CS 439 25F
Clustering: k-Means and t-SNE

• • •

Group 24
Andrew Menyhert [amm926]
Rida Mohammad [rm1724]
Kaushik Murali [km1526]

Project Description

- This project investigates the label summaries in a Graph City of Paper Citation Network, where each building is a fixed point of degree peeling.
- It identifies salient labels within each fixed point using measures such as TF-IDF and examines their correlations with fixed point peel values and sizes (within a logarithmic factor). The project further explores whether these relationships enable predicting a paper's most probable building association based on the set of words in its title and abstract using Naïve Bayes classification.
- Finally, by representing each fixed point as a meta-node, the project evaluates their topological importance through PageRank, presenting an overall ranking on important labels.

Dataset Description and Sample Snapshots

- **Please note the following dataset size requirement for your project:**
  - **For tabular datasets (DataFrames), the dataset must be at least 8 MB in size.**
    - Example: A dataset with 100,000 records, each containing 20 features stored as 4-byte integers or floats, totals approximately 8 MB.
    - If your dataset is smaller than 8 MB, you may consider augmenting it by adding additional data or derived features (e.g., polynomial features such as squares or interaction terms).
  - **For image datasets** , we consider a typical 32 × 32 pixel image with four RGBA channels (4 Bytes per pixel), roughly 4 KB. To meet the 8 MB minimum requirement, **you will need at least 2,000 such images.**

- Bonus Policy:
  - To incentivize the use of larger datasets, for every additional 8 MB beyond the initial 8 MB, you will receive 3 bonus points, up to a maximum of 30 bonus points.

- Please ensure your dataset meets these requirements before submission.

Dataset Description and Sample Snapshots

- Dataset Name: ogbn-papers100M
  - a directed citation graph of 111 million papers indexed by MAG
  - Vertices: 111M papers
  - Edges: 1.6B citations

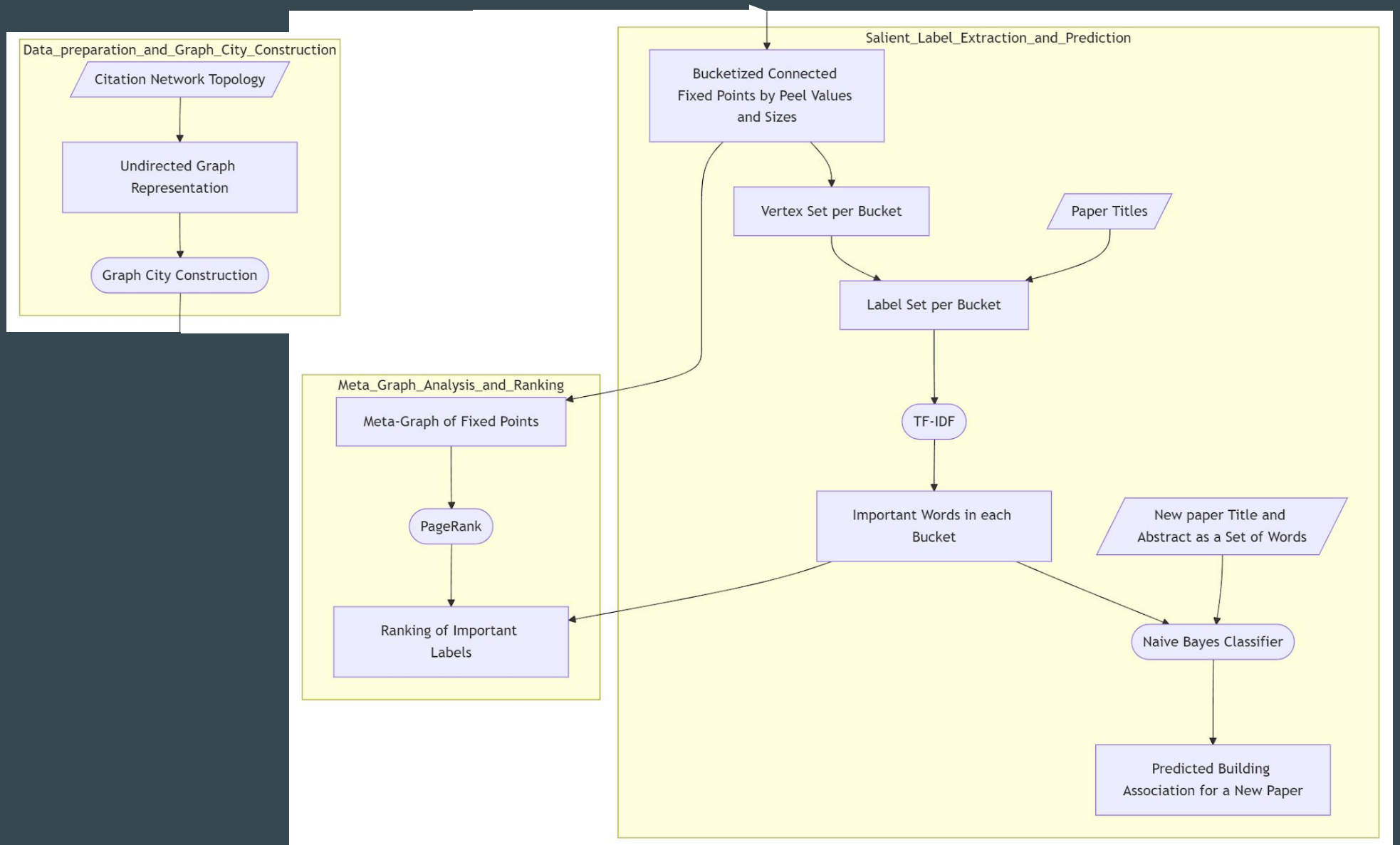
  - Snapshot of dataset format
  - Sample plots of data statistics (deg dist, avg deg, cc sizes dist, avg cc sizes, ...)

