Station-to-User Transfer Learning: Towards Explainable User Clustering Through Latent Trip Signatures Using Tidal-Regularized Non-Negative Matrix Factorization

Liming Zhang[1], Andreas Zfle[1], and Dieter Pfoser[1]

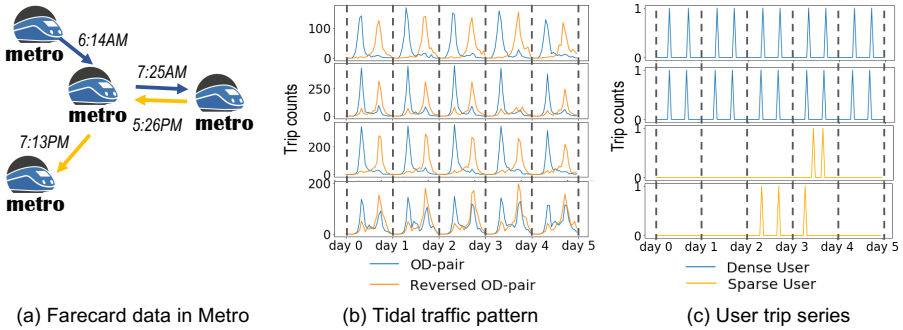
George Mason University, Fairfax VA 22030, USA
{lzhang22,azufle,dpfoser}@gmu.edu

Abstract. Urban areas provide us with a treasure trove of available data capturing almost every aspect of a population's life. This work focuses on mobility data and how it will help improve our understanding of urban mobility patterns. Readily available and sizable farecard data captures trips in a public transportation network. However, such data typically lacks temporal modalities and as such the task of inferring trip semantic, station function, and user profile is quite challenging. As existing approaches either focus on station-level or user-level signals, they are prone to overfitting and generate less credible and insightful results. To properly learn such characteristics from trip data, we propose a Collective Learning Framework through Latent Representation, which augments user-level learning with collective patterns learned from station-level signals. This framework uses a novel, so-called Tidal-Regularized Non-negative Matrix Factorization method, which incorporates domain knowledge in the form of temporal passenger flow patterns in generic Non-negative Matrix Factorization. To evaluate our model performance, a user stability test based on the classical Rand Index is introduced as a metric to benchmark different unsupervised learning models. We provide a qualitative analysis of the station functions and user profiles for the Washington D.C. metro and show how our method supports spatiotemporal intra-city mobility exploration.

Keywords: Urban Mobility · Matrix Factorization · temporal modalities · Spatial-temporal analysis.

1 Introduction

Traffic conditions in urban areas across the world remain a global challenge. According to the 2019 INRIX Traffic Scorecard [23], people in large cities across the world, such as Rio de Janeiro, Paris, and Chicago waste an average of more than 150 hours stuck in traffic, wasting hundreds of billions of USD and creating unnecessary greenhouse gas emissions. With two thirds of the population living in urban areas by 2050 [7], our future is characterized by mega cities in which urban mobility becomes a critical concerns. For these reasons and others, many



(a) Farecard data in Metro (b) Tidal traffic pattern (c) User trip series

Fig. 1: (a) A toy example of metro network. Each arrow indicates a trip with the timestamp; (b) Tidal traffic patterns observed within OD-pairs and associated reversed OD-pair (defined in Section 3); (c) dense users have strong recurrent pattern of commuting trips, while sparse users have weaker commuting pattern and are hard to be learned.

recent studies have focused on modeling and predicting human mobility in urban scenarios as surveyed in [28].

Paramount to improving urban mobility is to understand the purpose of trips of people [29]. Although the idea to directly collect trip purpose data through travel survey is a practice with a long history [20], ubiquitous computing and crowdsourcing data [29,8,10,9] provide a complementary way beside conventional time-consuming and costly field surveying. Yet, a data driven approach is challenging, as available trip information does not specify the purpose of a trip.

The goal of this work is to tackle this challenge by providing data-driven methods to learn the purpose of trips. More specifically, this work focuses on utilizing public transport data like Metro farecard data for urban mobility analysis. Millions of trip records in these Metro farecard data (Figure 1) provide data including: Card ID, Entry station, arrival stations, entry time, and arrival time. This data poses a number of unique challenges: (1) The purpose of a metro station is user-dependent, as the home-station of one user may be the work-station of another, and the recreation-station of yet another user; (2) a Card IDs do not injectively map to unique users, as one user may hold multiple cards including non-reusable temporary card, or multiple cards purchased at different time; (3) Trips may be missing, as user may choose to other modes of transport for certain trips. These challenges blur the signal of each individual user.

To address these challenges, we propose a "Station-to-User Transfer Learning" framework based on a novel domain-specific "Tidal-Regularized Non-Negative Matrix Factorization (TR-NMF)" machine learning model. This framework defines similarities of users by mapping them into a latent station feature space. Creation of this feature space exploits knowledge about "tidal" behavior of users having recurrent morning and evening peaks [3]. We also propose first-of-its-kind clustering stability test as a cross-model evaluation metric to promote future benchmarking in station and user clustering researches.

The remainder of this paper is organized as follows: After surveying related work in Section 2, we introduce the used datasets and formalize the problem of explainable user clustering in Section 3. Section 4 introduces our new Station-to-User transfer learning framework with its novel Tidal-Regularized Non-Negative Matrix Factorization to achieve explainable clustering based on trip semantics, and a clustering stability test metric. Next, in Section 5, we provide both quantitative and qualitative evaluations of our approach. Finally, in Section 6, we emphasize our contributions and future direction of works.

2 Related work

Early works on metro farecard data focus on descriptive statistics to characterize tidal pattern and dominant stations [17,10]. To learn the function of region in stations, Solutions have been proposed to infer the function of regions of a city based on individual mobility data in [34]. This work uses topic modeling to map point of interest and user visits of a region to latent topics. These latent topics that are leveraged to assess similarity between regions. Following this approach, it has been shown in [35] that the function of a region changes over time, and that it is paramount to consider temporal dynamics. Specifically using Smart Card data, latent factor based solutions have shown capable of recognizing daily patterns, such as weekdays, weekends, and holidays [33].

Related to our approach, a recent matrix factorization based approach to infer the temporal functions of regions (or stations) has been proposed in [30]. This approach has been leveraged to identify tidal patterns of human mobility in [26]. These works have in common that their goal is to infer the function of regions or stations. Our goal is to go a step further and to identify the "function" or signature of individual users, to assess the similarity of users to cluster them into groups of similar types of users. Towards inferring user-specific activities, solutions have been proposed in [8] using trajectory data and stop points. However, using only origin, destination, and time information available in metro farecard data, it is not possible to infer stops at specific points of interest to directly infer the purpose of a trip.

