Unsupervised Instance and Subnetwork Selection for Network Data

Lin Zhang
International Digital Economy Academy (IDEA)
Shenzhen, China
zhanglin@idea.edu.cn

Nicholas Moskwa
Biological Sciences
University at Albany - SUNY
New York, USA
nmoskwa@albany.edu

Melinda Larsen
Biological Sciences
University at Albany - SUNY
New York, USA
mlarsen@albany.edu

Petko Bogdanov
Computer Science
University at Albany - SUNY
New York, USA
pbogdanov@albany.edu

***Abstract*—Unlike tabular data, features in network data are interconnected within a domain-specific graph. Examples of this setting include gene expression overlaid on a protein interaction network (PPI) and user opinions in a social network. Network data is typically high-dimensional (large number of nodes) and often contains outlier snapshot instances and noise. In addition, it is often non-trivial and time-consuming to annotate instances with global labels (e.g. disease/normal). How can we jointly select discriminative subnetwork and representative instances for network data without supervision?**

We address these challenges within an unsupervised framework for joint subnetwork and instance selection in network data, called UISS, via a convex self-representation objective. Given an unlabeled network dataset, UISS identifies representative instances while ignoring outliers. It outperforms state-of-the-art baselines on both discriminative subnetwork selection and representative instance selection, achieving up to 10% accuracy improvement on all real-world data sets we evaluate. When employed for exploratory analysis in RNA-seq network samples from multiple studies it produces interpretable and informative summaries.

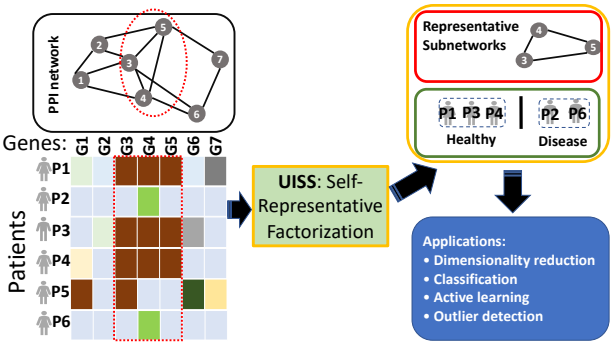


Figure 1: An illustrative example for unsupervised instance and subnetwork learning from a set of instances—patients' gene expression, sharing a common network structure—the human protein-protein interaction network (PPI) associating genes of known interactions. Our goal is to jointly select representative instances and discriminative subnetworks which can then be used for various downstream tasks.

I. Introduction

Network data abounds in a wide range of application domains: from activation snapshots of the human brain to the global state of sensor and online social networks. Different from tabular data, network data includes a shared structure associating features (nodes) in addition to their values. This structured feature space enables robust and interpretable solutions in a host of tasks, such as subnetwork selection [10, 34], network-aware distance measures [3] and summarization [37].

A network sample (i.e., instance) consists of feature values (node weights) and a structure shared by all instances. Consider the example in Fig. 1 of gene expression for 6 patients ($P1 - P6$), where the expression levels for each gene ($G1 - G7$) are employed as features and the human interactome is the structure associating the features. Our goal is to select a subspace of well-connected subnetworks which also "distinguish" the natural instance clusters. One such feature set in the example includes genes $\{G3, G4, G5\}$ forming a connected PPI subnetwork which also elucidates well-pronounced clusters of patients: $\{P1, P3, P4\}$ and $\{P2, P6\}$. Another goal is to identify outliers, such as patient $P5$ who does not "align" well with the natural clusters.

It is important to note that network instance data is different from network structure data employed for structural subgraph mining and classification since the latter employs subgraph occurrences as features as opposed to node values within a fixed structure [21]. Our setting is also different from collective classification [35] typically arising in social media mining [9], where individual nodes are instances as opposed to the whole network snapshot.

One important task in network instance data analysis is the prediction of the global state (label) of network instances [10, 47] with applications in neuroscience [6], gene expression on protein-protein interaction (PPI) networks [15] and analysis of the state of transportation and sensor networks [5]. Existing work typically focuses on the supervised setting, where each sample is characterized by a global label (e.g., disease/healthy). To tackle the typical high dimensionality existing work exploits the network structure to identify connected discriminative sub-spaces (subgraphs and associated feature values) to construct classifiers which generalize and offer interpretability [10, 34].

While it may be easy to obtain a network instance, obtaining

reliable global labels for supervision may be challenging due to expensive human-in-the-loop tests (e.g., doctor assessment for cognitive impairments [1]), or complex global states (e.g., regimes of operation of infrastructure networks such as road networks and power grids). In such scenarios unsupervised methods can both elucidate important subnetworks and representative instances to be labeled (akin to active learning [36]), while avoiding outlier and redundant samples.

We propose UISS, an unsupervised approach to jointly learn discriminative subgraphs and representative instances for network data. We formalize the problem as a self-representative factorization of the data which also maximizes the connectivity of selected nodes and the discriminative power of selected instances. In particular, we select subnetworks and instances that can reconstruct the full data (self-representation) and discount loss due to outlier instances. We enforce network locality for selected nodes by a quadratic penalty based on the network Laplacian. Finally, the representative power of selected instances is promoted via an orthogonality constraint.

The main contributions of this paper are as follows:

1. Significance: We address an unexplored problem of unsupervised learning for network samples by jointly detecting discriminative subgraphs and representative instances.

2. Robustness: UISS is robust to noise and outliers, outperforming baselines in feature and instance selection and achieving up to 10% accuracy improvement in real-world data.

3. Interpretability and applicability: UISS discovers interpretable subspaces such as discriminative city-center localities in bike rental behavior, and reconstructs the developmental timeline in RNA-seq samples from organoid development.

II. Related Work

Unsupervised feature selection: Methods in this category employ filtering [20] and embedding [45], however, the resulting features are not guaranteed to be discriminative for the actual (withheld) instance classes. Learning discriminative features based on cluster (pseudo) labels have been proposed when actual instance labels are absent [28]. Other discriminative feature selection methods [33, 27, 14] directly exploit cluster information: UDFS [45] maximizes the cluster separation and compactness; RUFS [33] learns jointly clusters and feature subsets by non-negative matrix factorization; NFDS [28] imposes orthogonality constraints to avoid cluster ambiguity; and MCFS [7] preserves across-cluster structure within a spectral embedding. Gu *et al.* [17] developed feature selection method using the structure among instances. While all above methods optimize the discriminative power of features, (i) they cannot select subnetworks in network samples,(ii) they are not robust to outliers which have been shown to impact unsupervised learning [29]; and (iii) they expect the number of clusters as an input parameter. We overcome these shortcomings by combining network smoothness and cluster-representative instance selection via an orthogonality constraint and by self-representation.

Feature selection in networked settings has also been proposed for cases when the instances are associated within a known network [25, 43, 26, 44]. However, this setting is different (and complementary) to ours in that the network structure associates instances as opposed to features.

Instance and feature co-selection: Feature and instance selection are co-dependent tasks since representative instances highlight relevant features and vice versa. CoSelect [39] and gActive [23] learn features and instances effectively but require global label information and are applicable in supervised settings only. Li *et al.* [24] developed an alternative unsupervised model, called ALFS which, however, does not consider a network structure among features or the existence of outliers, rendering it less robust than our solution as we demonstrate empirically. Our settings also differ from supervised active learning [36], which has been widely employed in text [40] and network data [39]. Active learning in a fully unsupervised setting has been pursued via transductive experimental design (TED) [46] which performs instance selection guided by reconstruction error. Following a similar minimum-reconstruction-error approach, ALFS [24] was proposed to jointly select representative instances and features. Similar to co-selection, the above approaches do not consider a network structure among features and are sensitive to noise and outliers due to their reliance on the Frobenius reconstruction norm [42].

