

# Representation Learning for Spatial Graphs

Zheng Wang\*, Ce Ju<sup>†</sup>, Gao Cong\*, Cheng Long\*

\*School of Computer Science and Engineering, Nanyang Technological University, Singapore

<sup>†</sup>IDG, Baidu Inc., Beijing, China

{wang\_zheng, gaocong, c.long}@ntu.edu.sg, juce@baidu.com

**Abstract**—Recently, the topic of graph representation learning has received plenty of attention. Existing approaches usually focus on structural properties only and thus they are not sufficient for those spatial graphs where the nodes are associated with some spatial information. In this paper, we present the first deep learning approach called **s2vec** for learning spatial graph representations, which is based on denoising autoencoders framework (DAF). We evaluate the learned representations on real datasets and the results verified the effectiveness of **s2vec** when used for spatial clustering.

## I. INTRODUCTION

Spatial graphs are graphs where each node is associated with a location [1]. One well-known instance of spatial graph is the location-based social networks (LBSN) [2]–[4], where the users in social networks, which correspond to nodes, are associated with some location information, e.g., online “check-ins” in Brightkite, geo-tagged tweets in Twitter, geo-tagged photo albums in Foursquare. Some other instances of spatial graph are commonly used in ecology and evolution [1].

We consider the problem of spatial subgraph embedding based on deep representation learning. Once the embedding vectors of the graphs are computed, they could be used for a variety of downstream graph analysis tasks, including spatial clustering [2], outlier detection [4] and spatial classification. For instance, in social network analysis, we often gained subgraph-structured communities by performing community search or detection algorithms. Clustering these communities are very beneficial for some practical applications such as social marketing, as studied in [5], people with close social relationships tend to purchase in places that are also physically close. Advertisers can target limited resources to the communities with similar spatial and structural characteristics to boost sales figures and achieve maximize revenue.

Existing graph representation learning methods can be categorized into node representation [6]–[9] and subgraph representations [10]–[13]. However, almost all these prior works focus on structural analysis and do not consider the spatial features. Some alternative approaches for spatial graph similarity search include graph kernels [14] and graph matching [15], which, however cannot scale-up for large graphs due to their high computational cost. Additionally, these approaches do not consider the global structure of the graphs and are sensitive to noise.

In this paper, we propose an unsupervised spatial graph to **vector** approach called **s2vec** for learning representations of spatial subgraphs based on the LSTM denoising autoencoders

framework (DAF). Specifically, we first sample a set of paths on each subgraph via random walks. Inspired by [16], a spatial information aware loss function is proposed based on negative log likelihood, which captures the similarity based on spatial proximity. Finally, we demonstrate the effectiveness of our approach by comparing our learned representations with those based some baselines when used for a spatial clustering task. To the best of our knowledge, **s2vec** is the first method that supports spatial subgraph representation, making it widely applicable to various downstream graph analysis tasks.

## II. RELATED WORK

Our work is closely related to the literature of representation learning for graphs. The traditional methods in this field represent a graph as an associated matrix (e.g., adjacency matrix) or a collection of nodes and edges. In recent years, inspired by the success of word2vec [17], modern learning methods attempt to embed nodes into high-dimensional vectors in a continuous space so that nodes with similar representation vectors share similar structural properties such as DeepWalk [6], node2vec [7], LINE [8] and DeepCas [9].

Another line of related work comes from the representation of subgraph structures [10]–[13]. Many of the approaches are inspired by the huge success of representation learning and deep neural networks applied to various domains. For example, subgraph2vec [11] borrows the key idea from document embedding methods. The biggest difference between our method and the above works is that these model structures are designed without considering the spatial information and can not be directly used for spatial graph embedding.

In addition, some new graph kernels such as Weisfeiler-Lehman subtree kernel (WL) [14] and Deep Graph Kernel (DGK) [12] have been proposed to characterize the similarity of different network structures. Others are motivated by representation learning of images to embed graph using convolutional neural networks (CNNs) [10], [18].