Non-negative Matrix Factorization (NMF) based solutions have also been proposed for other problem related to urban mobility, such as predicting road traffic [11,31] and predicting metro traffic demand [6]. These works provide powerful solutions to predict traffic, but lack explanation of patterns. To capture spatial and temporal mobility patterns, existing works [10,21,25] use NMF to explain temporal patterns in daily life, such as commuting pattern that concentrates on morning and evening, and explain the function of urban spatial urban areas. In a recent work by Wang et al. [27], a context-aware tensor decomposition is used to explain urban mobility over space and time using a tensor factorization approach. These works have in common that they allow to model similar spatial and temporal urban dynamics, such as days having similar mobility patterns and regions having similar function. However, these approaches do not allow to assess similarity among individual users and passengers. In contrast, our approach un-

wraps the signatures of individual metro users, allowing to cluster similar users to explain individual users and the purpose of their trips.

3 Problem Definition

In this section, basic definitions and problem setup are presented to formalize our problem.

Definition 1 (Trip Database). *Let $\mathcal{U}$ be a set of metro users, let $\mathcal{S}$ be a set of metro stations, $\mathcal{OD} = \mathcal{S} \times \mathcal{S}$ denote the set of all origin-destination station pairs, and let $\mathcal{T}$ be a set of time intervals or epochs. A trip database $\mathcal{D}$ is a collection of tuples $(u, (o, d), t) \in \mathcal{U} \times \mathcal{S} \times \mathcal{S} \times \mathcal{T}$, where $u \in \mathcal{U}$ is a user, $(o, d) \in \mathcal{OD}$ is an OD-pair, $o \in \mathcal{S}$ is the origin station, $d \in \mathcal{S}$ is the destination station,$t \in \mathcal{T}$ is the start time of the trip.*

With trip database, we can define temporal flow matrix for all OD-pairs which aggregate trip data and stores the number of trips, grouped by OD-pairs, for each time epoch.

Definition 2 (OD-pair Temporal Flow Matrix). *is denoted as $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{OD}| \times |\mathcal{T}|}$ matrix, such that:*

$$\boldsymbol{V}_{(o,d) \in \mathcal{OD}, t \in \mathcal{T}} = |\{x \in \mathcal{D} | x.o = o \wedge x.d = d \wedge x.t = t\}|$$

We further define a temporal flow matrix for each user, which aggregates the number of trips per user grouped by time epochs oblivious of stations.

Definition 3 (User Temporal Flow Matrix). *is denoted as $\boldsymbol{U} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{T}|}$, matrix such that:*

$$\boldsymbol{U}_{u \in \mathcal{U}, t \in \mathcal{T}} = |\{x \in \mathcal{D} | x.u = u \wedge x.t = t\}|$$

Given a OD-pair temporal flow matrix $\boldsymbol{V}$ and a user temporal flow matrix $\boldsymbol{U}$, **our problems of clustering users are as follows: separate users to different groups that maximize the internal similarity of travel.**

4 Station-to-User (S2U) Transfer Learning Framework

To better cluster users based on the purpose of their trips, we propose a framework to learn the temporal signature between stations and users in a collective manner. The diagram of this Station-to-User (S2U) Learning Framework is shown in Figure 2 and has three main steps.

Step 1: Tidal-regularized matrix factorization: Factorization of the OD-pair temporal flow matrix $\boldsymbol{V}$ (c.f. Definition 2) to find temporal latent features $\boldsymbol{H}$ and latent trip features $\boldsymbol{W}$. To obtain interpretable features, we employ a tidal-regularized loss function to better fit the (empirically grounded) tidal

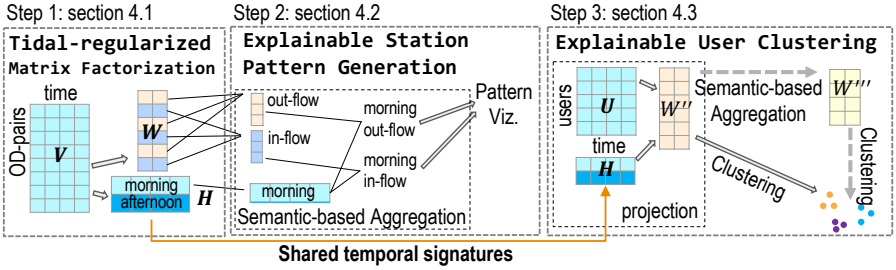


Fig. 2: Station-to-User (S2U) Learning Framework

pattern in urban mobility. More details on this matrix factorization approach are found in Section 4.1.

Step 2: Decomposing Tidal Features: Semantic-based Aggregation of latent trip features $\boldsymbol{W}$ to get tidal features of in- and out-flow weights $\boldsymbol{W}'$ for each station station.The weights indicate, for example, the degree to which a station is a work destination (inflow) or a home destination (outflow).Details of this approach are described in Section 4.2.

Step 3: Projection and clustering of user: Mapping temporal behavior of users $\boldsymbol{U}$ into the space of tidal features H. This yields a matrix $\boldsymbol{W}''$ containing the tidal features of each user. The reason for mapping users into the station space is that tidal features of stations are more stable and less noisy as shown by our experimental evaluation. This approach allows to provide explainable behavioral difference between users, even for users with only a few observed trips, thus making user clustering results more stable and informative. More details of this step are found in Section 4.3.

4.1 Tidal-regularized Non-negative Matrix Factorization (TF-NMF)

We decompose matrix $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{OD}|\times|\mathcal{T}|}$ into two non-negative matrices $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{OD}|\times K}$ and $\boldsymbol{H} \in \mathbb{R}^{K\times|\mathcal{T}|}$, such that

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H} =: \hat{\boldsymbol{V}},$$

where K is a positive integer, $\mathcal{OD}$ is the set of origin-destination pairs, and $\mathcal{T}$ is the set of temporal epochs (c.f. Definition 1). To find $\boldsymbol{W}$ and $\boldsymbol{H}$ we minimize a loss function $\mathcal{L}$ defined by the mean square approximation error and the L_1 and L_2 norms of $\boldsymbol{W}$ and $\boldsymbol{H}$ as follows [13]:

$$\mathcal{L} = \sum_i \sum_t (\boldsymbol{V}_{i,t} - \hat{\boldsymbol{V}}_{i,t})^2 + \alpha\eta(\|W\|_1 + \|H\|_1) + \alpha(1-\eta)(\|W\|_2 + \|H\|_2)$$

, where $\|\cdot\|_1$ is the L_1 norm of a matrix, $\|\cdot\|_2$ is L_2 (or Frobenius norm) of a matrix, and α, η are hyper-parameters.

Motivated by a *tidal traffic pattern* observed in urban areas (cf. [24,1,26]), we observe that the tidal-traffic pattern has strong temporal peaks, as the morning

commute happens before $11am$, while the reverse afternoon commute happens after $2pm$.