Learning with network data: Closest to our setting are recent supervised methods for labeled network data [10, 11, 34, 47]. They adopt sampling or optimization to learn subgraphs with high discriminative power for supervised network state prediction. Unsupervised learning for network data, however, presents a new challenge which has not been studied in the literature to the best of our knowledge.

III. Preliminaries

A network instance $\mathcal{S}_i = \{V_i, E_i, \mathbf{X}_i\}$, is the triplet containing nodes V_i, edges E_i and node/feature values $\mathbf{X}_i$. A dataset $\mathcal{DS} = \{S_1 \dots S_n\}$ is a set of n network instances. All instances share the same network structure. Edges in $\mathcal{S}_i$ are weighted by the fraction of their occurrences in instances $\mathbf{M}_{pq} = 1/n \sum_i E_i(p,q)$. The graph Laplacian matrix of $\mathcal{S}_i$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{M}$, where $\mathbf{D}$ is the diagonal matrix of weighted node degrees, i.e., $\mathbf{D}_{pp} = \sum_q \mathbf{M}_{pq}$. The data matrix $\mathbf{X} = [X_1, X_2, ...X_n]^T \in \mathbb{R}^{n \times m}$ of a dataset $\mathcal{DS}$ is comprised of network instance values X_i in its rows.

IV. Problem Formulation

Given a dataset $\mathcal{DS}$ of n unlabeled network instances with m nodes, one of our goal is to select p ($p \ll n$) instances which are representative of the underlying clusters in the dataset and also to discover a connected subgraph subspace of q ($q \ll m$) nodes. Intuitively, the above design objective assumes that some instances form (an unknown number of) clusters, while others are outliers and do not belong to clusters.

Self-representative factorization: We model the joint instance and subgraph selection as a sparse self-representative factorization of the observations $\mathbf{X}$. In traditional matrix factorization $\mathbf{X} = \mathbf{UV}$ [13], the two factors can be thought

of as latent indicators of row and column clusters. Such representations are efficient and compact, but are also shown to exchange points from separate low-dimensional structures in the data, leading to sub-optimal clustering performance [31]. Instead, we propose a self-representative reconstruction based on the principle that a matrix can be represented by factors selected from its own rows and columns. In particular, we introduce selectors for features $\mathbf{P} \in \mathbb{R}^{m\times k}$ and instances $\mathbf{Q} \in \mathbb{R}^{k\times n}$, leading to:

$$\begin{aligned} &\underset{\mathbf{P},\mathbf{Q}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{V}\right\|_{2,1} + \lambda_1 \left\|\mathbf{P}\right\|_{2,1} + \lambda_2 \left\|\mathbf{Q}^T\right\|_{2,1}, \\ &\quad s.t.\ \mathbf{U}=\mathbf{X}\mathbf{P}, \mathbf{V}=\mathbf{Q}\mathbf{X}, \mathbf{P}\geq 0, \mathbf{Q}\geq 0, \end{aligned} \tag{1}$$

where the $L_{2,1}$ norm is employed to ensure robustness to outliers and noise through balance parameters λ_1 and λ_2. The key distinction from existing latent factorization is that factors are selected from the data columns and rows themselves. One can also view our co-selection strategy as imposing a reduced-rank structure on a CUR-like decomposition [31] via a bi-linear factorization employing $\mathbf{P}$ and $\mathbf{Q}$. Such a low-rank structure promotes a compact and interpretable latent space of the instances and subnetworks. Note that subnetwork and instance selection is based on thresholding $\mathbf{P}$'s rows and $\mathbf{Q}$'s columns rather than a fixed number of components pre-specified by the user. In particular, the importance of subnetworks and instances can be quantified based on the corresponding row norm of $\mathbf{P}$ and column norm of $\mathbf{Q}$.

Discriminative power using an orthogonal instance subspace: So far our objective ensures instance and subnetwork selection, but does not ensure that the selected instances are representative of the underlying clusters as opposed to outliers. Instead of explicitly modeling cluster memberships and deciding on the number of clusters apriori, we enforce orthogonality on the selected instances via a constraint of the form $\mathbf{V}\mathbf{V}^T = \mathbf{I}$. Columns of $\mathbf{V}$ act as cluster indicators, but also as a dictionary basis to reconstruct all instances, and thus this orthogonal constraint within the factorization ensures separation of selected instances. While the important role of orthogonality in traditional factorization clustering has long been recognized [13], we impose orthogonality within a self-representative factorization, leading to a superior performance in both feature and instance selection for network data.

Graph connectivity: We also impose "smoothness" of the subnetwork selection with respect to the common graph structure $\mathcal{S}$ to encourage connectivity in the feature space via a trace norm: $\text{Tr}(\mathbf{P}^T\mathbf{L}\mathbf{P})$, where $\mathbf{L}$ is the graph Laplacian matrix of the nodes' network.

Combining all modeling goals we obtain:

$$\begin{aligned} &\underset{\mathbf{P},\mathbf{Q}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{V}\right\|_{2,1} + \lambda_1 \left\|\mathbf{P}\right\|_{2,1} + \lambda_2 \left\|\mathbf{Q}^T\right\|_{2,1} + \lambda_3 \text{Tr}(\mathbf{P}^T\mathbf{L}\mathbf{P}), \\ &s.t.\ \mathbf{U}=\mathbf{X}\mathbf{P}, \mathbf{V}=\mathbf{Q}\mathbf{X}, \mathbf{V}\mathbf{V}^T=\mathbf{I}, \mathbf{P}, \mathbf{Q}\geq 0. \end{aligned} \tag{2}$$

The objective unifies our design goals: joint (i) subnetwork selection via $\mathbf{P}$, and (ii) discriminative instance selection via $\mathbf{Q}$, where discriminative power is promoted by $\mathbf{V}$'s orthogonality.

V. Optimization Algorithm

Our objective from Eq. 2 is jointly convex, however, it is non-trivial to develop gradient-based solutions for it directly because the $L_{2,1}$ norm is non-smooth. Instead, we propose an alternating optimization algorithm which efficiently updates the instance selector $\mathbf{P}$ and feature selector $\mathbf{Q}$ in turn until convergence.

Update of Q: To update $\mathbf{Q}$, we fix $\mathbf{P}$ and obtain:

$$\begin{aligned} &\underset{\mathbf{V},\mathbf{Q}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{V}\right\|_{2,1} + \left\|\mathbf{V}-\mathbf{Q}\mathbf{X}\right\|_F^2 + \lambda_2 \left\|\mathbf{Q}^T\right\|_{2,1}, \\ &\quad s.t.\ \mathbf{V}\mathbf{V}^T=\mathbf{I}, \mathbf{Q}\geq 0 \end{aligned} \tag{3}$$

We set the gradient w.r.t $\mathbf{Q}$ of Eq. 3 to zero:

$$(\mathbf{Q}\mathbf{X}-\mathbf{V})\mathbf{X}^T + \lambda_2\mathbf{Q}\mathbf{\Pi} = \mathbf{0}, \tag{4}$$

where $\mathbf{\Pi}$ is a diagonal matrix: $\mathbf{\Pi}_{ii} = (2\left\|\mathbf{Q}^i\right\|_2)^{-1}$. Thus, we obtain a closed-form solution for Eq. 4: $\mathbf{Q} = \mathbf{V}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda_2\mathbf{\Pi})^{-1}$. Note that $(\mathbf{X}\mathbf{X}^T + \lambda_2\mathbf{\Pi})$ is invertible since $\mathbf{X}^T\mathbf{X}$ is positive definite and $\mathbf{\Pi}$ is a non-negative diagonal matrix.