## III. DEFINITIONS AND PRELIMINARIES

In this section, we present definitions and preliminaries necessary to understand the problem solved and the denoising autoencoders model used in our solution.

### A. Definitions

**Definition 1 (Spatial Graph):** Let  $G = (V, E)$  be a graph, where  $V$  denotes its vertex set and  $E \subseteq (V \times V)$  denotes its edge set.  $G$  is called a spatial graph if each vertex  $v \in V$

corresponds to a tuple  $(id, x, y)$ , where  $id$  is the identification number of  $v$  and  $(x, y)$  is the location of  $v$ .

**Definition 2 (Gromov-Hausdorff Distance):** Given two spatial graphs  $G_1 = (V_1, E_1)$  and  $G_2 = (V_2, E_2)$ , the one-sided graph distance from  $G_1$  to  $G_2$  is defined as:

$$\tilde{d}_H(G_1, G_2) := \max_{v_1 \in V_1} \min_{v_2 \in V_2} d(v_1, v_2),$$

where  $d(\cdot, \cdot)$  is the distance on Euclidean space. The bidirectional graph distance between  $G_1$  and  $G_2$  is defined as

$$d_H(G_1, G_2) := \max(\tilde{d}_H(G_1, G_2), \tilde{d}_H(G_2, G_1))$$

is then called Gromov-Hausdorff distance of spatial graph.

### B. Problem statement

Given a set of spatial subgraphs  $\mathbb{G} = \{G_1, G_2, \dots\}$ , we aim to compute an embedding vector of each subgraph such that if two subgraphs are spatial similar w.r.t. Gromov-Hausdorff distance, then they are also similar w.r.t. the Euclidean distance based on their embedding vectors. Specifically, given small enough  $\delta, \epsilon > 0$  and for any spatial subgraph  $G_i$  and  $G_j$ ,

$$\text{if } d_H(G_i, G_j) \leq \delta, \text{ we have } d(\phi(G_i), \phi(G_j)) \leq \epsilon,$$

where  $d_H(\cdot, \cdot)$  is Gromov-Hausdorff distance,  $d(\cdot, \cdot)$  is Euclidean distance and function  $\phi(\cdot)$  is our embedding algorithm. Suppose  $n$  is the embedding dimension, the matrix of representations of the subgraph sets  $\mathbb{G}$  is denoted as  $\Phi \in \mathbb{R}^{|\mathbb{G}| \times n}$  in the following sections. Note that we did not preserve the structures of these subgraphs and assume they are similar. For example, they are all the same  $k$ -core subgraphs obtained by community search algorithms,

### C. Preliminaries of denoising autoencoders framework

We briefly introduce the denoising autoencoders framework, which is proposed by Vincent et al. [19]. The principle behind it is to be able to reconstruct data from an input of corrupted or noisy data. In order to force the hidden layer to discover more robust features and prevent it from simply learning the identity, and thus the output will be a more refined version of the input data. Precisely, the denoising autoencoders framework has the following steps, as shown in Figure 1.

- **Distribution:** Corrupt the initial input  $x$  by sampling from a conditional distribution  $q_{\mathcal{D}}(\tilde{x}|x)$  which stochastically maps input data  $x$  to a corrupted version  $\tilde{x}$ .
- **Encoder:** Map the corrupted input  $\tilde{x}$  to a latent representation  $h := f_{\theta}(\tilde{x})$ .
- **Decoder:** Reconstruct a data  $z := g_{\theta^*}(h)$  from the latent representation  $h$ .
- **Loss:** Compute the average reconstruction error  $\mathcal{L}(x, z)$  over a training set.

