, michalis}@cs.ucr.edu kalinam@ucr.edu

***Abstract*—The proliferation of smartphones has lead researchers towards using them as an observational tool in psychological science. However, there is little effort towards protecting user privacy in these analyses. The overarching question of our work is: *Given a set of sensitive user features, what is the minimum amount of information required to group similar users?* Our contributions are two fold: we introduce privacy surfaces that combine sensitive user data at different levels of temporal granularity. Second, we introduce MIMiS, an unsupervised privacy-aware framework that clusters users as homogeneous groups with respect to their temporal signature. In addition, we explore the trade-off between intrusiveness and prediction accuracy. We extensively evaluate MIMiS on real data across a variety of privacy surfaces. MIMiS identified groups that are highly homogeneous w.r.t. user mental health scores and their academic performance.**

## I. Introduction

Smartphones come equipped with various sensors providing us the capability to continuously collect and analyze user data at scale. The fundamental questions in this work are a) *can we use such smartphone behaviors as proxies for estimating a user's mental state and well being?*; If so, b) *what is the minimum amount of sensitive information that we need in order to achieve that?* Sensor information is privacy intrusive and can compromise the safety of the user. To tackle the issue of privacy, we introduce the concept of *privacy surfaces* that combine user sensor information at varying levels of intrusiveness as shown in figure 1. We define intrusiveness based on the fineness of temporal granularity. We propose MIMiS, a privacy-aware unsupervised framework, which is based on PARAFAC2 decomposition. It takes as input a) smartphone data for a set of users and b) a privacy surface configuration, and produces groups of users that exhibit similar behavior over time. The clusters correlate very well with self-reported mental health states, such as depression and stress.

**Related Work:** We studied changes in user temporal behaviors and their relationship to real-life events [1] in our previous work. [2] used smartphones to study individual differences with regards to mental health. [3] used multi-set mining for mining electronic health records of patients and deriving time-evolving phenotypes.

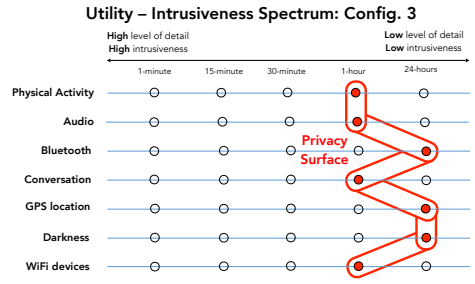



Fig. 1. An indicative **privacy surface** with low level of intrusiveness.

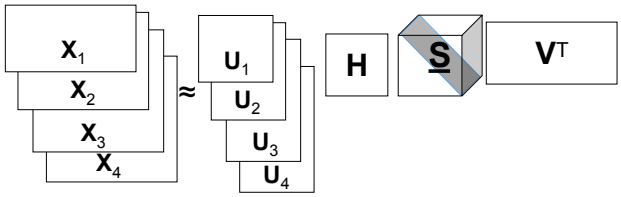



Fig. 2. PARAFAC2 decomposition of a multi-set

## II. Our Approach

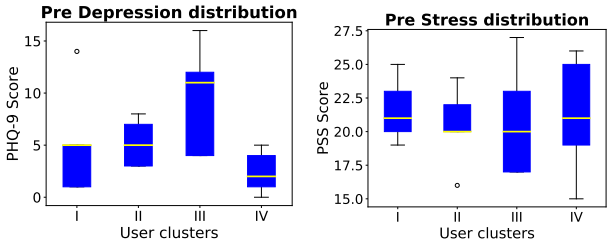
We propose to cast the problem of preserving privacy in smartphone data as an instance of tensor analysis. An $n$-mode tensor is a generalization of a matrix in $n$ dimensions. Multi-set is an irregular tensor with one incomparable mode with slices $\mathbf{X}_k$. We use PARAFAC2 decomposition to successfully deal with multi-set representation of data. PARAFAC2 decomposes each slice $\mathbf{X}_k$ as shown in Figure 2: $\mathbf{X}_k \approx \mathbf{U}_k\mathbf{S}_k\mathbf{V}^T$

where $k = 1, ..., K$, $\mathbf{U}_k \in \mathbb{R}^{I_k \times R}$, $\mathbf{S}_k \in \mathbb{R}^{R \times R}$ is a diagonal and $\mathbf{V} \in \mathbb{R}^{J \times R}$. The cross product of $\mathbf{U}_k^T\mathbf{U}_k$ is invariant regardless of the $k$ involved. This relaxes the CP model's invariance of the factor $\mathbf{U}_k$, thus preserving the uniqueness of the solution. For the above constraint to hold, each $\mathbf{U}_k$ factor us decomposed as: $\mathbf{U}_k = \mathbf{Q}_k\mathbf{H}$

### A. Dataset

StudentLife [2] is a 10-week study conducted on 48 Dartmouth students during 2013 spring quarter. We use the following from the dataset: smartphone sensors, pre and post mental health surveys, and academic performances. The smartphone sensor data include GPS, WiFi, accelerometer, call log, SMS, conversation, audio, phone lock, phone charge, and darkness.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 




(a) Per-cluster depression score. (b) Per-cluster stress score .

Fig. 3. Box plots showing dispersion of depression and stress scores per cluster for fig. 1 privacy surface, computed using PARAFAC2.