To handle the orthogonality constraint in Eq. 3, we introduce an intermediate variable $\mathbf{W}$ which approximates $\mathbf{V}$, leading to two subproblems:

$$\begin{cases} \underset{\mathbf{V}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{V}\right\|_{2,1} + \left\|\mathbf{V}-\mathbf{Q}\mathbf{X}\right\|_F^2 + \left\|\mathbf{V}-\mathbf{W}\right\|_F^2 \text{ (a)} \\ \underset{\mathbf{W}}{\operatorname{argmin}} \left\|\mathbf{V}-\mathbf{W}\right\|_F^2,\ s.t\ \mathbf{W}\mathbf{W}^T=\mathbf{I} \qquad \text{(b)} \end{cases} \tag{5}$$

We set the gradient w.r.t. $\mathbf{V}$ in Eq. 5(a) to $\mathbf{0}$:

$$\mathbf{U}^T(\mathbf{U}\mathbf{V}-\mathbf{X})\mathbf{\Theta} + \mathbf{V} - \mathbf{Q}\mathbf{X} + \mathbf{V} - \mathbf{W} = \mathbf{0}, \tag{6}$$

where $\mathbf{\Theta}$ is diagonal with elements $\mathbf{\Theta}_{ii} = \frac{1}{2\|\mathbf{X}^i-\mathbf{U}\mathbf{V}^i\|_2}$ when $\mathbf{X}^i - \mathbf{U}\mathbf{V}^i \neq \mathbf{0}$, and $\mathbf{0}$ otherwise. Since $\mathbf{U}^T\mathbf{U}$ is symmetric and positive semi-definite, its eigendecomposition $\mathbf{U}^T\mathbf{U} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$ has real eigenvalues (diagonal of $\mathbf{\Sigma}$) and real orthonormal eigenvectors in $\mathbf{A}$. Employing this decomposition, we can rewrite (6) as

$$\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\mathbf{V}\mathbf{\Theta} + 2\mathbf{V} = \mathbf{\Psi}, \tag{7}$$

where $\mathbf{\Psi} = \mathbf{U}^T\mathbf{X}\mathbf{\Theta} + \mathbf{Q}\mathbf{X} + \mathbf{W}$. Since $\mathbf{A}$ is orthonormal, we can multiply Eq. 7 by $\mathbf{A}^T$ on the left side, which after substituting $\mathbf{E} = \mathbf{A}^T\mathbf{V}$ can be simplified to $\mathbf{\Sigma}\mathbf{E}\mathbf{\Theta} + 2\mathbf{E} = \mathbf{A}^T\mathbf{\Psi}$. Individual elements of $\mathbf{E}$ can then be updated as follows : $\mathbf{E}_{ij} = \left[\mathbf{A}^T\mathbf{\Psi}\right]_{ij}/(\mathbf{\Sigma}_{ii}\mathbf{\Theta}_{jj} + 2)$ and $\mathbf{V}$ can be obtained as $\mathbf{V} = \mathbf{A}\mathbf{E}$.

There exists a closed-form solution $\mathbf{W} = \mathbf{R}\mathbf{I}\mathbf{T}^T$ solution for Eq. 5(b) due to Viklands *et al.* [41], where $\mathbf{R}$ and $\mathbf{T}$ are the left and right singular vectors of $\mathbf{V}$ in an SVD decomposition.

Update of P: The subproblem w.r.t $\mathbf{P}$ becomes:

$$\underset{\mathbf{P}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{X}\mathbf{P}\mathbf{V}\right\|_{2,1} + \lambda_1 \left\|\mathbf{P}\right\|_{2,1} + \lambda_3 \text{Tr}(\mathbf{P}^T\mathbf{L}\mathbf{P})$$

We again employ an auxiliary $\mathbf{U}$ to approximate $\mathbf{X}\mathbf{P}$:

$$\underset{\mathbf{P},\mathbf{U}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{U}\mathbf{X}\right\|_{2,1} + \lambda_1 \left\|\mathbf{P}\right\|_{2,1} + \lambda_3 \text{Tr}(\mathbf{P}^T\mathbf{L}\mathbf{P}) + \left\|\mathbf{U}-\mathbf{X}\mathbf{P}\right\|_F^2, \tag{8}$$

Algorithm 1 UISS

Input: Training data $\mathbf{X}$, and parameters$(\lambda_1, \lambda_2, \lambda_3, k)$
Output: Selection matrices $\mathbf{P}, \mathbf{Q}$

1: Initialize $\mathbf{P}, \mathbf{Q}$, $\mathbf{\Theta}$, $\mathbf{\Pi}$ and $\mathbf{K}$ to identity matrices; $\mathbf{U}$, $\mathbf{V}$, to random matrices;
2: **while** $\mathbf{P}, \mathbf{Q}$ not converged **do**
3: $(\mathbf{A}, \mathbf{\Sigma}) = evd\left(\mathbf{U}^T\mathbf{U}\right)$
4: $\mathbf{\Theta}_{ii} = \left(2\left\|\mathbf{X}^i - \mathbf{U}\mathbf{V}^i\right\|_2\right)^{-1}$ if $\mathbf{X}^i - \mathbf{U}\mathbf{V}^i \neq \mathbf{0}$
5: **while** $\mathbf{W}$ and $\mathbf{V}$ have not converged **do**
6: $\mathbf{\Psi} = \mathbf{U}^T\mathbf{X}\mathbf{\Theta} + \mathbf{Q}\mathbf{X} + \mathbf{W}$
7: $\mathbf{E}_{ij} = \left[\mathbf{A}^T\mathbf{\Psi}\right]_{ij}/(\mathbf{\Sigma}_{ii}\mathbf{\Theta}_{jj} + 2)$
8: $\mathbf{V} = \mathbf{A}\mathbf{E}$
9: $(\mathbf{R}, \mathbf{T}) = svd\left(\mathbf{V}\right)$
10: $\mathbf{W} = \mathbf{R}\mathbf{I}\mathbf{T}^T$
11: $\mathbf{Q} = max[\mathbf{V}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda_2\mathbf{\Pi})^{-1}, 0]$
12: $\mathbf{\Pi}_{ii} = (2\left\|\mathbf{Q}^i\right\|_2)^{-1}$ if $\mathbf{Q}_i \neq 0$
13: $(\mathbf{B}, \mathbf{\Lambda}) = evd\left(\mathbf{V}\mathbf{V}^T\right)$
14: $\mathbf{P} = max[(\mathbf{X}^T\mathbf{X} + \lambda_1\mathbf{G} + \lambda_3\mathbf{L})^{-1}\mathbf{X}^T\mathbf{U}, 0]$
15: $\mathbf{\Xi} = \mathbf{X}\mathbf{P} + \mathbf{K}\mathbf{X}\mathbf{V}^T$ $\mathbf{H}_{ij} = [\mathbf{\Xi}\mathbf{B}]_{ij}/(\mathbf{K}_{ii}\mathbf{\Lambda}_{jj} + 1)$
16: $\mathbf{U} = \mathbf{H}\mathbf{B}^T$
17: $\mathbf{K}_{ii} = (2\left\|\mathbf{X}_i - \mathbf{U}_i\mathbf{X}\right\|_2)^{-1}$ if $\mathbf{X}_i - \mathbf{U}_i\mathbf{X} \neq 0$
18: **return** $\mathbf{P}, \mathbf{Q}$

and alternate between $\mathbf{P}$ and $\mathbf{U}$ updates. We set the gradient w.r.t. $\mathbf{U}$ in Eq. 8 to $\mathbf{0}$:

$$\mathbf{K}(\mathbf{U}\mathbf{V} - \mathbf{X})\mathbf{V}^T + \mathbf{U} - \mathbf{X}\mathbf{P} = 0, \tag{9}$$