The objective function of denoising autoencoders framework is

$$\arg \min_{\theta^*, \theta} \mathbb{E}_{q^0(x, \tilde{x})} \mathcal{L}(x, z)$$

where  $z = g_{\theta^*}(f_{\theta}(\tilde{x}))$  and  $q^0(x, \tilde{x})$  denotes the empirical joint distribution associated to our training inputs. Minimizing the

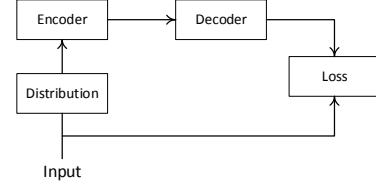


Fig. 1. Denoising Autoencoders Model

objective function by stochastic gradient descent method, the denoising autoencoders framework reconstruct the uncorrupted data from the corrupted one.

## IV. METHOD

Our proposed method for solving this problem is inspired by the denoising autoencoders framework and techniques of Seq2Seq and the spatial proximity computation.

### A. Random walk sampling

In this subsection, we will introduce our path sampling method by which we sample a group of paths on each subgraph  $G \in \mathbb{G}$ . The path sampling method is simplified from a popular graph sampling method called Random Walk with Jump [9]. The reason for the simplification of our sampling is in order to sample a set of connected paths. We walk from a starting node  $v_{t:=start}$  picked uniformly from the whole subgraph. In the next step, we have a given probability  $\mathbb{P}$  to walk to one of the adjacent node selected uniformly at random. The uniform selection means in the following way that supposes the degree of the current node  $v$  is  $deg(v)$ , then we have the probability transition from the current node  $v$  to its adjacent node  $u$  is

$$p(u \in N(v)|v) = \frac{deg(u) + \eta}{\sum_{s \in N(v)} (deg(s) + \eta)},$$

where  $N(v)$  denotes the set of  $v$ 's outgoing neighbors.  $\eta$  is a smoother. Otherwise, we need to stay at the current node with probability  $1 - \mathbb{P}$ . The walk stops at the end node  $v_{t:=end}$  after enough number of nodes are visited. According to our construction of the path, each path of the subgraph could be represented as  $\mathcal{T}_G := v_{t:=start} \cdots v_{t:=end}$ , which is also regarded as a user visiting order-spatial sequence. In the following sections, for each subgraph  $G \in \mathbb{G}$ ,  $set(\mathcal{T}_G)$  corresponds to the set containing all sampling paths on the subgraph  $G$ .

### B. The latent representation of subgraph

For each subgraph  $G \in \mathbb{G}$ , we construct a corrupted subgraph  $\tilde{G}$  which is added random noise  $\delta \cdot \mathcal{N}(0, 1)$  on each node of the initial subgraph  $G \in \mathbb{G}$ , where  $\delta > 0$  is the magnitude of noise and  $\mathcal{N}(0, 1)$  is the standard normal distribution. Thus, the Gromov-Hausdorff distance between the initial subgraph  $G$  and its corresponding corrupted subgraph  $\tilde{G}$  is bounded by the magnitude  $\delta$ . This can be proved directly by definition. Also, we construct the corrupted path set  $set(\mathcal{T}_{\tilde{G}})$  of  $\tilde{G}$ . Each path  $\mathcal{T}_{\tilde{G}}$  of  $set(\mathcal{T}_{\tilde{G}})$  only differs the Gaussian noisy displacement on each node of the initial path  $\mathcal{T}_G$  of  $G$ .

In the sequence encoder-decoder model, we need to input the discrete token into the model and thus we partition the spatial space into cells of equal size. For each node on the path, if it falls into one of the cells and then we provide this node a token. We only keep the cells which are sufficiently hit by sample points. These cells are referred to as hot cells  $C$ . The sequence of token of the path  $\mathcal{T}_G$  is denoted as  $\mathcal{S}_G = \{s_1, s_2, \dots\}$ . The sequence set is denoted as  $set(\mathcal{S}_G)$ .

