DeepKmeans: Integrating k-means and neural network to decipher spatial domains from and spatially variable genes

A Preprint

Ke Wang *
School of Artificial Intelligence
Nanjing University
201300008@smail.nju.edu.cn

January 29, 2023

Abstract

The development of single-cell sequencing technology has provided new tools for studying genome and cell diversity. However, due to the large number of cells and high-dimensional, dense data with a lot of noise, cluster analysis of this data remains a challenging task. The original k-means algorithm has the disadvantage of not adapting to high-dimensional data and being highly sensitive to the initial clustering center. To address these problems, we propose a clustering method based on VAE dimensionality reduction and Louvain algorithm initialization, which incorporates biological knowledge. The Louvain algorithm can use cell-to-cell similarity information to initialize the clustering center, while VAE can effectively reduce the dimensionality of the data and reduce noise interference, resulting in more accurate clustering results. This paper will detail the implementation of this method and demonstrate its effectiveness through experiments.

Keywords clustering · single-cell sequencing technology · initialization

1 Introduction

Single-cell sequencing technology is a rapidly emerging technology in recent years, which can sequence the genome at the single cell level and provide new tools for studying genome and cell diversity. However, due to the large number of cells, high-dimensional data density and a lot of noise, clustering analysis of it is still a challenging task.

The development of single-cell sequencing technology has provided new tools for studying genome and cell diversity. These technologies allow us to study the gene expression levels and cell states of individual cells and to carry out more detailed analysis of differences between cells. However, due to the large number of cells, high-dimensional data density, and a lot of noise, clustering analysis is still a challenging task. The original k-means algorithm has the inadequacy of not being able to adapt to high-dimensional data and being highly sensitive to the initial cluster centers, which limits its application in single-cell sequencing data analysis.

To address these problems, we propose a clustering method based on VAE dimensionality reduction and Louvain algorithm initialization, combining biological knowledge.VAE (variational autoencoder) is a generative model that can effectively reduce the dimensionality of data, reduce noise interference, and make the clustering results more accurate. The Louvain algorithm can use the similarity information between cells to initialize the clustering centers. This method can effectively avoid the problems of traditional k-means algorithms and make the clustering results more accurate.

In the experiment, we used high-dimensional data obtained from single-cell sequencing technology and performed clustering analysis. The results show that using the Louvain algorithm for clustering analysis can more accurately discover cell types and clustering structures.

* 王科

2 Materials and Methods

2.1 Data sources and preprocess

The Dorsolateral Prefrontal Cortex (DLPFC) dataset is a dataset of gene expression data from the human brain, specifically from the DLPFC region. It was generated by the Allen Institute for Brain Science and is available on their website. The dataset includes data from 12 different layers of the DLPFC region and is commonly used for gene expression analysis and to study the molecular and cellular organization of the brain. There are many papers and studies that have used this dataset to investigate various aspects of brain biology and disease. Some references that use this dataset include[Ma et al., Miller et al.].

We use the processed data from

`https://edward130603.github.io/BayesSpace/articles/maynard_DLPFC.html`[Zhao et al.]

The study presents a Bayesian model-based analysis of spatial transcriptomics data from the human dorsolateral prefrontal cortex (DLPFC) to reveal cell type-specific patterns of gene co-expression. The study used a six-layered human DLPFC dataset to discover the gene expression patterns in different cell types and to identify cell type-specific modules of co-expressed genes. The study also applied the BayesSpace algorithm to the DLPFC dataset to identify spatially restricted gene expression patterns and cell type-specific transcriptional modules.

2.2 Traditional methods

VAE (Variational Autoencoder) is a generative model that uses neural networks to reduce high-dimensional data. VAE is composed of two parts: an encoder and a decoder. The encoder maps high-dimensional data to a low-dimensional space, while the decoder reconstructs low-dimensional data into high-dimensional data. The key of VAE is that it uses the idea of variational inference, by maximizing the marginal log likelihood to learn the low-dimensional representation of high-dimensional data. In bioinformatics, VAE algorithm is widely applied in genomics data analysis, transcriptomics data analysis, single-cell data analysis, and other fields. A VAE-based method [Lopez et al.] was introduced for analyzing single-cell transcriptomics data and identifying cell subpopulations. Another VAE-based unsupervised deep embedding method [Cai et al.] was proposed for clustering analysis, which is applied to identify cell subpopulations in scRNA-seq data.