where $\mathbf{K}$ is a diagonal matrix with elements $\mathbf{K}_{ii} = (2\left\|\mathbf{X}_i - \mathbf{U}_i\mathbf{V}\right\|_2)^{-1}$ when $\mathbf{X}_i - \mathbf{U}_i\mathbf{X} \neq 0$, and $\mathbf{K}_{ii} = 0$ otherwise. Similar to the solution for $\mathbf{V}$, we employ the eigendecomposition of $\mathbf{V}\mathbf{V}^T = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^T$ and multiply by $\mathbf{B}$ on the right, thus simplifying Eq. 9 to $\mathbf{K}\mathbf{H}\mathbf{\Lambda} + \mathbf{H} = \mathbf{\Xi}\mathbf{B}$, where $\mathbf{H} = \mathbf{U}\mathbf{B}$ and $\mathbf{\Xi} = \mathbf{X}\mathbf{P} + \mathbf{K}\mathbf{X}\mathbf{V}^T$. After updating the elements of $\mathbf{H}$ according to $\mathbf{H}_{ij} = \frac{[\mathbf{\Xi}\mathbf{B}]_{ij}}{\mathbf{K}_{ii}\mathbf{\Lambda}_{jj}+1}$, we get a closed-form solution for $\mathbf{U}$: $\mathbf{U} = \mathbf{H}\mathbf{B}^T$. We set the gradient w.r.t. $\mathbf{P}$ in Eq. 8 to $\mathbf{0}$:

$$\mathbf{X}^T(\mathbf{X}\mathbf{P} - \mathbf{U}) + \lambda_1\mathbf{G}\mathbf{P} + \lambda_3\mathbf{L}\mathbf{P} = 0, \tag{10}$$

where $\mathbf{G}$ is diagonal with elements $\mathbf{G}_{ii} = (2\left\|\mathbf{P}_i\right\|_2)^{-1}$ when $\left\|\mathbf{P}_i\right\|_2 \neq 0$, and $\mathbf{G}_{ii} = 0$ otherwise. Thus, $\mathbf{P}$ has the following closed-form solution:

$$\mathbf{P} = (\mathbf{X}^T\mathbf{X} + \lambda_1\mathbf{G} + \lambda_3\mathbf{L})^{-1}\mathbf{X}^T\mathbf{U}, \tag{11}$$

where $\mathbf{X}^T\mathbf{X} + \lambda_1\mathbf{G} + \lambda_3\mathbf{L}$ is invertible due to all summands being positive semi-definite matrices.

Overall algorithm, complexity and convergence: The detailed steps of the UISS are presented in Algorithm 1 and follow our update derivations above. We considered both iteration to convergence for each of the sub-problems independently and iterations of one update for each of the variables. The former strategy results in faster convergence for $\mathbf{W}$ and $\mathbf{V}$, hence this is the one we present in the Alg. 1. It is important to note that all updates to diagonal matrices (Steps 5, 15, 21) are applied after they are re-initialized to all-zeroes at each iteration.

The complexity of our algorithm is determined by the number of iterations to convergence (in both loops) and the complexity of individual steps in the loops. The most costly steps involve the SVD operation in (Step 10) and the matrix inversions (Steps 13 and 17). While svd has a super-quadratic complexity in the worst case, efficient implementations enable good scalability in practice. The matrix inversions can also be implemented efficiently due to the low-rank updates in each iteration. Note that in Step 13 $\mathbf{X}\mathbf{X}^T$ and in Step 17 $\mathbf{X}^T\mathbf{X}+\lambda_3\mathbf{L}$ are constants and can be inverted once and re-used in subsequent iterations, thus exploiting the sparse structure via the Sherman-Morrison formula for sparse inverse updates:

$$(\mathbf{A} + \gamma\mathbf{D})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{d}\mathbf{d}^T\mathbf{A}^{-1}}{1 + \mathbf{d}^T\mathbf{A}^{-1}\mathbf{d}}, \tag{12}$$

where A corresponds to the constant matrices in the two steps, $\mathbf{d} = \sqrt{\gamma diag(\mathbf{D})}$ is a column vector and the square root is applied element-wise thus $\gamma\mathbf{D} = d^T d$. Therefore, we can reduce the complexity of inversion as $O\left(mn_N\right)$, where n_N is the number of non-zero elements in $\mathbf{X}$. The pair $(\mathbf{D}, \gamma)$ from the above formula corresponds to $(\mathbf{\Pi}, \lambda_2)$ in Step 12 and $(\mathbf{G}, \lambda_1)$ in Step 16 respectively. *We will publish the implementation of UISS at the camera-ready paper.*

All subproblems in the Alg. 1 are solved in an alternating manner and have closed-form solutions, where $\{\mathbf{P}, \mathbf{Q}, \mathbf{U}, \mathbf{V}\}$ are optimal solution of Eq. 2. These closed-form solutions guarantee the overall optimization sequence converges to the primal-dual optimal solutions. Meanwhile, the constraints in Alg. 1 will converge to zero, i.e. $\left\|\mathbf{P}^{iter+1} - \mathbf{P}^{iter}\right\|_F \rightarrow 0, \left\|\mathbf{Q}^{iter+1} - \mathbf{Q}^{iter}\right\|_F \rightarrow 0$. Therefore, we can obtain the global convergence of UISS because it features both sequence convergence and constraint convergence [19].

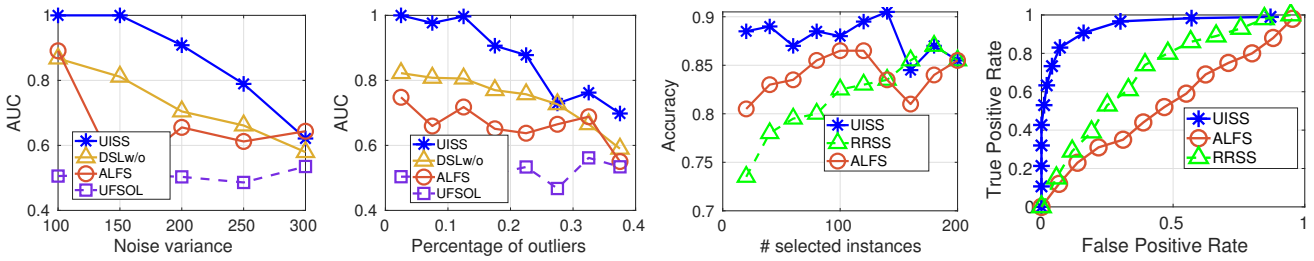
Dataset	$\lvert\mathcal{V}\rvert$	$\lvert E\rvert$	$\lvert\mathcal{DS}\rvert$	Hidden Classes
Synthetic	100	637	400	positive/negative
Bike [2]	142	1,723	299	weekday/weekend
Embryo [15]	1,321	5,227	34	tissue layers
CCT [8]	4,665	270,571	184	weekday/weekend
ADNI [1]	6,216	683,760	173	AD/NC
Liver [22]	7,383	251,916	123	disease/control

Table I: Summary of evaluation datasets' statistics.

VI. Experimental Evaluation

A. Datasets.

We employ both synthetic and real-world datasets for evaluation and summarize their statistics in Table. I. We synthesize geometric networks by uniformly sampling nodes in a unit square and connecting nodes at distances smaller than a threshold of $\tau = 0.2$. We select well-connected ground truth subgraphs and generate balanced set of instances labeled by two (hidden) global states. Particularly, nodes from the ground truth are assigned a value between $[50, 100]$ for positive instances and $[-100, -50]$ for negative ones. Then we add Gaussian noise to all nodes, where ground truth and non-ground-truth nodes have different mean values set to 10 and



(a) Feature selection (noise) (b) Feature selection (outliers) (c) Instance selection (d) Outlier detection (ROC)

Figure 2: Accuracy of recovering ground truth feature subgraphs (measured as AUC) in synthetic data for (a): varying noise level and (b): varying number of outliers (noise $\sigma = 120$). (c): Utility of selected instances without supervision for classifier learning after annotation (5-fold cross-validation accuracy, noise $\sigma = 100$). (d) Quality of instance selectors as outlier detectors based on inverse ranking of selected instances.

