TooManyCells identifies and visualizes relationships of single-cell clades

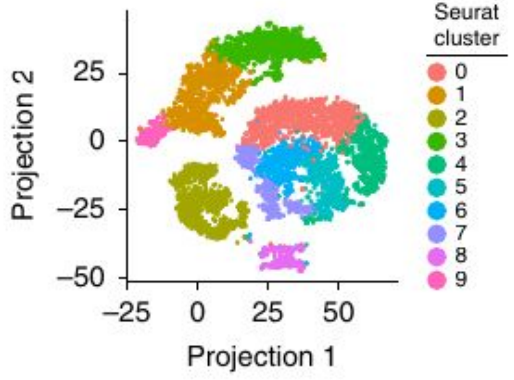
Gregory W. Schwartz et al. (2020/04)

Outlines

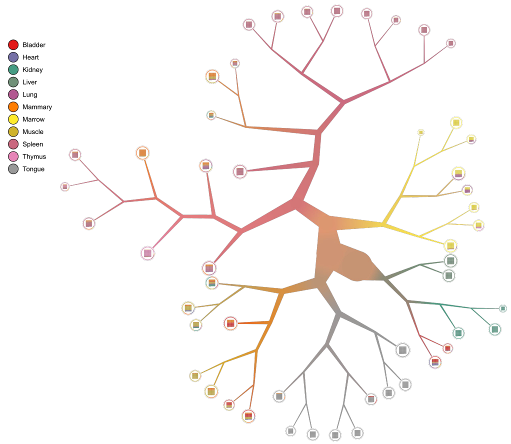
1. Describe the method in human language
2. The Development of key concepts
3. How it works
4. How it helps my project

1 TooManyCells: Clustering Algorithm

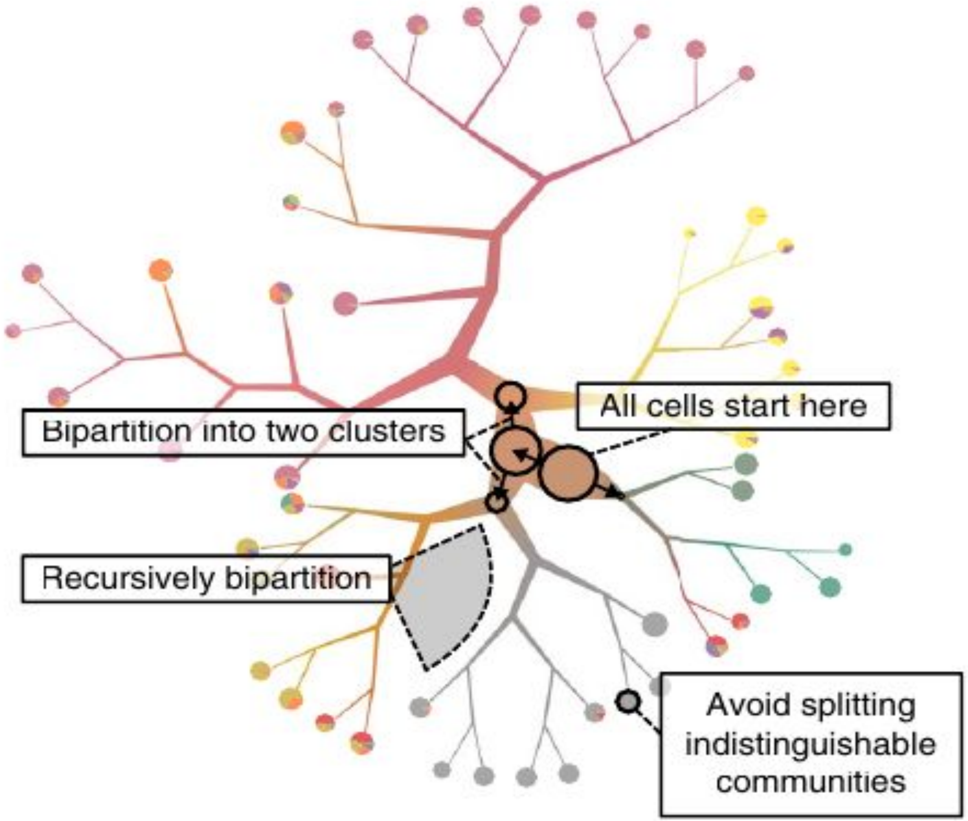
Seurat Nearest neighbor clustering determine a unique position of each cell based on their coordination in latent space (e.g. PCA).