The Louvain algorithm is a method for community detection in networks, which can partition the nodes of a network into different communities. The basic idea of the algorithm is to iteratively re-organize the community structure in the network to maximize the sum of the weights of edges within communities. In bioinformatics, the Louvain algorithm is commonly used to analyze biological networks, such as protein-protein interaction networks and metabolic networks. By using the Louvain algorithm, it is possible to discover community structures in biological networks, thereby improving our understanding of the functions and mechanisms of biological systems.Louvain agglomerative hierarchical clustering[Seth et al.] provides a robust approach to efficiently discover cluster-specific frequent biomarkers, i.e., overlapping biomarkers from single-cell RNA sequencing data.

The major advantage of k-means in bioinformatics is its ability to identify clusters of similar samples, which can be used to identify subpopulations of cells, differentially expressed genes, or other biologically meaningful groups. Handhayani and Hiryanto's study using gene expression of human colorectal carcinoma had shown that the genes were meaningfully grouped into three. Additionally, k-means is computationally efficient and easy to implement, making it a popular choice for large-scale data analysis in bioinformatics.

2.3 Improvement

K-means has two main disadvantages[Jain]:

1. It assumes that clusters are spherical and equally sized, which may not always be the case in real-world data.
2. It is sensitive to the initial placement of centroids, which can lead to different results depending on the initialization method used.

Meanwhile the Louvain algorithm is a popular method for community detection in complex networks, but it does have some limitations. One disadvantage is that it may not perform well on sparse networks. For example, a study by Traag et al. found that the Louvain algorithm can have poor performance on sparse networks, especially when the edge density is low. They proposed an extension of the Louvain algorithm, called the Leiden algorithm, which addresses this issue.In contrast, one of VAE's main advantages is its ability to handle missing data, or sparse data, by using the reparameterization trick. This allows the model to learn a probabilistic mapping between the data and its latent representation, which can be used to generate new samples.

Because VAE can learn useful features from sparse data, it can achieve better results when used in conjunction with clustering algorithms.After dimensionality reduction using VAE, using Louvain algorithm as the initialization of clustering centers for k-means clustering algorithm further improves clustering efficiency. Louvain algorithm can quickly discover community structures. This ensures the quality of initial clustering centers, maximizing the quality of clustering results. Finally, using the k-means clustering algorithm can achieve efficient clustering results. K-means is an iterative clustering algorithm that can quickly converge to the final result. Because k-means is computationally efficient and can handle large-scale data, it is widely used in the field of bioinformatics.

In summary, using VAE dimensionality reduction, Louvain initialization of clustering centers, and k-means clustering can effectively process sparse data and improve clustering efficiency and quality

2.4 Experimental Design and Evaluation

The control group is scDeepCluster[Zhao et al.].We will demonstrate the performance of DeepKmeans on the dorsolateral prefrontal cortex (DLPFC) samples, and compare it with the scDeepCluster to highlight its accuracy and efficiency. The evaluation indicators use ARI (Adjusted Rand Index) and NMI (Normalized Mutual Information). ARI is a measure of similarity between two clustering results. It ranges from -1 to 1, where a value of 1 indicates that the two clusterings are identical, a value of 0 indicates that the two clusterings are independent, and a negative value indicates that the two clusterings are dissimilar. The formula for ARI is:

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]} \tag{1}$$

where RI is the Rand Index, which measures the number of pairs of elements that are either in the same cluster in both clusterings or in different clusters in both clusterings, and Expected_RI is the expected value of RI under the assumption that the two clusterings are independent. Normalized Mutual Information (NMI) is another measure of similarity between two clustering results. It ranges from 0 to 1, where a value of 1 indicates that the two clusterings are identical, and a value of 0 indicates that the two clusterings are independent. The formula for NMI is:

$$NMI = \frac{2\,\mathcal{I}(C_1, C_2)}{\mathcal{H}(C_1) + \mathcal{H}(C_2)} \tag{2}$$

where $\mathcal{I}(C_1, C_2)$ is the mutual information between the $\mathcal{H}(C_1)$ is the entropy of the first clustering, and $\mathcal{H}(C_2)$ is the entropy of the second clustering.