Methods	FS	IS	Netw.	Description
ALFS [24]	✓	✓		Optimal reconstruction
DSLw/o [47]	✓		✓	Sparse self-representation
UFSOL [18]	✓			Locality preservation
RRSS [32]		✓		Subspace learning
UISS	✓	✓	✓	**Our method**

Table II: Summary of competing techniques. FS: feature selection; IS: instance selection; Network: employs the inter-feature network.

70 respectively. We also inject outlier instances in the data for some of our experiments.

Nodes in the *Bike* [2] dataset are bicycle rental stations in Boston and edges connect stations based on a distance threshold. Nodes' values correspond to the number of check-outs in a day, where we employ the last 299 days of the data. We also employ two gene expression datasets: *Liver* metastasis [22] and Embryo [15]. Their network structures are PPI networks [12], while node features correspond to gene expression values with hidden global labels: healthy/normal subjects in Liver and tissue type in Embryo. The *ADNI* [1] dataset contains fMRI resting state measurements for subjects labeled by AD: suffering Alzheimer's disease and NC: healthy normal controls. The graph structure associates functional links (nodes) with their level of coherence (feature values). Nodes are connected if the corresponding functional links share a brain region. *CCT* contains city cellular HTTP traffic data records [8] where nodes are stations, hourly requests are feature values, and node pairs are connected based on a distance threshold. Hidden global labels reflect if the snapshot occurred during workday hours (8am-16pm) or off hours.

B. Experimental setup.

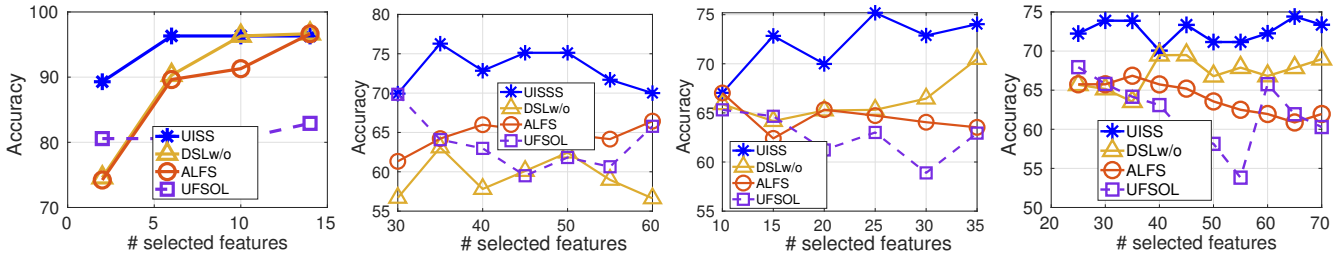
Baselines: All competing techniques are summarized in Tbl. II. We employ several recent feature selection methods to conduct comparison experiments, ALFS [24], employs data reconstruction to find the instances and features that minimize the reconstruction error; DSLw/o is a variant of the state-of-art supervised discriminative subgraph learning method [47] which we restrict to the unsupervised setting by removing the SVM-like supervision from its objective. UFSOL [18] selects features by preserving the local structure in the data. We also employ two recent unsupervised instance selectors: ALFS [24] which selects instances and features for minimum-reconstruction error and RRSS [32] which selects instance based on sparse subspace learning.

Metrics: Following the typical experimental setup in unsupervised feature [18] and instance [24] selection, we first perform these tasks in an unsupervised manner, then reveal the hidden class labels and measure the accuracy of the selected features and/or instances. To allow for fair comparison, we employ the same classifier SVM (linear kernel, $C = 1$) for all methods' selections and report average accuracy of 50 runs.For datasets with ground-truth features (Liver and Synthetic) and synthetically injected outliers, we also measure the ROC of recovering these ground truth elements respectively.

C. Experiments on synthetic network data.

Subnetwork selection: We first evaluate the ability of competing techniques to detect injected ground truth (GT) feature subgraphs in synthetic data. The AUC for ground truth subgraph detection with increasing noise variance is presented in Fig. 2a. While all methods' performance decreases with increasing noise variance, UISS's ability to recover the injected GT subgraphs consistently dominates that of baselines. Though DSLw/o uses the network structure for subnetwork selection, similar to UISS, it does not optimize unsupervised discriminative power and is sensitive to noise and outliers. The comparatively lower accuracy of UFSOL is due to the low robustness to noise of its cluster-based feature selection strategy. Similarly ALFS deteriorates quickly with the noise variance due to its preference for high-variance features dominating its Frobenius reconstruction loss. UISS is less sensitive to noise due to the $L_{2,1}$ reconstruction loss, the use of the network structure among features and the imposed representative power via orthogonality.

We also study the robustness of subnetwork selection for increasing the number of outlier instances and provide comparison in Fig. 2b. UISS dominates and degrades more gracefully with the number of outlier compared to alternatives. This robustness can be attributed to the combination of orthogonal-



(a) Bike (b) Embryo (c) ADNI (d) CCT

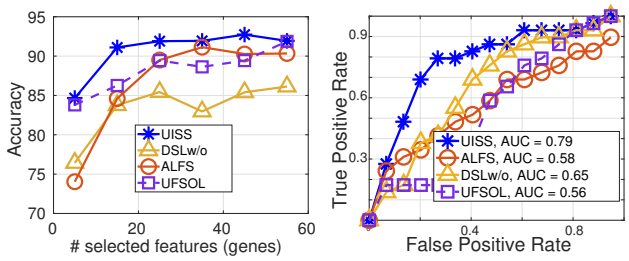
Figure 3: Comparison of classification accuracy on real world datasets with increasing number of selected features.

ity of selected instances for self-representation and the $L_{2,1}$ reconstruction cost, collectively discarding outliers as unable to represent other instances and incurring high reconstruction error. All baseline methods' AUC suffers at all levels of outliers since they treats all instances as cluster members and do not explicitly model the existence of outliers.

Instance selection: Next we evaluate the utility of competing techniques as discriminative instance selectors (Fig. 2c). *Can selected instances inform accurate classifiers if global labels are added post-selection?* Among the instance selector baselines, ALFS can also perform feature selection while RRSS selects representative instances based on all features. Note that while ALFS and UISS internally re-weigh features top select instances, the final classifiers we train in this experiment are based on all features in selected instances to enable a fair comparison with RRSS.

When all training instances are available, the accuracy of competitors is the same as they employ the same SVM classifier and all features. As we constrain the methods to select smaller number of instances for labeling, the baselines' performance degrades, while that of UISS increases slightly and remains above the all-instance accuracy even when we select only 10% of the total instances (Fig. 2c). The advantage of our model stems from the dependence of the instance selector $\mathbf{Q}$ on the feature selector $\mathbf{P}$ which enforces smoothness in the network structure $\mathcal{S}$ as well as the imposed orthogonality on the selected representative instances. In other words, while we do not explicitly exclude features, UISS treats the available features deferentially which in turn informs better instance selection. This advantage is to some extent noticeable for ALFS as well, although its inability to consider the network structure among features leads to significantly degraded accuracy for decreasing number of selected instances.

Outlier detection. We also employ instance selectors to detect outlier instances injected in the dataset (Fig. 2d). We invert the instance selection order reported by each method and plot the ROC for predicting outlier instances (the dataset is balanced, i.e. the same number of outliers and non-outlier instances). UISS is better at detecting outliers than ALFS, since the latter employ Frobenius reconstruction strategies, thus forcing outliers' inclusion among important instances. Although RRSS employs an $L_{2,1}$-norm fitting function, it does not to consider the structure among features when informing the instance



(a) Classification accuracy (b) Ground truth features

Figure 4: Classification accuracy (a) and ROC for selecting ground truth genes (b) in the Liver dataset.

selection. Our choice of $L_{2,1}$ norm for reconstruction and the orthogonality regularization on the instance association matrix allows UISS to "spot" outlier instances better.