We incorporate this *a-priori knowledge* into our NMF approach by adding **a tidal-regularized (TR) loss to the generic NMF loss function.** It acts as a soft regularization to guide learned temporal signatures towards a better fit to such a tidal pattern. To understand the tidal regularized loss, we partition factor matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ to separate tidal features corresponding to daily morning and evening peaks. This approach is illustrated in Figure 3 and described as follows.

(i) Grouping latent features by temporal semantics. Generic NMF does not consider (or understand) temporal ordering, as temporal epochs (columns in $\boldsymbol{U}$ and $\boldsymbol{V}$) are treated as nominal (but not ordinal) variables. We sort latent features by their temporal semantics to understand and guide the learning process. We exploit that matrix $\boldsymbol{H}$ provides the temporal semantics of each latent feature: It describes each temporal epoch (such as each hour), by K latent features. Assuming tidal patterns, we expect some latent features to have larger weights in the morning epochs, and some latent features to have larger weights in the evening epochs. As an example in the upper right of Figure 3, the scatter plot shows the temporal semantics of 6 latent features Washington D.C. metro data. We observe that for latent feature 1 and 2, we clearly observe morning hour semantics. For features 5 and 6, the semantic feature is on the afternoon hours. Feature 3 are not clearly delineated between morning and afternoon. Feature 6 describe late afternoon and evening semantics.

We swap lines in matrix $\boldsymbol{H}$ such that the first $k \leq K$ feature correspond to the morning features, and the last $k' \leq K-k$ features correspond to the evening features. To ensure that this swapping of columns in $\boldsymbol{H}$ does not affect the factor product $\boldsymbol{V}$, we perform the same swaps among lines of $\boldsymbol{W}$ using the following observation.

Lemma 1. *Let $W \in R^{m \times K}$, $H \in R^{K \times n}$. Further, let $x, y \leq K$, let W' be obtained by swapping lines x and y in W, and let H' be obtained by swapping columns x and y, then*

$$WH = W'H'$$

Proof. Let $V = WH$, and let $V' = W'H'$. For any cell v_{ij} in V is derived by matrix multiplication as

$$v_{ij} = \sum_{k=1}^{K} w_{ik}h_{ki} = \sum_{k=1, k\neq x,y}^{K} w_{ik}h_{ki} + w_{xk}h_{kx} + w_{yk}h_{ky}$$

Equivalently, we obtain

$$v'_{ij} = \sum_{k=1}^{K} w_{ik}h_{ki} = \sum_{k=1, k\neq x,y}^{K} w_{ik}h_{ki} + w_{yk}h_{ky} + w_{xk}h_{kx}.$$

Since $w_{xk}h_{kx} + w_{yk}h_{ky} = w_{yk}h_{ky} + w_{xk}h_{kx}$ by commutativity of multiplication, we get $v_{ij} = v'_{ij}$ for any $i, j \leq k$. Thus $V = V'$.

Lemma 1 allows us to assume, without loss of generality, that columns of $\boldsymbol{W}$ and lines of $\boldsymbol{H}$ are grouped into morning features first and afternoon features last.

(ii) Grouping Symmetric OD-Pairs: Symmetric origin-destination pairs $a, b \mathcal{OD}, \forall a = (u, v), b = (v, u)$ have opposite direction of tidal traffic flow. To make learned latent tidal features fitting to tidal traffic flow, we aims to minimize the difference between morning flow of a and evening flow of b which is in the symmetric origin-destination of a. We drop all the OD-pairs with $u = v$, since it is meaningless for a user entering and exiting the same station.

(iii) Temporal partitions. Based on (i) and (ii), we can partition $\boldsymbol{W}$ and $\boldsymbol{H}$; five partitions for $\boldsymbol{W}$ according to OD direction and temporal ordering of latent components as shown in Figure 3, in which the white cells are weights of non-commuting signatures. Accordingly, there are five partitions for $\boldsymbol{H}$ based on temporal ordering of latent components and time of the day, shown in Figure 3, in which the white cells are non-commuting signatures. Figure 3 also shows how the tidal-regularized loss is calculated by minimizing the differences between Reverse OD-Pairs' total morning and afternoon commute flows. For example, the reconstructed morning trip matrix (u, v) is $\boldsymbol{W}^1\boldsymbol{H}^1$. Instead of calculating the difference for each t, we need the total accumulated morning trip flow for OD-pair (u, v) as $\boldsymbol{r}^{morning}_{(u,v)} = \sum_{t=1}^{t'} \left(\boldsymbol{W}^1_{(u,v),}, \boldsymbol{H}^1_{,t}\right)$. Similarly, its Reverse OD-Pair (u', v') the total afternoon flow as $\boldsymbol{r}^{afternoon}_{(u',v')} = \sum_{t=t'+1}^{T} \left(\boldsymbol{W}^4_{(u',v'),}, \boldsymbol{H}^4_{,t}\right)$. The next step is to add all Reverse OD-Pairs' differences, $\left(\boldsymbol{r}^{morning}_{(u,v)} - \boldsymbol{r}^{afternoon}_{(u',v')}\right)^2$. Given that the combinations $\boldsymbol{H}^3$ and $\boldsymbol{H}^2$ have as little flow as possible, we regularize them to (close to) zero.

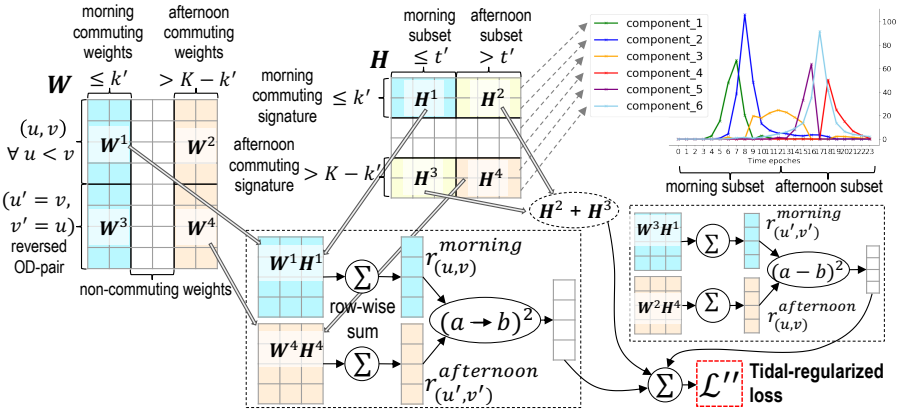


Fig. 3: Partitions of weight matrix $\boldsymbol{W}$ and latent component matrix $\boldsymbol{H}$ by Reverse OD-Pairs and temporal ordering, and compositions of tidal-regularized loss

In summary, the Tidal-regularized loss is formulated as follows.