In general, ARI is more robust to the effect of different numbers of clusters, while NMI is more robust to the effect of different cluster sizes. Both ARI and NMI can be used to evaluate the performance of a clustering algorithm, with NMI being more popular in bioinformatics.

3 Results

3.1 Analysis

In the study, we applied DeepKmeans to analyze a specific dataset of 12 slices from the Dorsolateral Prefrontal Cortex (DLPFC). The results of this analysis were evaluated using two commonly used metrics, the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI). The accuracy of the algorithm, as measured by these metrics, is presented in detail in Table 1. Additionally, we have also included a figure 1 to visualize the results of the analysis in terms of the time consumed.The maximum value of ARI is 0.5593, the minimum value is 0.3008, and the average value is 0.43. The maximum value of NMI is 0.6204, the minimum value is 0.4437, and the average value is 0.53.

When comparing the results obtained using DeepKmeans with those obtained using scDeepCluster, as shown in figure 2, we can see that DeepKmeans is more accurate and efficient. Specifically, we can see that the accuracy of DeepKmeans, as measured by the ARI and NMI metrics, is both higher than that of scDeepCluster. Furthermore, we can also see that the time consumed by DeepKmeans is much lower than that of scDeepCluster, takes only one-fifth the time.

Overall, this study demonstrates the effectiveness and efficiency of the DeepKmeans algorithm in analyzing high-dimensional datasets in the field of clustering single-cell RNA-seq data.