D. Experiments on real-world network data

Subnetwork selection: Next we vary the number of features in the selected subnetwork and then add the hidden labels to quantify the cross-validation accuracy (based on selected features in all instances) in the network datasets (Fig. 3). UISS's accuracy dominates all baselines on all datasets and its accuracy is relatively stable around its optimal value. A second pattern that stands out is that there is no clear second best baseline. While DSLw/o dominated other baselines in identifying GT features in synthetic (Figs. 2a), its ability to select predictive features on real datasets does not consistently dominate non-network baselines: better on CCT, ADNI and Bike, but worse on Embryo where features are binary. The best improvement of UISS over baselines in these datasets is about 10% of accuracy and it is retained over different number of selected features.

In addition, we quantify the subgraph selection quality for the Liver dataset that has partial ground truth genes associated with the global disease state labels and reported in [10] (Figs. 4). We observe that the methods which consider the PPI network structure among features (genes), namely UISS and DSLw/o, are better at recovering known genes associated with the disease than alternative which do not consider this structure. In particular the AUC of UISS and DSLw/o are 0.79 and 0.65 respectively, while those of non-network alternatives are close to random (Fig. 4b). Without considering the network

structure, UFSOL and ALFS tend to select isolated nodes rather than a connected subgraph and those are not part of potential target pathways which form connected subgraphs in the PPI network. UFSOL achieves the second best performance after UISS in terms of classification accuracy (4a), however, its AUC for ground truth gene detection is the worst among competitors (Fig. 4b). It is important to note that the set of ground truth genes as discussed in [10] is far from complete, i.e. there are potentially more genes associated with the disease, and thus the competing techniques could be employed to detect more target genes and pathways of biological significance.

Instance selection: Similar to synthetic data, we also compare instance selectors on classification accuracy for increasing number of informative instances selected in an unsupervised manner (Fig. 5). The observed behavior in real-world dataset is similar to that in synthetic data. Namely, UISS dominates alternatives for small number of selected instances and it achieves its optimal accuracy without using all available instances. This is a very promising result for applications in which acquiring annotations is expensive or time consuming, since employing UISS would enable the creation of accurate classifiers with minimal number of instances.

Outlier detection: We also evaluate the ability of competing techniques to detect synthetically injected outliers in real-world datasets in Fig. 6. As in synthetic, we invert the scores for instances to include, i.e. the lowest-ranking instances are deemed highest-scoring outliers. Injected outlier instances have random feature values with similar mean value to instances in the respective real-world dataset. UISS achieves close-to-optimal performance on the Bike and CCT datasets, while ALFS behaves close to random due to its sensitivity to outliers. RRSS is the second best method on the Bike dataset as it explicitly models outliers for instance selection. The behavior of competitors on CCT is also close to random, although in this dataset their initial FPR growth is steeper and later flattens due to ranking some outliers among the most important instances.

E. Joint feature and instance selection.

To enable a fair comparison with non-network baselines in our feature and instance selection experiments from the previous subsections, we sub-selected one of those dimensions and used the other one fully. The performance of UISS improves both when some number of features and a subset of all available instances are excluded. The optimal subsets in both dimensions are likely inter-independent, and hence, in what follows, we study this dependence by an exhaustive sub-selection of both instances and features in the CCT dataset. Note that this dataset is very high dimensional and learning reliable predictive models on it is likely to benefit from both feature and instance selection. The only baseline that performs feature and instance selection jointly is ALFS, and hence, we focus on comparison between UISS and ALFS in this experiment.

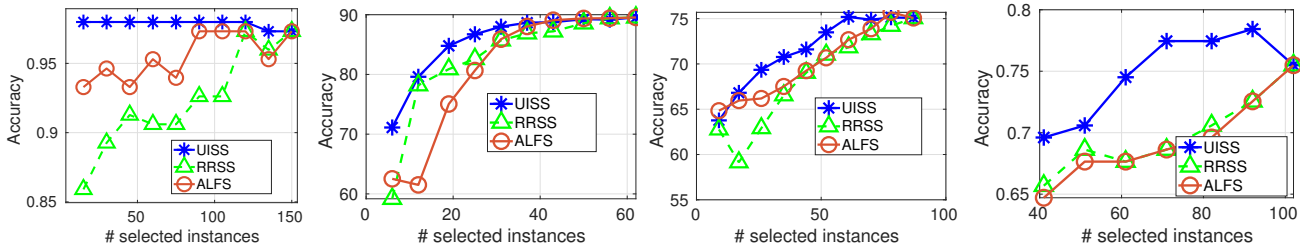
We vary the number of selected features from 10% to 100% (top to bottom in Fig. 6c) and the number of instances between 5% and 50% (left to right in Fig. 6c) in the CCT dataset and present the cross-validation accuracy after revealing the hidden labels for the selected feature subsets by the methods (Fig. 6c). UISS achieves its optimal accuracy (81%) with as few as 30% of the features and 30% of the instances. To get to a similar performance ALFS employs 100% of the instances and 20% of the features, i.e. three times the number of instances and 10% less features. Overall, both methods benefit from the joint selection of features and instances, however, UISS utilizes the network structure and further employs regularization making it more robust to outliers.

F. Performance on non-network data.

UISS is the first unsupervised feature and instance selector for network data. However, one important question about its performance is whether its advantage is only due to the availability of network structure among the data features. To test this, we turn to a common image ("non-network") dataset USPS [4]. This dataset was previously adopted for both feature and instance selection by some of our baselines and also more broadly in the machine learning literature. We compare a non-network version of UISS ($\lambda_3 = 0$) to all baselines for feature and instance selection cross-validation accuracy (Fig. 7), following the same protocol as the one adopted for network data. Interestingly, when employed for instance selection, UISS once again dominates all baselines when limiting the number of instances employed to train a classifier (Fig. 7b). The feature selection experiment renders our method still remains competitive with state-of-art feature selection methods when they are restricted to use between 1% and 10% of the available features (Fig. 7a). While these experiments show promise for the generality of UISS beyond network data, investigation of more non-network datasets is necessary to confirm its utility in such settings. Extensive non-network experiments are beyond the scope and space constraints of the current manuscript and we plan to include such analysis in an extended journal version.

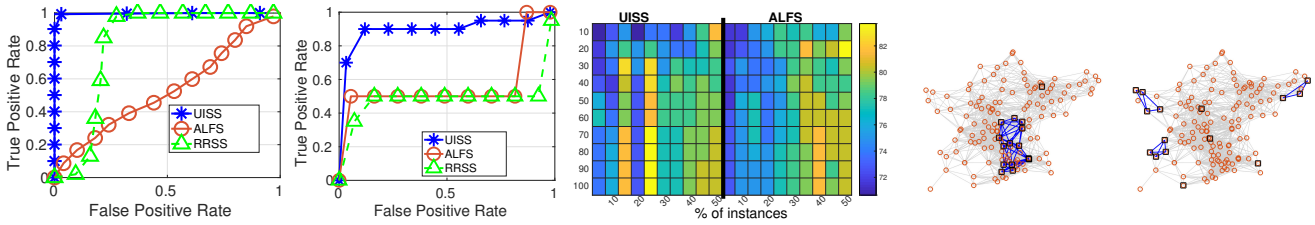
G. UISS at work: salivary gland organoids and bike sharing.