$$\begin{aligned}\mathcal{L}'' =& \gamma \sum_{(u,v)=(1,2)}^{u\leq|V|-1,\forall u<v} \big((\boldsymbol{r}^{morning}_{(u,v)} - \boldsymbol{r}^{afternoon}_{(u',v')})^2 + (\boldsymbol{r}^{morning}_{(u',v')} - \boldsymbol{r}^{afternoon}_{(u,v)})^2\big) \\ &+ \rho\big(\|\boldsymbol{H}^3\|_F^2 + \|\boldsymbol{H}^2\|_F^2\big) \\ \boldsymbol{r}^{morning}_{(u,v)} &= \sum_{t=1}^{t'} \Big(\boldsymbol{W}^1_{(u,v),}\boldsymbol{H}^1_{,t}\Big),\ \boldsymbol{r}^{afternoon}_{(u',v')} = \sum_{t=t'+1}^{T} \Big(\boldsymbol{W}^4_{(u',v'),}\boldsymbol{H}^4_{,t}\Big) \\ \boldsymbol{r}^{morning}_{(u',v')} &= \sum_{t=1}^{t'} \Big(\boldsymbol{W}^3_{(u',v'),}\boldsymbol{H}^1_{,t}\Big),\ \boldsymbol{r}^{afternoon}_{(u,v)} = \sum_{t=t'+1}^{T} \Big(\boldsymbol{W}^2_{(u,v),}\boldsymbol{H}^4_{,t}\Big)\end{aligned} \tag{1}$$

The total regularization loss $\mathcal{L}$ is defined as $\mathcal{L} = \mathcal{L}' + \mathcal{L}''$. We use Tensorflow to develop our new algorithm and directly utilize the *Autodiff* features, which automatically compute the gradient update rules to optimize W and H. To terminate training, we either use the Mean-Square-Error $\mathcal{L}'$ (setting a minimum threshold) or use a fixed number of training steps. Post training, we convert the latent representation to a unit vector (cf. [32]) as follows: $h_{k,t} = \frac{h_{k,t}}{\sqrt{\sum_t^T h_{k,t}^2}},\ w_{i,k} = w_{i,k}\sqrt{\sum_t^T h_{k,t}^2}$

4.2 Decompose modalities as station functions and clustering stations with explainable temporal modality

Relating signatures to station functions: To cluster stations, we need to convert the OD-pair flow matrix to in-flows and out-flow for each station. To determine the function of a station (home, work, tourism, etc.) we examine the in- and out-flow of stations (cf. [10,4]). In-flow symbolizes attracting people for, e.g., work, and we refer to this as "attractivity" function of a place. Out-flow indicates people leaving, e.g., from home or hotels, and we refer to this as "generativity" function of a place. However, different from existing works, we want to distinguish not only between commuting and other functions, but also between different types of commuting (flexible work hours, etc.) for a station. We decompose station functions based on the meaning of temporal signatures $\boldsymbol{H}$ identified by TR-NMF.

Finding explainable station functions is achieved by "*semantics-based aggregation*" (cf. Figure 4) that cherry-picks specific temporal signatures according to our partitions and their peak hours. We can select only one type of temporal signature and use its corresponding weights to recover the in- and out-flow of each station. Temporal signatures are used to explain trip semantics, such as morning ($\boldsymbol{H}^1$) and afternoon commutes ($\boldsymbol{H}^4$) (cf. Figure 3). Furthermore, within both $\boldsymbol{H}^1$ and $\boldsymbol{H}^4$, there could be multiple rows of temporal signatures for different hours, e.g., $7am$, $8am$. Figure 4 gives an example for temporal signatures peaking at $7am$ and aggregates the OD-pair weights to recover the in- and out-flow for stations. The recovered flows inherit the strong semantic meaning for different

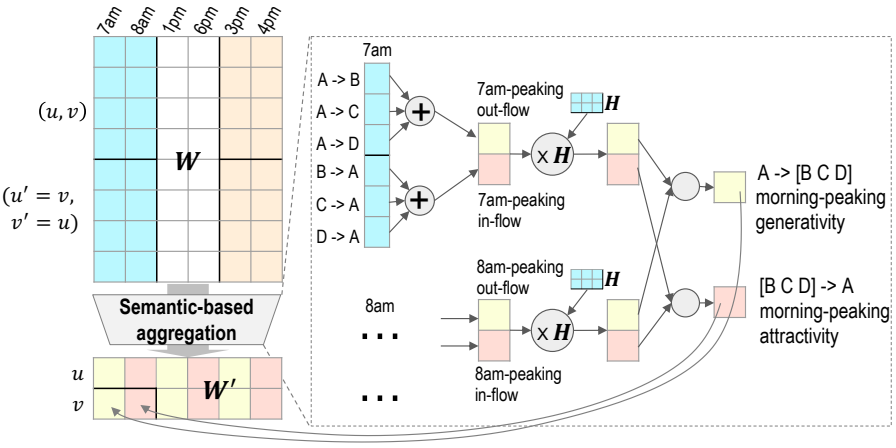


Fig. 4: Semantics-based aggregation operation to get explainable generativity and attractivity station functions

stations. Stations with large early hour in-flows are strong attractivity places, where stations with large out-flows are strong generativity places. Section 5.3 will give examples for Washington D.C.

4.3 Learning user clusters

To find the projection of users is the same dimension reduction task as NMF. We implement a part of a multiplicative update rule to project a user supported by the learned temporal modalities:

$$w''^{(i+1)}_{u,k} \leftarrow w''^{(i)}_{u,k} \frac{(\boldsymbol{U}_u \boldsymbol{H}^T)_{u,k}}{(\boldsymbol{W}^{(i)} \boldsymbol{H} \boldsymbol{H}^T)_{u,k}} \tag{2}$$

where $w''_{u,k}$ are weights of shared temporal signature of a matrix $\boldsymbol{W}''$ for user ID u and temporal modalities k (column). i is the number of iterations needed until a convergence criterion (like Mean Square Error) is met. More details on multiplicative update rules can be found in, e.g., [16,32].

After finding the weight matrix $\boldsymbol{W}''$ for users, a clustering algorithm like k-Means++ [2] can be used to cluster them. Since each weight is associated with shared temporal signature, we can again use a semantics-based aggregation to obtain $\boldsymbol{W}'''$ for a different explainable clustering.

4.4 User clustering stability test

To assess the performance of clustering methods, we introduce a novel domain-specific quantitative evaluation metric called "Clustering Stability Test". Although many different models are proposed in existing work using farecard data, no domain-specific quantitative evaluation metric is available to compare model

performance. Different model assumptions and procedures prevent cross-model comparisons for clustering.