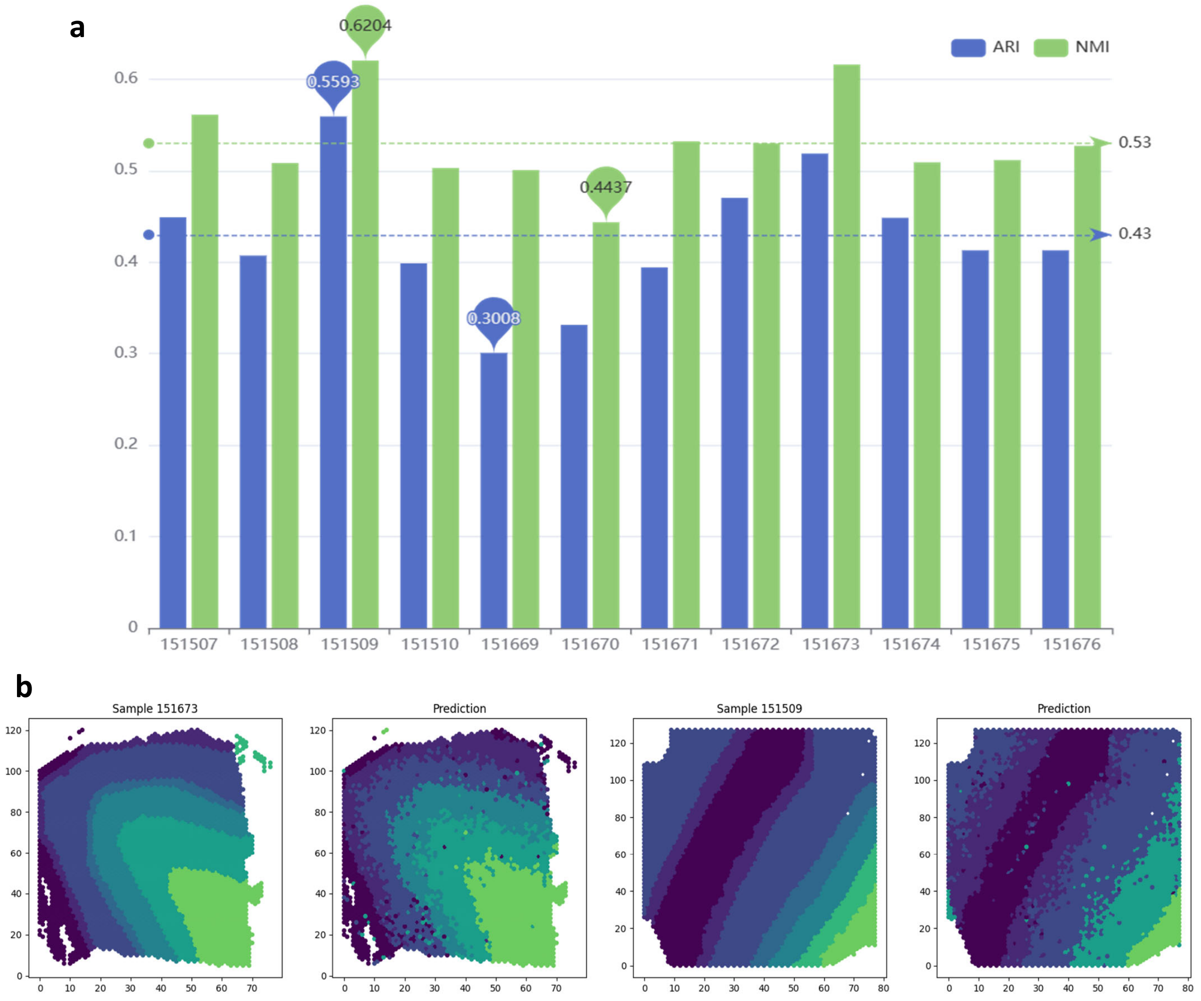

Figure 1: Spatial domains and SVGs detected in the LIBD human dorsolateral prefrontal cortex data. **a,** Histogramb of clustering accuracy in all twelve samples. The ARI and NMI is used to compare similarity between cluster labels from each method against the manually annotated layers for all twelve samples. The maximum, minimum, and average values are marked in the graph. **b,** Ground-truth segmentation and cluster assignments generated by DeepKmeans of cortical layers and white matter (WM) in the DLPFC section 151509(left) and 151673(right).

3.2 Bioinformatics Sense

Through comparing the clustering results of different methods, we can find that using VAE for dimensionality reduction, initializing the clustering centers with the Louvain algorithm, and finally completing the clustering with k-means has significant advantages in terms of clustering accuracy and time efficiency. In this experiment, the clustering results of this method reached a high level in both ARI and NMI indicators, which indicates that the clustering results have good accuracy.In addition, the excellent performance of this method on high-dimensional sparse data also shows the robustness and efficiency of VAE in dimensionality reduction for sparse data.

In terms of biological significance, this method can effectively cluster biological data, and the clustering results can help us discover similarities and differences between genomes, proteins, or other biological samples. By comparing the clustering results of different methods, we can better understand the biology of gene expression, protein interactions, and other processes. This information helps us better understand the structure and function of biological systems, and ultimately contributes to drug development and disease diagnosis.