Questions to be Addressed

- Which vertex labels are considered “salient” within a Graph City Building? (A building corresponds to what is called a fixed point of graph degree peeling.)
- Do the “salient” vertex labels in each building exhibit correlations with the corresponding building peel values and/or building sizes (within a log factor)?
  If such relationships exist, is it possible to predict the most probable building association for a given paper based on the words it contains?

- Viewing each building as a meta-node, constructing an intersection graph of buildings. Which are more “relevant” to a query in terms of PageRank?
- Analyze the impact of citation directions between “local vertices” in a building and between them and “shared vertices”
- Based on the PageRank distribution of each building, can one construct an overall ranking of “salient” labels of the entire citation network?
- Are there any other “importance” measurements, e.g., betweenness centrality, that can be used to extract topic semantics?

Data Transformation Flowchart



Exploration Algorithms Used

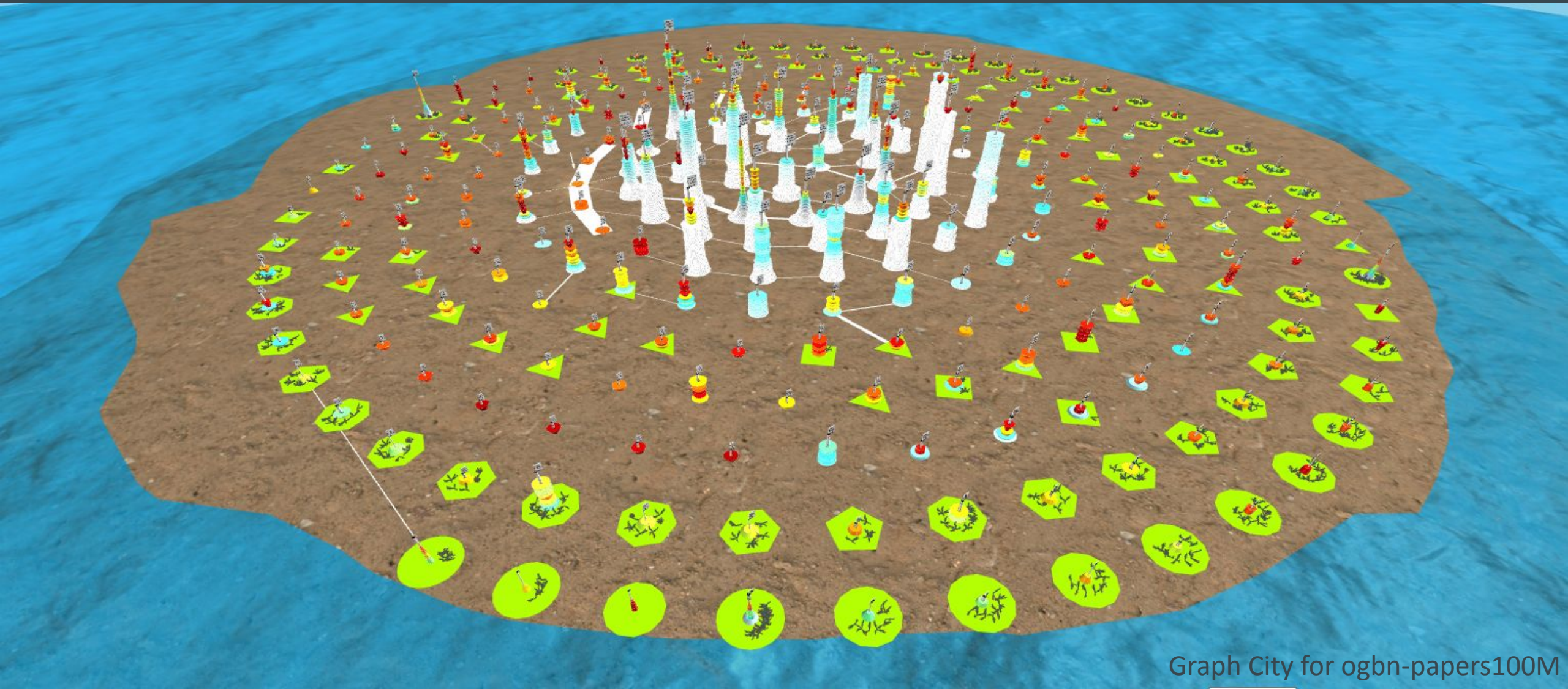
- Data Cleaning
- Graph Cities Processing
  - Vertex Peeling (k-core)
  - Iterative Graph Edge Partition (fixed points of graph degree peeling)
- Label Summarization and Prediction
  - TF-IDF for “salient” labels appearing in each building
  - Naïve Bayes Classifier for the prediction of the probability of a paper being associated with a building
  - PageRank to assign a level of “importance” to the buildings in the Graph City