Different metrics such as potential (sum of squared distances of samples to their closest cluster center) [12], log-likelihood score [12], perplexity score (information measure of generative probabilistic models) [18], AIC [12], and BIC [12], are used to assess clustering quality based on model assumption or information theory but they are not able to judge the stability of a clustering. Various works exist to test the stability of a clustering, e.g., [15,22]. Our proposed metric is based on Adjusted Rand Index (ARI) [14], a well-known measurement of the similarity between two clusterings with the same number of clusters K. For a set of data, like users $\mathcal{U}$, one clustering result assigns a set of group labels to each user with $X = \{x_1, x_2, \ldots, x_u\}$, while another clustering result assigsn a set of labels $Y = \{y_1, y_2, \ldots, y_u\}$. It can compute a ARI score $ARI_{x,y}$ based on these two label sets with random permutation of cluster label orders (cf. [14]). $ARI_{x,y}$ is a value with range $[-1, +1]$, where 0 indicates complete random labeling, $+1$ stands for a perfect match, and -1 indicates complete reversed labeling. However, generic ARI only tells if two clustering sets are similar. It cannot tell which method is better in terms of stability, which is something our new method addresses.

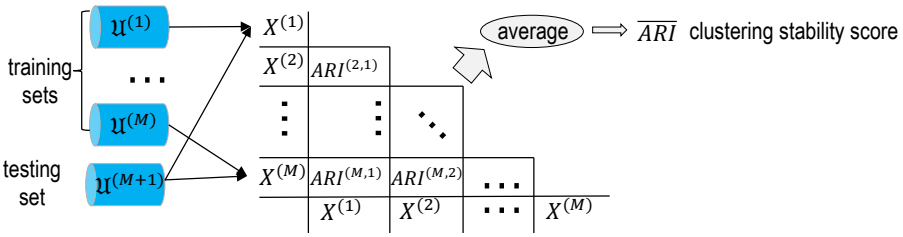


Fig. 5: Clustering stability test

Clustering stability in the context of semantic-poor farecard data aims to get a stable clustering, which is resilient to long-tail users who have abnormal or sparse behaviors, and without considering the internal processes. For example, we consider StoU framework as a whole method that output user labels, not just the internal k-means++ method. A stable clustering should capture those most significant behaviors without overfitting to those long-tail users. This means that if a large portion of users are unobserved, the user clustering labels do not change. So, our proposed procedure in Figure 5 is similar to a typical classification problem that different training data are used to predict testing data. This approach is also inspired by some input randomization works [15,22]. We partition original user ID set $\mathcal{U}$ to non-overlapping M training sets $\{\mathcal{U}^{(m)}\}_{m=1}^{M}$, and another non-overlapping testing set $\mathcal{U}^{(M+1)}$. Each method are applied to a mixed set that add a training set with the testing set, $\mathcal{U}'_m = \mathcal{U}^{(m)} + \mathcal{U}^{(M+1)}$. A label set $X^{(m)}$ can be got for $\mathcal{U}'_m$. For each pair of $X^{(m_i)}$ and $X^{(m_j)}$, we can get a ARI score $ARI^{(m_i,m_j)}$. Then, the average of all the paired ARI scores is used

as the "clustering stability score $A\bar{R}I$" to compare across different clustering methods. $A\bar{R}I$ is going to be in range $[-1, +1]$, where $+1$ indicates a perfectly stable method.

5 Experiments and Results

In this section, we compare Collective Learning Framework using three real-world datasets and two competing methods. Section 5.1 introduces our experimenting settings. Evaluation of competing methods and our proposed method are compared using the new clustering stability metric in Section 5.2. Then qualitative evaluations with spatial-temporal visualization are shown in Section 5.3 including explainable station clustering and user clustering results that decode urban mobility pattern in Washington D.C. as a case study.

5.1 Experimental settings

Real-world datasets: we utilize three real-world farecard mobility datasets from Washington D.C.. Metro farecard data is from Washington Metropolitan Area Transit Authority (WMATA), which covers metropolitan area of Washington D.C. in one week from May-01-2016 to May-07-2016 Each fare record only contains limited information, which are an anonymous card ID, an entry station ID, an exit station ID, an entry timestamp, and an exit timestamp. Some preprocessing is done to convert timestamps to 24 hour time snapshot of a day. There are total of about 3.57 million trip records, and about 0.8 million unique user IDs. Taxi data is collected from different taxi agencies required by open-data initiative of Washington D.C. metropolitan []. It includes data from Washington D.C.'s bike-sharing serving with 401 stations. To preprocess these data, we convert raw timestamps to 24 hours for $\mathcal{T}$. For metro data, we only keep sparsy users who have less or equal to 3 trips per week in this experiments. Since taxi data do not have stations, we use grid cells of 0.02 degree (about $2Km$) by 0.02 degree to build OD-pair temporal flow matrix. For metro, we do random selection without replacements to get 10 training sets with $50,000$ users per set and $10,000$ users in testing set. For taxi, similarly, we get 8 training sets with $10,000$ per set and $2,000$ users in testing set. For bike, we get 10 training sets with 300 users per set and 30 users in testing set. Table 1 is some descriptive summaries of our experiment datasets.

Metric: we use the clustering stability test procedure to get the average or median ARI scores of non-overlapping training sets. The higher ARI score is,

Data	Total users	Total trip	Training sets	Users per training	Users per testing
Metro	516,976	845,700	10	$50,000$	$10,000$
Taxi	89,237	89,237	8	$10,000$	$2,000$
Bike	3,032	51,325	8	$10,000$	$2,000$

Table 1: Descriptive summaries of experiment datasets

the better a method is. Because we introduce random splitting to get training sets, median (MED) and median absolute deviation (MAD) of a few dozens of runs are used to eliminate impacts from outlying cases.

Competing methods: Two competing methods are used with one controlled experiment method: 1) a naive model using raw trip flow of each time epoch as clustering features in KMeans++, noted as "Naive"; 2) a baseline model using NMF on temporal trip counts feature proposed in [5], and apply KMeans++ clustering on reduced weights matrix, noted as "NMF"; 3) a controlled experiment for S2U framework which identically replicates a training set for clustering stability test, noted as "Control". Its goal is to show the effectiveness of S2U.

5.2 Quantitative comparisons with clustering stability test

What follows is a discussion of the performance of S2U framework compared to competing methods. Given the low complexity of Matrix Factorization and KMeans++, all the running times are around a few seconds.