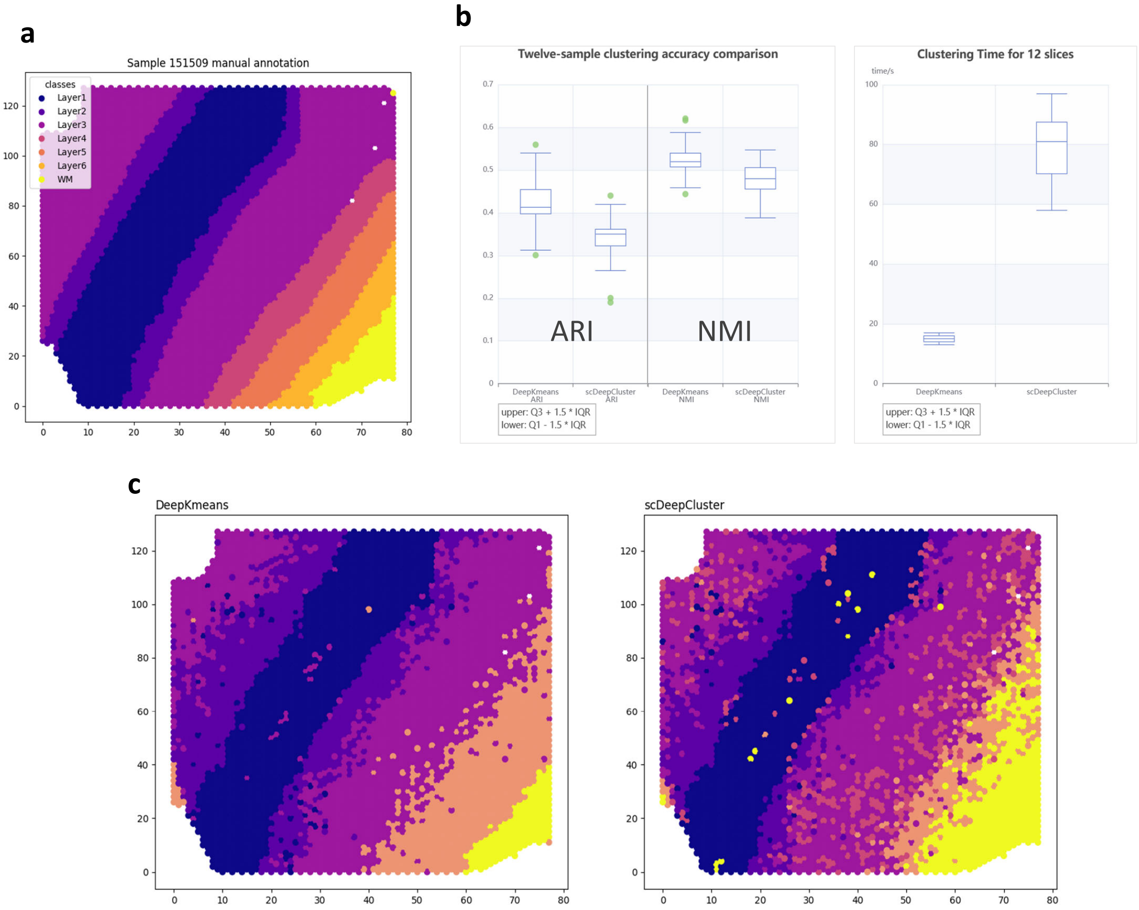

Figure 2: DeepKmeans improves the identification of layer in human dorsolateral prefrontal cortex tissue. **a,** Ground truth. We highlight the manually annotated six DLPFC layers and white matter (WM) in sample 151509 from the spatialLIBD dataset. **b,** Boxplot of clustering accuracy(left) and time consumed(right) in all 12 sections of the DLPFC dataset in terms of adjusted rand index (ARI) and normalized mutual information (NMI) scores for seven methods. In the boxplot, the center line, box limits and whiskers denote the median, upper and lower quartiles, and 1.5‘Ů interquartile range, respectively. **c,** Cluster assignments generated by DeepKmeans(left) and scDeepCluster(right) in the DLPFC section 151509.

Table 1: Clustering accuracy and time consumed of DeepKmeans and scDeepCluster.

Sample	DeepKmeans			scDeepCluster		
	NMI	ARI	Time(s)	NMI	ARI	Time(s)
151507	0.5613	0.4491	14	**0.5468**	**0.4400**	60
151508	0.5084	0.4072	14	0.4605	0.3601	72
151509	**0.6204**	**0.5593**	15	0.5055	0.3599	85
151510	0.5030	0.3987	14	0.5211	0.4172	75
151669	0.5007	0.3008	13	0.3880	0.2004	65
151670	0.4437	0.3315	13	0.4400	0.1902	58
151671	0.5321	0.3943	17	0.5061	0.3650	97
151672	0.5298	0.4703	16	0.4570	0.3388	84
151673	0.6161	0.5188	16	0.5051	0.3603	92
151674	0.5091	0.4485	16	0.4742	0.3403	92
151675	0.5115	0.4129	15	0.4852	0.3080	86
151676	0.5271	0.4129	15	0.4508	0.3276	78

4 Discussion

4.1 Limitations

In this study, we investigated the use of VAE dimensionality reduction, Louvain initialization of clustering centers, and K-means clustering methods and evaluated their application in bioinformatics. The experimental results showed that this method has significant improvement in terms of clustering efficiency and accuracy. We also discussed the biological significance of this method and looked at future research directions.

In summary, our research demonstrates that DeepKmeans is an effective clustering method that can improve clustering efficiency and accuracy. However, this study also has some limitations, such as the size and complexity of the dataset. One main disadvantage is that the Louvain algorithm is sensitive to the resolution parameter. A study by Mucha et al. found that the algorithm can produce different community structures for different values of the resolution parameter, which can affect the interpretability of the results.