Computational Platform and Software Libraries Used

- Preprocessing
  - Python: OGB, NumPy, Pandas, NetworkX, NLTK, spaCy
- Graph Cities
  - C++: The Boost Graph Library
  - Python: NumPy, SciPy, Matplotlib, Pandas
  - JavaScript: Node.js, Three.js, D3.js
- Label Summarization and Prediction
  - Python: NetworkX, PySpark, scikit-learn, NLTK, spaCy, wordcloud

Gantt Charts

Visualizations



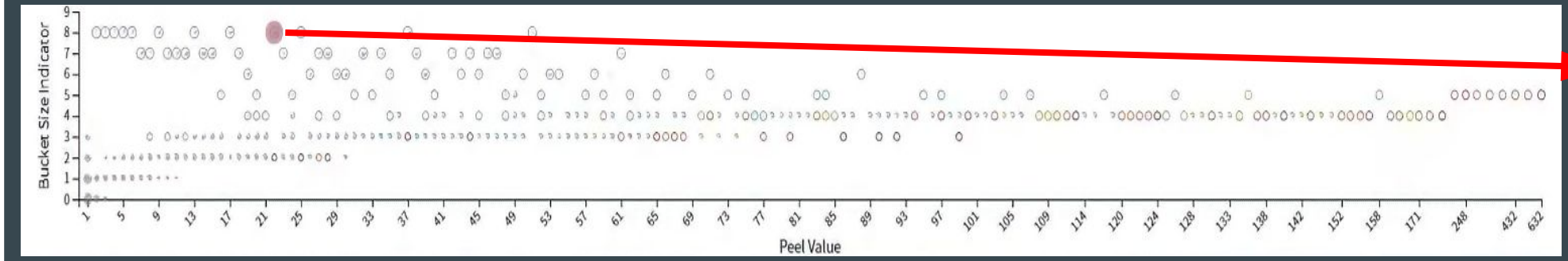
Graph City for ogbn-papers100M

Visualizations

- Label Summaries as Word Clouds of selected fixed points

Dash Board

Bucketized Fixed Points Grid Map Selector



Label Summary

Word Cloud

Predict the probability of a paper being associated with a building / building bucket

Title

Abstract

Papers Titles with Selected Labels

Interesting Findings and Conclusions

Deliverables

- Source Code (either in a .zip directory or a link to a GitHub repo.)
- Video Demo (at most 3-4 minutes, with captions or voice over)
- Project Report (a .pdf file following ACM / IEEE conference format)
- Presentation Slides (a .pptx or .pdf file)

Future Work Extensions

References

- James Abello, Haoyang Zhang, Daniel Nakhimovich, Chengguizi Han, and Mridul Aanjaneya. 2022. Giga Graph Cities: Their Buckets, Buildings, Waves, and Fragments. *IEEE Computer Graphics and Applications* 42, 3 (2022), 53–64. DOI:https://doi.org/10.1109/MCG.2022.3172650
- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv preprint arXiv:2005.00687* (2020).
- Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly Media Inc.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, (2020), 357–362. DOI:https://doi.org/10.1038/s41586-020-2649-2
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, (2020), 261–272. DOI:https://doi.org/10.1038/s41592-019-0686-2
- Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- Wes McKinney and others. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Austin, TX, 51–56.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
- Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, and others. 2016. Apache spark: a unified engine for big data processing. *Communications of the ACM* 59, 11 (2016), 56–65.