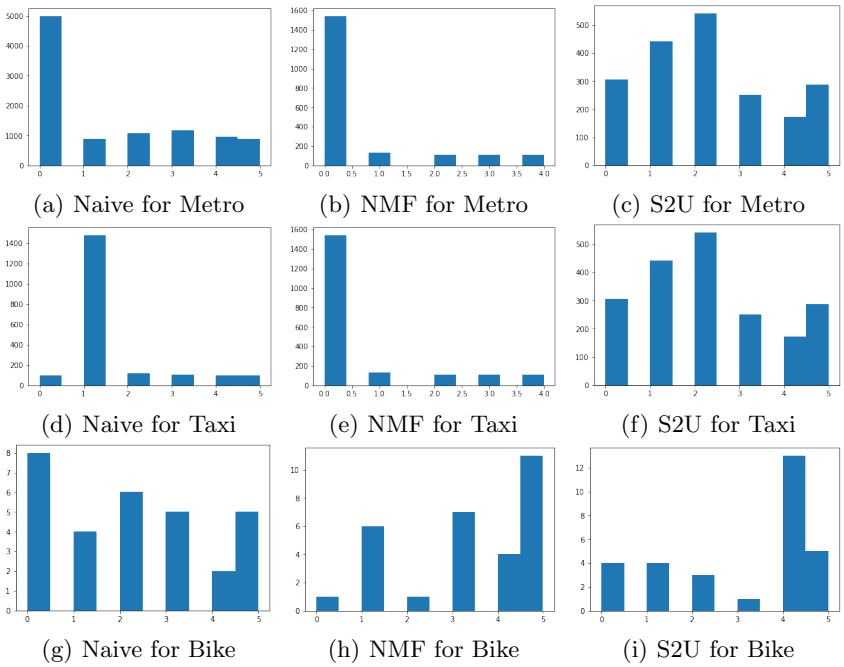
data	ARI scores	Naive	NMF	S2U	Control
Metro	[MED,MAD] of Mean	0.5217, 0.0474	0.6501, 0.0342	**0.7019, 0.0477**	0.8034, 0.0590
	[MED,MAD] of Median	0.5504, 0.0523	0.5815, 0.0333	**0.6496, 0.0821**	0.7347, 0.1400
Taxi	[MED,MAD] of Mean	0.5417, 0.0466	0.6605, 0.0804	**0.8117, 0.0388**	1.0, 0
	[MED,MAD] of Median	0.4781, 0.0222	0.6079, 0.0239	**0.8150, 0.0421**	1.0, 0
Bike	[MED,MAD] of Mean	0.5727, 0.1308	0.5412, 0.1147	**0.6347, 0.0836**	0.7846, 0.0921
	[MED,MAD] of Median	0.5697, 0.1293	0.5525, 0.1169	**0.6272, 0.0844**	0.7816, 0.1284

Table 2: Comparisons using user clustering stability test

Table 2 contains performances of different methods. In each table cell, first value is median (MED) of a hundred of experiment runs, while second value is median absolute deviation (MAD) value in a hundred of runs. Fig. 6 shows distributions of clustering labels for each data and each model. The main finding is that our S2U outforms other two competing methods for all three data in both MED of Mean ARI and MED of Median ARI. Even if we subtract the MAD scores from MEDs (which is 95% lower bound), S2U have a discounted lower bound of condifence close to two competing methods. For example of Metro, S2U have a subtracted value of $0.7019 - 0.0477 = 0.6542$, and NMF have a MED of 0.6501. For taxi data, the gain is even larger with S2U's lower bound of $0.8117 - 0.0388 = 0.7729$ and NMF's MED of Mean ARI of 0.6605. And, we are also confident in this result by examining Control model, which is consistantly much higher than normal S2U results. Additional proof of a better clustering is the less skewed distribution of clustering labels in Figures 6(c) and 6(f) compared to others of Naive and NMF models, which means that each cluster capture more meaningful patterns of sparsy users.. By checking clustering labels of Metro and Taxi with Naive and NMF in Fig. 6(a), 6(b), 6(d), and 6(e), it can be observed that there is a quite dominating cluster for both data. This is an indicator that

these two methods do not capture the real patterns because raw user temporal flow matrix $\boldsymbol{U}$ or decomposed $\boldsymbol{U}$ by NMF model are not informative for these sparsy users. That is why we could not conclude that NMF is just as good as raw $\boldsymbol{U}$ features for Metro, even though it gains 0.13 of MED of mean ARI. For bike data, clustering labels are more evenly distributed in different groups. It is not quite changed for S2U and competing methods.

Looking at more details of the results, by comparing Naive model with NMF model, NMF already performs better for Metro and Taxi data with 0.13 higher for MED of mean ARI scores and 0.03 higher for MED of median ARI scores. If we consider the variance in random splitting, the improvement of median ARI is not significant since MAD of NMF's median ARI is 0.03. But, the variance of NMF's mean ARI is only 0.03, so this is a significant improvement if we use NMF model for user clustering compared to raw features. But, NMF is not as good as Naive for bike data with about 0.01 to 0.02 lower on both MED of mean ARI scores and MED of median ARI scores. Of course, if we consider MADs of 0.1308 and 0.1147, the difference of 0.02 is smaller than variance. It is not a significant gain, and it is hard to say NMF is better than Naive approach. For Taxi data, there is a huge improvement for cluster label distribution, while MED of mean ARI improve a lot by 0.15 and MED of median ARI improve by a huge value of 0.21. The possible reason is that Taxi data are already quite cluster-able and dominated by commuting patterns that our method fit into this commuting pattern strongly.



(a) Naive for Metro (b) NMF for Metro (c) S2U for Metro

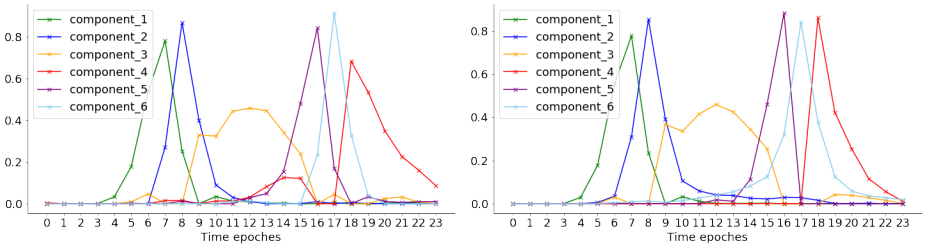
(d) Naive for Taxi (e) NMF for Taxi (f) S2U for Taxi

(g) Naive for Bike (h) NMF for Bike (i) S2U for Bike

Fig. 6: Histograms of User Clustering Labels (6 cluster number for all methods)

5.3 Qualitative evaluations

A few demonstrations of how our framework compare to others are shown in this part. In follow Figure 7, temporal signatures found by TR-NMF model is shown. The x axis is 24 hours of time epoches of a day, and y axis is the total signal in a time epoch. The left one 7(a) shows temporal signatures found by generic NMF algorithm, and the right one 7(b) shows temporal signatures found by TR-NMF. In both figures, components 1 & 2 are morning commute signatures, components 5 & 6 are afternoon commute signatures, and components 3 & 4 are non-commuting signatures. We can see the overall trends are more or less the same, while the improvement of TR-NMF is between the component 4 (red one of non-commuting signature) and the component 6 (skyblue one of afternoon commute signature). Component 4 lost a few signal around noon, while those signal are used to improve component 6, because those temporal features are temporally closer to component 6's peaking feature. This result of TR-NMF definitely show how tidal-regularized loss constrain the learning, and provides better explainable power to temporal pattern in metro data.