4.2 Future Work

Future research should continue to explore the application of this method in other fields and further improve the performance of the clustering algorithm. In future research, we can continue to explore the use of VAE dimensionality reduction clustering methods in other fields and conduct further research to improve the accuracy and reliability of the clustering results. In addition, we can also combine other clustering algorithms, such as density-based clustering algorithms, to improve the diversity of the clustering results. Additionally, we can explore the use of bioinformatics-specific evaluation metrics in clustering to better assess the biological significance of the clustering results.

5 Conclusion

This study used VAE dimensionality reduction, Louvain initialization of clustering centers, and K-means clustering algorithms in the field of bioinformatics for experimentation. The results of the experiment showed that using VAE dimensionality reduction has a good effect on sparse data, and using the Louvain algorithm to initialize the K-means clustering results can improve the clustering efficiency and robustness. By using ARI and NMI evaluation indexes, the results show that this method is superior to other algorithms.

The main contribution of this study is to propose a new dimensionality reduction clustering algorithm with good application prospects in the field of bioinformatics. In order to better verify the effectiveness of this method, experiments can be conducted on different bio-data sets in the future and combined with other evaluation indexes. This study provides a new idea and method for single-cell sequencing data analysis. In the future, improvements can be made to other clustering algorithms. In addition, it can try to study the mechanism of VAE in single-cell data analysis more in-depth.

References

Shaojie Ma, Mario Skarica, Qian Li, Chuan Xu, Ryan D. Risgaard, et al. Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. 377(6614):eabo7257. doi:10.1126/science.abo7257. URL `https://www.science.org/doi/10.1126/science.abo7257`. Publisher: American Association for the Advancement of Science.

Jeremy A. Miller, Song-Lin Ding, Susan M. Sunkin, Kimberly A. Smith, Lydia Ng, et al. Transcriptional landscape of the prenatal human brain. 508(7495):199–206. ISSN 1476-4687. doi:10.1038/nature13185. URL `https://www.nature.com/articles/nature13185`. Number: 7495 Publisher: Nature Publishing Group.

Edward Zhao, Matthew R. Stone, Xing Ren, Jamie Guenthoer, Kimberly S. Smythe, Thomas Pulliam, et al. Spatial transcriptomics at subspot resolution with BayesSpace. 39(11):1375–1384. ISSN 1546-1696. doi:10.1038/s41587-021-00935-2. URL `https://www.nature.com/articles/s41587-021-00935-2`. Number: 11 Publisher: Nature Publishing Group.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. 15(12):1053–1058. ISSN 1548-7105. doi:10.1038/s41592-018-0229-2. URL `https://www.nature.com/articles/s41592-018-0229-2`. Number: 12 Publisher: Nature Publishing Group.

Jinyu Cai, Wenzhong Guo, and Jicong Fan. Unsupervised deep discriminant analysis based clustering. URL `http://arxiv.org/abs/2206.04686`.

Soumita Seth, Saurav Mallik, Tapas Bhadra, and Zhongming Zhao. Dimensionality reduction and louvain agglomerative hierarchical clustering for cluster-specified frequent biomarker discovery in single-cell sequencing data. 13. ISSN 1664-8021. URL `https://www.frontiersin.org/articles/10.3389/fgene.2022.828479`.

Teny Handhayani and Lely Hiryanto. Intelligent kernel k-means for clustering gene expression. 59:171–177. ISSN 1877-0509. doi:10.1016/j.procs.2015.07.544. URL `https://www.sciencedirect.com/science/article/pii/S1877050915020736`.

Anil K. Jain. Data clustering: 50 years beyond k-means. 31(8):651–666. ISSN 0167-8655. doi:10.1016/j.patrec.2009.09.011. URL `https://www.sciencedirect.com/science/article/pii/S0167865509002323`.

V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. 9(1):5233. ISSN 2045-2322. doi:10.1038/s41598-019-41695-z. URL `https://www.nature.com/articles/s41598-019-41695-z`. Number: 1 Publisher: Nature Publishing Group.

Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. 328(5980):876–878. doi:10.1126/science.1184819. URL `https://www.science.org/doi/10.1126/science.1184819`. Publisher: American Association for the Advancement of Science.