According to the denoising autoencoders framework, we pick LSTM as the encoding function  $f_\theta(\cdot) := LSTM_\theta^{enc}(\cdot)$ . Then, the time-varying hidden vector  $h_t^{enc}$  is

$$h_t^{enc} := LSTM_\theta^{enc}(\tilde{s}_t, h_{t-1}^{enc}),$$

where  $\tilde{s}_t$  means the corrupted version at timestamp  $t$ . We let another LSTM as the decoding function  $g_{\theta^*}(\cdot) := LSTM_{\theta^*}^{dec}(\cdot)$ , then

$$h_t^{dec} := LSTM_{\theta^*}^{dec}(s_{t-1}, h_{t-1}^{dec}),$$

the objective function of our algorithm is inspired by the classic negative log likelihood loss [20],

$$\arg \min_{\theta, \theta^*} \mathbb{E}_{q^0(\mathcal{S}_G, \mathcal{S}_{\tilde{G}})} \mathcal{L}(\mathcal{S}_G, g_{\theta^*}(f_\theta(\mathcal{S}_{\tilde{G}}))).$$

The reconstruction error in our loss function is specifically defined as the spatial proximity aware loss [16],

$$\mathcal{L} := - \sum_{t=1}^{|\mathcal{S}_G|} \sum_{u \in C} w_{ust} \log \frac{\exp W_u^T \cdot h_t^{dec}}{\sum_{v \in C} \exp(W_v^T \cdot h_t^{dec})},$$

where the coefficients of the polynomial

$$w_{ust} = \frac{\exp(-||u - s_t||_2)}{\sum_{v \in C} \exp(-||v - s_t||_2)}$$

is the spatial proximity weight on cell  $v \in C$  when decoding target  $s_t$ .  $W_u^T$  is the projection matrix for cell  $u$  that projects  $h_t$  from the hidden state spaces into the token space. And,  $||u - s_t||_2$  denotes the Euclidean distance between the centroid coordinates of the cells.

Eventually, the graph embedding function  $\phi(G)$  is defined as the average of the latent representation of the sampling path for each subgraph  $G$ ,

$$\phi(G) := \frac{1}{|set(\mathcal{S}_G)|} \sum_{\mathcal{S}_G \in set(\mathcal{S}_G)} f_\theta(\mathcal{S}_G).$$

Thus, our desired subgraph matrix representation is

$$\Phi := (\phi(G_1), \dots, \phi(G_{|\mathbb{G}|})) \in \mathbb{R}^{|\mathbb{G}| \times n},$$

where  $n$  is the embedding dimension.

### C. Algorithm overview

Algorithm 1 presents the framework. The input includes a set of spatial subgraphs  $\mathbb{G} = \{G_1, G_2, \dots\}$ , the embedding size  $n$ , learning model  $M_\theta$  with encoding function  $f_\theta$  and decoding function  $g_\theta$ , where  $\theta$  is the global parameters and learning rate  $\alpha$  (line 1-4). During the iterative training process (line 5-11), we first sample a set of paths of vertices  $set(\mathcal{T}_G)$  from the set

of subgraphs (line 6). Similar to DeepWalk [6], the sampling process could be generalized as performing a random walk. Then, we stochastically map the original input  $set(\mathcal{T}_G)$  to a noisy version  $set(\mathcal{T}_{\tilde{G}})$  by adding a Gaussian noise which is subject to the conditional distribution  $q_D$  (line 7) and perform the tokenization to get  $set(\mathcal{S}_G)$  and  $set(\mathcal{S}_{\tilde{G}})$  respectively (line 8). Next, we get the reconstruction  $\mathcal{S}_{\tilde{G}}$ , which is computed by decoder component of the model  $M_\theta$  and an optimizer such as stochastic gradient descent (SGD) is used to optimize the parameters  $\theta$  (line 9-10). Finally, a vector representation of each subgraph is computed via the learned model  $M_\theta$  (line 12-15). Note that the ultimate output  $\Phi$  is a matrix of vector representations of spatial subgraphs.