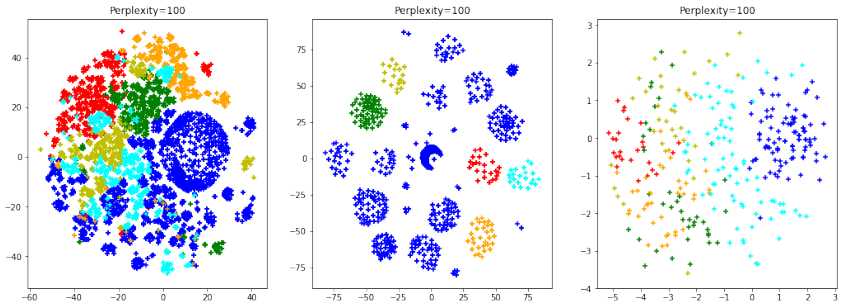


(a) Temporal signatures by generic NMF (b) Temporal signatures by TR-NMF

Fig. 7: Temporal signatures by generic Non-negative Matrix Factorization (NMF) and Tidal-Regularized Non-Negative Matrix Factorization (TR-NMF))

Explainable user clustering results t-Distributed Stochastic Neighbor Embedding (t-SNE) is a information-based machine learning visualization technique [19]. It can reduce dimension through non-linear manifold while preserving informative similarity pattern within data. With its popularity to visualize high dimension data, we use it to qualitatively demonstrate the intrinsic properties of raw data itself (first raw of Figure 8), and properties in transformed features by TR-NMF (second row of Figure 8). The x and y axis are reduced two features. Each point is a user. Different colors of points are clusering labels found previously (first row is Raw model and second row is S2U model). First row is the output of t-SNE using raw features, and second row is the output of t-SNE using TR-NMF transformed features. By comparing raw features and transformed features, we can see that TR-NMF features are more informative with more clear clustering patterns. For Metro (first column), raw users are pretty flattened, while transformed users are more concentrated. It is a similar case for Bike (third column). However, Metro and Bike are both challenging problem

themselves, since there are not strong clustering patterns in t-SNE transformed space. For Taxi, the raw features already contain strong similarity within several observable clusters, while TR-NMF did a good job to make those clusters condensed. Notice that there are more visually-appealing clusters in t-SNE space than S2U's clusters. But, for Naive model's clustering, blue point clustering includes most of the visually-appealing clusters. This is a minor problem for Taxi data since we can increase cluster number of S2U until the number reach a optimal value, however, increasing cluster number for Metro and Bike would not significantly improve clustering results.



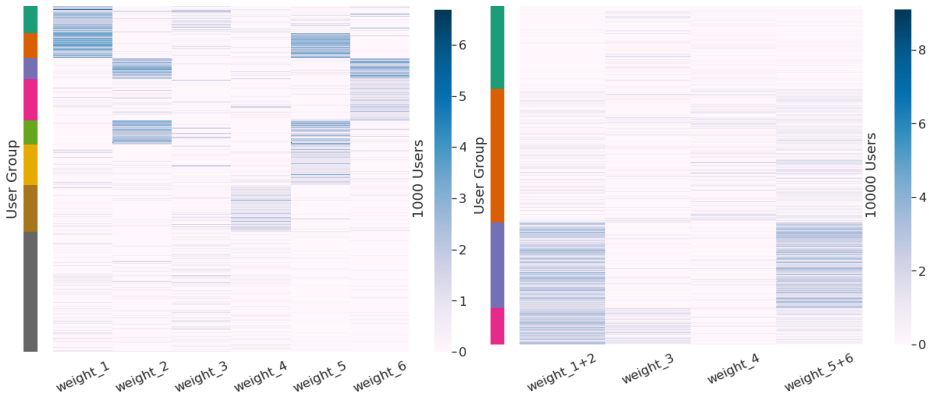
(a) Metro's Raw features (b) Taxi's Raw features (c) Bike's Raw features

(d) Metro with S2U (e) Taxi with S2U (f) Bike with S2U

Fig. 8: t-SNE visualization in which points are t-SNE transformed users using both raw and S2U-transformed users, and points' colors are based on Naive model and S2U model using 6 clusters.

Results of semantic-based aggregation of users: Two visualizations in Figure 9 demonstrate a qualitative performance of explainable clustering results based on S2U's raw weights and also semantic-based aggregated weights. In both heatmaps of sub-figures, each row of the heatmap represents a user, and each column is a weight for corresponding temporal signatures. The darker the blue is, the larger a weight value is, whose value can be found in the right-side color bar. Different colors in left-side color bars are cluster labels of different user groups. The x axis is noted by the same index of its associated temporal signatures. The right sub-figure 9(b) is noted by its combined temporal signatures' index, like

$weight\_1+2$ means this weight is got by semantic-based aggregating temporal signatures 1 and 2. Both sub-figures have a strong similarity between a user's weights of temporal signatures and other users in the same clusters. Based on this explainable clustering results, we can interpret human daily life in Metro data. For example in sub-figure 9(a), the brown cluster (the second group from top) are users who have early working hour $7am$ and early back-home hour around $3pm$. the pink cluster (the third group from top) contains users who have later working hour $8am$ and later back-home hour $5pm$. We can use explainable clustering to support many such analysis of users.



(a) S2U raw weights (b) Semantic-based aggregated weights

Fig. 9: Explainable User Clustering Labels (6 clusters for for flows before aggregation and 4 clusters for flows after aggregation)

Results of generated semantic-based station pattern: Using the tidal-regularized loss, the following visualization in Figure 10 shows that our TR-NMF could support more explainable station patterns combined with semantic-based aggregation. Both sub-figures demonstrate station locations (circles) on the area around The White House (the background map). The bigger the size of a circle, and the lighter a blue color is, the stronger attractivity pattern (recovered commuting in-flow for associated temporal signatures) is found. The black solid line is the Metro transit lines. The left sub-figure is based on the $7am$ commuting signature. We can see that station around The White House (mostly Federal Government offices) have more commuting flow which indicate a early working hour. In right sub-figure b), the station at Dupont Circle (a concentrated commercial area) has a larger $8am$ commuting in-flow, and stations around The White House decrease a little but still very strong. This example clearly illustrates how our S2U with TR-NMF can support explainable station patterns generation and intra-city function analysis.