Organoids: We next evaluate the ability of UISS to elucidate the organization of RNA sequencing (RNA-seq) networked samples combined from two different studies of mouse salivary gland development [38]. RNA-seq quantifies the levels of RNAs present in a tissue sample and the Tanaka study [38] employed RNA-seq to characterize the gland organoids derived from embryonic stem cells which are grown ex vivo and engineered to mimic in vivo organ development. Organoids offer rapid disease modeling with applications to infectious diseases from Zika to SARS-CoV-V2.This is a fitting use case for UISS, since while fusing multiple unlabeled, high-dimensional and noisy datasets may enable important new insights, it also warrants examination for outliers as well as artifacts of varying experimental protocols. The organoid data set includes 38 instances: 19 time points (some with



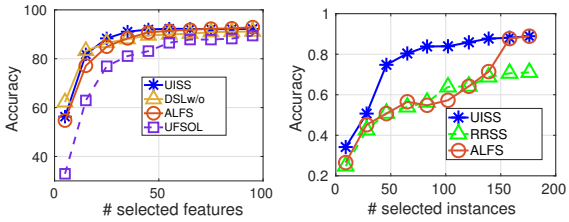
(a) Bike (b) Liver (c) ADNI (d) CCT

Figure 5: Comparison of classification accuracy with increasing number of selected instances on real world datasets.



(a) Outliers (Bike) (b) Outliers (CCT) (c) Joint FS and IS (d) UISS (e) UFSOL

Figure 6: (a),(b): Outlier detection on real-world datasets. (c) Accuracy comparison of joint unsupervised feature and instance selection between UISS (left) and ALFS (right) on the CCT dataset. The number of selected features varies on the vertical axis. UISS achieves over 82% accuracy with as few as 30 features and 30 instances, while ALFS requires all instances to attain the same performances. (d),(e): Visualization of the subgraphs selected by UISS and UFSOL on the Boston Bike dataset.



(a) Feature selection (b) Instance selection

Figure 7: Feature and instance selection accuracy on a non-network dataset USPS [4].

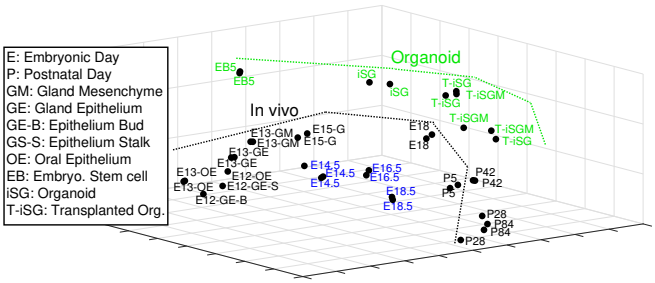


Figure 8: UISS employed to analyze networked RNA-seq samples from salivary gland development. Samples are annotated post-analysis with (i) development time (E12-P84), and (ii) sample types GM/GE/EB/OE/(T)iSG. The embedding learned by UISS reconstructs the developmental timeline and isolates expected stem cell outliers (EB) and differentiates between organoid (iSG) and in vivo development. ($k = 20, \lambda_1 = 0.1, \lambda_2 = 0.1, \lambda_3 = 10^3$).

replicates), spanning from embryonic day 12 (E12) to 12 weeks into adulthood (P84), as well as different cell types (GE/GM/OE) and manipulations (iSG/T-iSG) labeled in Fig. 8. We employed the fold changes in RNA-seq counts as input features after a regularized logarithm transformation [30].

We apply UISS on the data and obtain a 3D embedding of the instances by employing PCA on the learned data projection $\mathbf{XP}$ presented in Fig. 8. It is important to note that the labels in the figure were not used by UISS, i.e. the analysis is fully unsupervised. Embryonic stem cell derived embryoid bodies (EBs) are pluripotent and able to become all known cell types. Not surprisingly, UISS renders them based on their RNA-seq as outliers compared to other samples. After the EBs are induced to differentiate into salivary glands (iSG) in vitro their transcriptional landscape changes to become more similar to developing glands. Using the time course labels (E12 through P84), we can place experimental data in frame with the known developmental (in vivo) progression (black dotted line). Organoids (iSG) are transcriptionally similar to an early salivary gland differentiation stage around E15-16, however, once upon transplantation in vivo (samples T-iSG) the organoid overcome a developmental wall furthering their maturation. The T-iSG transcriptomes should be comparable to postnatal salivary glands from five days after birth (P5) to 84 days after birth (P84) from the Gluck study [16].

Interestingly, we detect a wider range of variability in maturation of the transplanted samples, which could represent either

variability in the transplanted organoids or differential responses of the individual organoids to the in vivo environment. Although the representation by UISS cannot readily explain the variable gene expression by the organoids, this example of staging organoids illustrates its ability to map relationships between biological datasets. As RNA-seq is broadly used in biomedical sciences, UISS will have broader applicability in comparison of biological datasets. It is important to note that, PCA on the UISS' embedding of the data $\mathbf{XP}$ retains over 99% of the variance in the first component, while PCA on the raw data X requires more than 12 components for 99% variance retention.

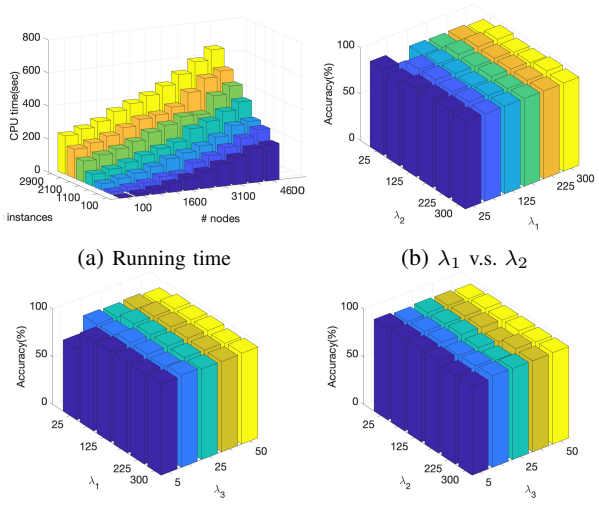
Bike sharing: We also investigate the interpretability of the feature selection by UISS in the Bike dataset. Figs. 6d, 6e visualize the selected subgraphs (highlighted nodes and edges) by UISS and UFSOL—its closest competitor in terms of accuracy when employing at least 6 features (see Fig. 3a). As expected, UISS selects a mostly connected feature subgraph due to the network smoothness regularization. This subgraph corresponds to a locality of bike rental stations in downtown Boston which can differentiate between weekday and weekend traffic (hidden classes in this dataset). UFSOL selects features which form multiple connected components which are less obvious to interpret and more importantly result in less predictive classifiers (as demonstrated in Fig. 3a).

H. Parameter sensitivity and running time

We evaluate UISS's scalability and sensitivity to various parameters. The experiment shows that our method's running time grows nearly linearly for both increasing number of instances and nodes. Additionally, we also investigate the sensitivity of UISS to its three hyper-parameters, i.e. λ_1, λ_2 and λ_3. All combinations of hyperparamters exhibit similar robust behavior. Please find more details at the supplement material.

We also investigate the sensitivity of UISS to its three hyper-parameters: λ_1, λ_2 and λ_3, which control the number of selected subnetwork, instances, and level of graphs smoothness in feature selection respectively. These parameters all have a physical meaning and can provide useful control of the end user who want to enforce each of the corresponding behaviors based expert knowledge about the domain in which she employs UISS. However, in some cases the optimal parameter setting may be hard to obtain apriori, a typical challenge with many unsupervised methods. Hence we analyze the accuracy stability under small variations of these parameters.

Based on our analysis presented in Fig. 9 the quality of UISS does not vary significantly with the parameter settings. For this analysis, we measure the performance in terms of feature selection accuracy by fixing one parameter (λ_3) and updating the other two. The overall performance on different combinations is stable and close to peak accuracy over large parameter ranges. The performance drops when at one of the parameter becomes significantly smaller (order of magnitude) than the other two. The reason for this behavior is that the three regularization terms in our objective have similar



(a) Running time (b) λ_1 v.s. λ_2

(c) λ_1 v.s. λ_3 (d) λ_2 v.s. λ_3

Figure 9: Parameter sensitivity and running time on the synthetic dataset.

contribution to the cost, and thus keeping their importance balanced results in optimal performance. In addition, small values of λ_1 and λ_2 will not enforce sufficient sparsity for the representative instance and feature selection and lead to poor performance due to negative effect of outliers and noise. Other combinations of hyperparamters exhibit similar robust behavior (in supplement material). Also, by default we set $k = 10$, but our analysis does not render UISS sensitive to third internal dimension of the self-representative factorization.