### Algorithm 1 Algorithm framework

**Input:**

- 1:  $\mathbb{G} = \{G_1, G_2, \dots\}$ : a set of spatial graphs need to be embedded
- 2:  $n$ : the embedding size
- 3: Learning model  $M_\theta$  with encoder  $f_\theta$  and decoder  $g_\theta$
- 4:  $\alpha$ : learning rate
- 5: **repeat**
- 6:   sample  $set(\mathcal{T}_G)$  by random walking from the set of subgraphs
- 7:   get noisy version  $set(\mathcal{T}_{\tilde{G}}) \sim q_D(set(\mathcal{T}_G) | set(\mathcal{T}_{\tilde{G}}))$
- 8:   get  $set(\mathcal{S}_G)$  and  $set(\mathcal{S}_{\tilde{G}})$  from  $set(\mathcal{T}_G)$  and  $set(\mathcal{T}_{\tilde{G}})$
- 9:   compute reconstruction  $\mathcal{S}_{\tilde{G}} = g_\theta(f_\theta(\mathcal{S}_{\tilde{G}}))$  from  $M_\theta$
- 10:    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\mathcal{S}_G, \mathcal{S}_{\tilde{G}})$
- 11: **until** No improvement on validation set
- 12: **for** each  $G_i \in \mathbb{G}$  **do**
- 13:   get representation  $\phi(G_i)$  from  $M_\theta$
- 14:    $\Phi \leftarrow \phi(G_i)$
- 15: **end for**

**Output:** Matrix of vector representations of spatial graphs  $\Phi \in \mathbb{R}^{|\mathbb{G}| \times n}$

## V. EXPERIMENTS

In this section, we use our representations for the task of spatial graph clustering on four real datasets and show the effectiveness of our model s2vec compared against several baseline approaches.

### A. Datasets

The experiments are conducted on four real datasets: Brightkite<sup>1</sup>, Gowalla<sup>1</sup>, Flickr<sup>2</sup> and Foursquare<sup>3</sup>. For all the datasets, each vertex represents a user and each link represents the friendship between two users. The statistics of each dataset are shown in Table I, where the average degree  $\hat{d}$ , the average betweenness centrality **BC** and the clustering coefficient **CC** of each vertex in the dataset are included. We randomly select 200 subgraphs of each dataset among those found by a community search algorithm [4].

TABLE I  
DATASET STATISTICS.

Type	Name	Vertices	Edges	$\hat{d}$	BC	CC
Real	Brightkite	51,406	197,167	7.67	7.34E-5	0.1795
	Gowalla	107,092	456,830	8.53	3.40E-5	0.2487
	Flickr	214,698	2,096,306	19.5	1.65E-5	0.1113
	Foursquare	2,127,093	8,640,352	8.12	1.68E-6	0.1044

<sup>1</sup><http://snap.stanford.edu/data/index.html>

<sup>2</sup><https://www.flickr.com/>

<sup>3</sup>[https://archive.org/details/201309\\_foursquare\\_dataset\\_umh](https://archive.org/details/201309_foursquare_dataset_umh)

## B. Baselines

We compare s2vec with two baselines of graph representation, the first baseline is similar to Principal Component Analysis (PCA) [21], we arrange the spatial features of each vertex in the subgraph according to the order of longitude and latitude and then expand into a high-dimensional vector. In this case, vectors of inconsistent length are padded by a special value such as zero. After that, we perform PCA dimensionality reduction on all subgraphs to obtain new vector representations. Note that this algorithm considers the graph as a set of spatial vertices without structural features.

The second baseline is a kind of ablation of s2vec, adapting the sequence-to-sequence learning framework for autoencoding by using the same sequence for both the input and output [?], [13]. We denote these two baselines as **PCA** and **Vanilla-s2vec (V-s2vec)** respectively.

## C. Spatial Graph Clustering

Given a set of spatial graphs  $\mathbb{G}$ , the goal of graph clustering is to group graphs with similar structure and spatial features together. s2vec's representation vectors could be used along with conventional clustering algorithms such as DBSCAN with Euclidean distance for this purpose.