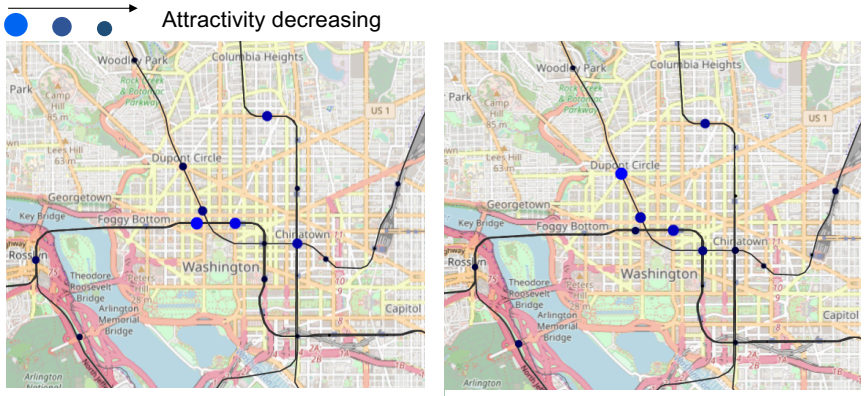


Fig. 10: Explainable stations pattern around White House at Washington D.C.

6 Conclusion

We propose a new Station-to-User (S2U) transfer learning framework to achieve a more explainable and stable learning of users in semantic-poor farecard data through transfering users to a latent feature space built with stations' temporal signatures. Also, we develop a novel Tidal-regularized Non-negative Matrix Factorization to guide the learning process towards a tidal-traffic commuting pattern that dominate urban transportation. To demonstrate the effectiveness of our work, we also set up a first-of-its-kind user stability test as a benchmarking evaluation metric to promote cross-model performance comparison. Lastly, we show that our framework improve 0.15 for mean ARI and 0.21 for median ARI in Taxi data experiment, and smaller margin of improvement for Metro and Bike data experiments. With visualization of t-SNE, we discuss the observations that reveal the power of S2U framework, and the difficulty of clustering tasks for Metro and Bike users. Finaly, we showcase how our explainable framework could support user behaviors analysis and station patterns analysis.

References

1. Alvizu, R., Zhao, X., Maier, G., Xu, Y., Pattavina, A.: Energy efficient dynamic optical routing for mobile metro-core networks under tidal traffic patterns. Journal of Lightwave Technology **35**(2), 325–333 (2016)

2. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Tech. rep., Stanford (2006)
3. Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C.: Human mobility characterization from cellular network data. Communications of the ACM **56**(1), 74–82 (2013)
4. Briand, A.S., Côme, E., Mohamed, K., Oukhellou, L.: A mixture model clustering approach for temporal passenger pattern characterization in public transport. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2015)
5. Carel, L., Alquier, P.: Non-negative matrix factorization as a pre-processing tool for travelers temporal profiles clustering. In: European Symposium on Artificial Neural Networks (2017)
6. Duan, Z., Lei, Z., Zhang, M., Li, H., Yang, D.: Understanding multiple days metro travel demand at aggregate level. IET Intelligent Transport Systems (2018)
7. of Economic, U.N.D., Affairs, S.: 2018 revision of world urbanization prospects. `https://population.un.org/wup/` (2018), accessed: 2019-06-09
8. Furletti, B., Cintia, P., Renso, C., Spinsanti, L.: Inferring human activities from gps tracks. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. pp. 1–8 (2013)
9. Gan, J., Zhang, J., Zheng, S.: Where you really are: User trip based city functional zone ascertainment. In: 2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC). pp. 1–8. IEEE (2018)
10. Gong, Y., Liu, Y., Lin, Y., Yang, J., Duan, Z., Li, G.: Exploring spatiotemporal characteristics of intra-urban trips using metro smartcard records. In: 2012 20th International Conference on Geoinformatics. pp. 1–7. IEEE (2012)
11. Han, Y., Moutarde, F.: Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. International Journal of Intelligent Transportation Systems Research **14**(1), 36–49 (2016)
12. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media (2009)
13. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of machine learning research **5**(Nov), 1457–1469 (2004)
14. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1), 193–218 (1985)
15. Lange, T., Roth, V., Braun, M.L., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural computation **16**(6), 1299–1323 (2004)
16. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001)
17. Liu, L., Hou, A., Biderman, A., Ratti, C., Chen, J.: Understanding individual and collective mobility patterns from smart card records: A case study in shenzhen. In: 2009 12th International IEEE Conference on Intelligent Transportation Systems. pp. 1–6. IEEE (2009)
18. Lucas-Cuesta, J.M., Fernández-Martínez, F., Moreno, T., Ferreiros, J.: Mutual information and perplexity based clustering of dialogue information for dynamic adaptation of language models. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 148–157. Springer (2012)
19. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
20. Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P.: Supporting large-scale travel surveys with smartphones–a practical approach. Transportation Research Part C: Emerging Technologies **43**, 212–221 (2014)

21. Poussevin, M., Tonnelier, E., Baskiotis, N., Guigue, V., Gallinari, P.: Mining ticketing logs for usage characterization with nonnegative matrix factorization. In: Big Data Analytics in the Social and Ubiquitous Context, pp. 147–164. Springer (2015)
22. Rakhlin, A., Caponnetto, A.: Stability of k-means clustering. In: Advances in neural information processing systems. pp. 1121–1128 (2007)
23. Reed, T.: Inrix global traffic scorecard (2019)
24. Taylor, M.A.: Network modelling of the traffic, environmental and energy effects of lower urban speed limits. Road & Transport Research **9**(4), 48 (2000)
25. Tonnelier, E., Baskiotis, N., Guigue, V., Gallinari, P.: Smart card in public transportation: Designing a analysis system at the human scale. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). pp. 1336–1341. IEEE (2016)
26. Troia, S., Sheng, G., Alvizu, R., Maier, G.A., Pattavina, A.: Identification of tidal-traffic patterns in metro-area mobile networks via matrix factorization based model. In: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 297–301. IEEE (2017)
27. Wang, J., Wu, J., Wang, Z., Gao, F., Xiong, Z.: Understanding urban dynamics via context-aware tensor factorization with neighboring regularization. IEEE Transactions on Knowledge and Data Engineering (2019)
28. Wang, J., Kong, X., Xia, F., Sun, L.: Urban human mobility: Data-driven modeling and prediction. ACM SIGKDD Explorations Newsletter **21**(1), 1–19 (2019)
29. Wang, P., Fu, Y., Liu, G., Hu, W., Aggarwal, C.: Human mobility synchronization and trip purpose detection with mixture of hawkes processes. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 495–503 (2017)
30. Wang, S., Xu, Y., Gao, S.: Revealing functional regions via joint matrix factorization based model. In: 2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC). pp. 205–209. IEEE (2016)
31. Wang, Y., Zheng, Y., Xue, Y.: Travel time estimation of a path using sparse trajectories. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 25–34 (2014)
32. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 267–273. ACM (2003)
33. Yang, C., Yan, F., Xu, X.: Daily metro origin-destination pattern recognition using dimensionality reduction and clustering methods. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). pp. 548–553. IEEE (2017)
34. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 186–194 (2012)
35. Zhang, K., Jin, Q., Pelechrinis, K., Lappas, T.: On the importance of temporal dynamics in modeling urban activity. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. pp. 1–8 (2013)