VII. Conclusion

In this paper we proposed and evaluated UISS: a general unsupervised approach for joint subnetwork and instance selection in network data. It performs interpretable subnetwork and instance selections via a self-representative factorization which enforces smoothness on the inter-feature network and robustness to outlier instances and noise. UISS dominated alternatives in both instance and subnetwork selection, often achieving its highest accuracy using significantly fewer features and instances than those employed by corresponding baselines. Meanwhile, our method is able to discover interpretable subnetwork in the data such as city center localities of bike rental behavior distinguishing between workdays and weekends.

VIII. Acknowledgement

The work is supported by the NSF Smart and Connected Communities (SC&C) grant 1831547 and by NIH award R01DE027953.

References

[1] ADNI project. http://www.adni-info.org/.
[2] Hubway data visualization challenge:. http://hubwaydatachallenge.org.
[3] V. Amelkin, P. Bogdanov, and A. K. Singh. A distance measure for the analysis of polar opinion dynamics in social networks. In *Proc. of the Intl. Conference on Data Engineering (ICDE)*. IEEE, 2017.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. 1997.

[5] P. Bogdanov et al. Mining heavy subgraphs in time-evolving networks. In *Proc. of the IEEE Intl. Conference on Data Mining*, ICDM '11, 2011.

[6] P. Bogdanov and other. Learning about learning: Mining human brain sub-network biomarkers from fMRI data. *PLoS One*, 2017.

[7] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *Proc. of ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD '10, 2010.

[8] X. Chen et al. Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale. In *Communications (ICC), 2015 IEEE Intl. Conference on*, 2015.

[9] K. Cheng, J. Li, and H. Liu. Unsupervised feature selection in signed social networks. In *Proc. of the SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD '17, 2017.

[10] X. H. Dang et al. Learning predictive substructures with regularization for network data. In *2015 IEEE Intl. Conference on Data Mining*, pages 81–90, Nov 2015.

[11] X. H. Dang, A. K. Singh, et al. Discriminative subnetworks with regularized spectral learning for global-state network data. In *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 8724*, ECML PKDD 2014, pages 290–306, 2014.

[12] R. Dannenfelser, N. R. Clark, and A. Ma'ayan. Genes2fans: connecting genes through functional association networks. *BMC Bioinformatics*, 13(1):156, Jul 2012.

[13] C. Ding et al. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proc. of the 12th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 126–135, New York, NY, USA, 2006. ACM.

[14] L. Du and Y.-D. Shen. Unsupervised feature selection with adaptive structure learning. In *Proc. of the 21th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 209–218, New York, NY, USA, 2015. ACM.

[15] J. Dutkowski and T. Ideker. Protein networks as logic functions in development and cancer. *PLoS comp. biology*, 7(9):e1002180, 2011.

[16] C. Gluck, S. Min, et al. Rna-seq based transcriptomic map reveals new insights into mouse salivary gland development and maturation. *BMC genomics*, 17(1):923, 2016.

[17] Q. Gu, Z. Li, and J. Han. Joint feature selection and subspace learning. In *Proc. of the Intl. Joint Conf. on Artificial Intelligence*, IJCAI'11, 2011.

[18] J. Guo, Y. Quo, X. Kong, and R. He. Unsupervised feature selection with ordinal locality. In *IEEE Intl. Conference on Multimedia and Expo (ICME)*, pages 1213–1218, July 2017.

[19] B.-S. He, L.-Z. Liao, and H. Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92:103–118, 01 2002.

[20] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Proc. of the Intl. Conference on Neural Information Processing Systems*, NIPS'05, pages 507–514, Cambridge, MA, USA, 2005. MIT Press.

[21] C. Jiang, F. Coenen, and M. Zito. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(1):75–105, 2013.

[22] D. H. Ki et al. Whole genome analysis for liver metastasis gene signatures in colorectal cancer. 121:2005–12, 11 2007.

[23] X. Kong et al. Dual active feature and sample selection for graph classification. In *Proc. of the 17th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 654–662, New York, NY, USA, 2011. ACM.

[24] C. Li, X. Wang, W. Dong, J. Yan, Q. Liu, and H. Zha. Joint active learning with feature selection via cur matrix decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2018.

[25] J. Li, R. Guo, C. Liu, and H. Liu. Adaptive unsupervised feature selection on attributed networks. KDD '19, New York, NY, USA, 2019. Association for Computing Machinery.

[26] J. Li, X. Hu, J. Tang, and H. Liu. Unsupervised streaming feature selection in social media. CIKM '15, New York, NY, USA, 2015. Association for Computing Machinery.

[27] J. Li, X. Hu, L. Wu, and H. Liu. *Robust Unsupervised Feature Selection on Networked Data*, pages 387–395.

[28] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2012.

[29] H. Liu and S. Yan. Robust graph mode seeking by graph shift. In *Proc. of the 27th Intl. Conference on Intl. Conference on Machine Learning*, ICML'10, pages 671–678, USA, 2010. Omnipress.

[30] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[31] M. W. Mahoney and P. Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[32] F. Nie, H. Wang, H. Huang, and C. H. Q. Ding. Early active learning via robust representation and structured sparsity. In *IJCAI*, pages 1572–1578. IJCAI/AAAI, 2013.

[33] M. Qian and C. Zhai. Robust unsupervised feature selection. In *Proc. of the Intl. Joint Conf. on Artificial Intelligence*, IJCAI '13, 2013.

[34] S. Ranu, M. Hoang, and A. Singh. Mining discriminative subgraphs from global-state networks. In *Proc. of the 19th ACM SIGKDD intl. conference on Knowledge discovery and data mining*, pages 509–517. ACM, 2013.

[35] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[36] B. Settles. Active learning literature survey. Technical report, 2010.

[37] A. Silva, P. Bogdanov, and A. Singh. Hierarchical in-network attribute compression via importance sampling. In *Proc. of the 31st IEEE Intl. Conference on Data Engineering (ICDE)*, 2015.

[38] J. Tanaka, M. Ogawa, , et al. Generation of orthotopically functional salivary gland from embryonic stem cells. *Nature communications*, 9(1):1–13, 2018.

[39] J. Tang and H. Liu. CoSelect: feature selection with instance selection for social media data. In *Proc. of the SIAM Intl. Conference on Data Mining*, 2013.

[40] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, Mar. 2002.

[41] T. Viklands. Algorithms for the weighted orthogonal procrustes problem and other least squares problems. 2006.

[42] S. Wang, J. Tang, and H. Liu. Embedded unsupervised feature selection. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2015.

[43] X. Wei, B. Cao, and P. S. Yu. Unsupervised feature selection on networks: A generative view. AAAI'16, page 2215–2221. AAAI Press, 2016.

[44] X. Wei, S. Xie, and P. S. Yu. *Efficient Partial Order Preserving Unsupervised Feature Selection on Networks*, pages 82–90.

[45] Y. Yang et al. L2,1-norm regularized discriminative feature selection for unsupervised learning. In *Proc. of the Intl. Joint Conference on Artificial Intelligence*, 2011.

[46] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Proc. of the Intl. Conference on Machine Learning*, ICML '06, pages 1081–1088, New York, NY, USA, 2006. ACM.

[47] L. Zhang and P. Bogdanov. DSL: discriminative subgraph learning via sparse self-representation. In *Proc. of the SIAM Intl. Conference on Data Mining (SDM)*, 2019.