The lack of ground-truth makes it a challenging problem to evaluate the effectiveness of spatial graph clustering. To overcome it, we use the Gromov-Hausdorff distance to measure the similarity of spatial graphs and perform DBSCAN algorithm to find the results as its ground-truth.

**Evaluation Metric.** In order to quantitatively measure the clustering accuracy, a standard clustering evaluation metric, namely, Adjusted Rand Index (ARI) [22] is used. The ARI values lie in the range  $[-1, 1]$ . A higher ARI means a higher correspondence to the ground-truth results. We convert it into a percentage value for easy understanding.

The results of spatial graph clustering using s2vec and other baselines are presented in Table II. We observe that s2vec outperforms all the other compared approaches highly significantly. In particular, it outperforms the PCA techniques by at least 10%. This is because s2vec fully considers maintaining similarity in original space.

TABLE II  
SPATIAL GRAPH CLUSTERING.

Dataset	Brightkite	Gowalla	Flickr	Foursquare
PCA	49.76	49.45	49.99	48.85
V-s2vec	57.14	59.81	69.59	54.60
s2vec	<b>62.49</b>	<b>63.07</b>	<b>73.64</b>	<b>63.61</b>

## VI. CONCLUSION AND FUTURE WORK

In this paper, we learn the latent representation of the spatial subgraphs in the denoising autoencoders framework and apply our latent representations in the spatial graph clustering task achieving the effective results. The problem of spatial graph representation is the first time to be proposed and has a great potential application value. In the future, we will consider a more complex graph structure with more information on its nodes.

## REFERENCES

- [1] M. Dale and M.-J. Fortin, "From graphs to spatial graphs," *Annual Review of Ecology, Evolution, and Systematics*, vol. 41, pp. 21–38, 2010.
- [2] J. Shi, N. Mamoulis, D. Wu, and D. W. Cheung, "Density-based place clustering in geo-social networks," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 2014, pp. 99–110.
- [3] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3–5, pp. 75–174, 2010.
- [4] Y. Fang, Z. Wang, R. Cheng, X. Li, S. Luo, J. Hu, and X. Chen, "On spatial-aware community search," *IEEE Transactions on Knowledge and Data Engineering*, 2018.
- [5] P. Manchanda, G. Packard, and A. Pattabhiramaiah, "Social dollars: The economic impact of customer participation in a firm-sponsored online customer community," *Marketing Science*, vol. 34, no. 3, pp. 367–387, 2015.
- [6] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.
- [7] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.
- [8] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.
- [9] C. Li, J. Ma, X. Guo, and Q. Mei, "Deepcas: An end-to-end predictor of information cascades," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 577–586.
- [10] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International conference on machine learning*, 2016, pp. 2014–2023.
- [11] A. Narayanan, M. Chandramohan, L. Chen, Y. Liu, and S. Saminathan, "subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs," *arXiv preprint arXiv:1606.08928*, 2016.
- [12] P. Yanardag and S. Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1365–1374.
- [13] A. Taheri, "Learning graph representations with recurrent neural network autoencoders," 2018.
- [14] C. Tan, L. Lee, and B. Pang, "The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter," *arXiv preprint arXiv:1405.1438*, 2014.
- [15] M. Aanjaneya, F. Chazal, D. Chen, M. Glisse, L. Guibas, and D. Morozov, "Metric graph reconstruction from noisy data," *International Journal of Computational Geometry & Applications*, vol. 22, no. 04, pp. 305–325, 2012.
- [16] X. Li, K. Zhao, G. Cong, C. S. Jensen, and W. Wei, "Deep representation learning for trajectory similarity computation," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 617–628.
- [17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [18] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.
- [19] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.
- [20] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
- [21] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1–3, pp. 37–52, 1987.
- [22] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.