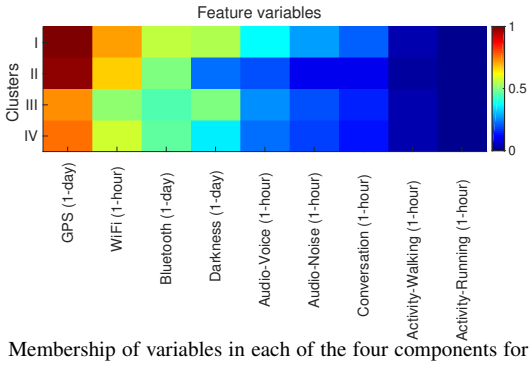



Fig. 4. Membership of variables in each of the four components for Config 3 from the factor matrix $\mathbf{S}_k$.

The mental health measures include standard psychometric scales for depression, stress and loneliness. We also use class schedules, GPA, deadlines per day, and Piazza participation for class to measure academic performance.

### *B. Proposed method: MIMiS*

Privacy surfaces investigate the trade-off between the intrusiveness and prediction capability. MIMiS finds user clusters which are coherent with user mental state.

**Step 1: Create privacy surface configurations**

We created three sets of privacy surfaces as multi-sets (time, user, feature), aggregated at five temporal granularities - 1-minute, 15-minute, 30-minute, 1-hour, and 1-day time bins. The finer the granularity, the higher the intrusiveness. Set 1 has all features and Set 2 has a reduced set of features, all at same granularity. Set 3 combines features with different granularities and a sample configuration is shown in Fig. 1.

**Step 2: PARAFAC2 for unsupervised user clustering**

We use the PARAFAC2 decomposition and its efficient implementation [3] to compute user clusters in an unsupervised fashion. We use AUTOTEN [4] to find the best approximation for rank $R$, the number of components.

*Model Interpretation*: We propose the following interpretation:

- For the common factor matrix $\mathbf{V}$, the non-zero values of each $r$th column indicate the *user membership* to the corresponding $r$th cluster.
- The diagonal $\mathbf{S}_k$ is a feature by cluster matrix that provides the *importance membership indicators* of the $k$th feature to each one of the $R$ clusters. Sorting the $R$ columns gives the feature importance to each cluster.
- Each $\mathbf{U}_k$ factor matrix provides the *temporal signature* of the cluster: each $r$th column of $\mathbf{U}_k$ reflects the temporal evolution of the $r$th cluster with respect the time granularity of feature $I_k$.

**Step 3: Validation using psychometric scales**

We analyze the components of the soft membership matrix $\mathbf{V}$ obtained from PARAFAC2 decomposition and consider the top-2 cluster memberships for each user. We hypothesize that members in a cluster share a similar mental state and temporal evolution. We compare depression and stress levels in the discovered clusters.

### *C. Exploratory analysis of discovered clusters*

We present an in-depth analysis of user clusters discovered using PARAFAC2 for privacy surface in fig 1. Fig 3 presents the box plots of depression and stress scores for top-10 users in each cluster, while fig 4 presents the order of importance of features in each cluster. The top-3 features are GPS, WiFi, and Bluetooth, those that characterize the physical mobility of users. We compared the temporal evolution of behavior in clusters against the academic information of the users - deadlines, piazza class forum participation, and overall GPA. Cluster 4 has the highest GPA of 3.56 and exhibits low mean depression score among clusters. Cluster 3 has mildly high depression and stress scores with a good overall GPA. We correlated the temporal evolution of features in each cluster with the deadlines of users and observed the following: cluster 1 displays low mobility and low sleep, which is also the case with cluster 4. Clusters 2 and 3 exhibit weak negative correlation with deadlines and have higher participation in Piazza online forum. Cluster 3 reported the highest loneliness and coincidentally, have lower GPA. This indicates a strong correlation between loneliness and academic performance. PARAFAC2 presented us with high-quality clusters where each cluster has a unique set of properties.

## III. CONCLUSIONS

We proposed the concept of privacy surfaces that combine sensors with different granularities to preserve privacy. Our proposed method MIMiS takes these privacy surfaces as input and employs multi-set decomposition in order to compute useful clusters. Our work is a step towards privacy-preserving analytics in an era of sensitive user information.

## IV. ACKNOWLEDGEMENTS

Research was supported by the Department of the Navy, Naval Engineering Education Consortium under award no. N00174-17-1-0005, DHS ST Cyber Security (DDoSD) HSHQDC-14-R-B00017 grant, NSF NeTS 1518878, and an an Adobe Data Science Research Faculty Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

## REFERENCES

[1] P. Devineni, E. E. Papalexakis, D. Koutra, A. S. Dogruöz, and M. Faloutsos, "One size does not fit all: Profiling personalized time-evolving user behaviors," in *ASONAM'17*, 2017, pp. 331–340.

[2] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *UBICOMP'14*. ACM, 2014, pp. 3–14.

[3] I. Perros, E. E. Papalexakis, F. Wang, R. Vuduc, E. Searles, M. Thompson, and J. Sun, "Spartan: Scalable parafac2 for large & sparse data," in *KDD '17*, 2017, pp. 375–384.

[4] E. E. Papalexakis, "Automatic unsupervised tensor mining with quality assessment," in *SDM'16*, 2016, pp. 711–719.