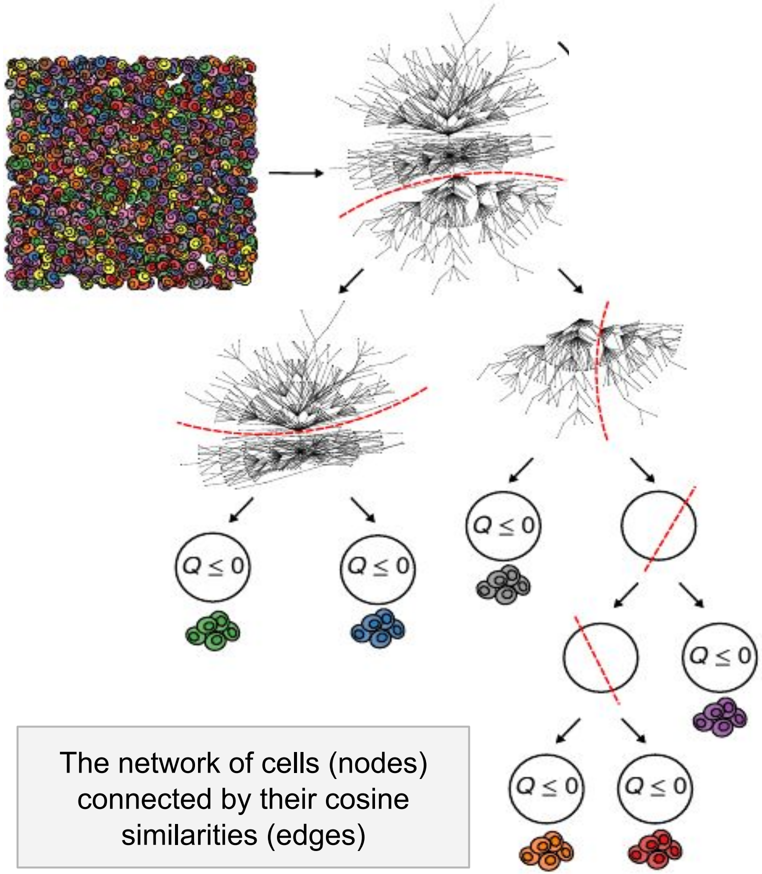


too-many-cells algorithm **recursively divides cells into two clusters each time** and relates clusters while branching the tree.



1 TooManyCells: Clustering Algorithm



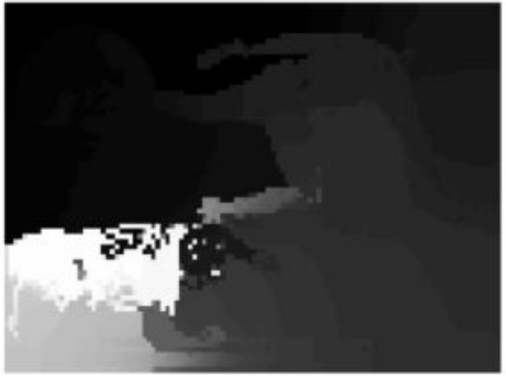


The network of cells (nodes) connected by their cosine similarities (edges)

2 Development of Key Concepts: Computer Vision

Normalized Cuts and Image Segmentation (Jianbo Shi et al. 2000)

- Similarity between pixels are calculated by color change in euclidean space ***A***.
- Define Normalized Cuts, Graph Laplacian Matrix ***L***
- Bring up a brilliant idea that it is possible to optimally biparition a picture by the second smallest eigenvector of matrix ***L***



2 Development of Key Concepts: Text Mining

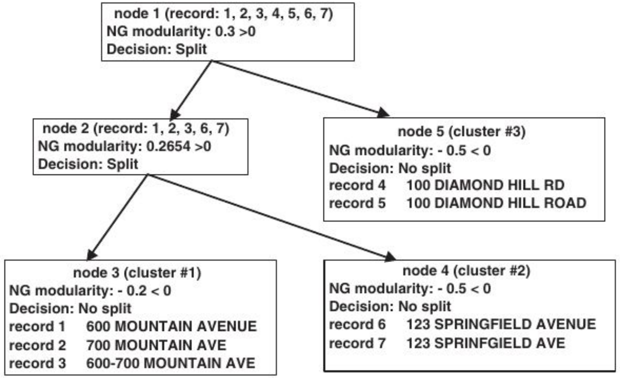
Normalized Cuts and Image Segmentation (Jianbo Shi et al. 2000)

- Similarity between pixels are calculated by color change in euclidean space ***A***.
- Define Normalized Cuts, Graph Laplacian Matrix ***L***
- Bring up a brilliant idea that it is possible to optimally biparition a picture by the second smallest eigenvector of matrix ***L***

Efficient Spectral Neighborhood Blocking for Entity Resolution (Liangcai et al 2011)

- Use q-gram, TF-IDF and cosine similarity to describe the similarities ***A*** among records
- Build “Graph” Laplacian Matrix ***L*** based on ***A***
- Fast way to calculate the second eigenvector in **sparse** matrix
- Introduce Newman-Girvan modularity from social network research to decide when to stop splitting.

Record#	Address
1	600 MOUNTAIN AVENUE
2	700 MOUNTAIN AVE
3	600-700 MOUNTAIN AVE
4	100 DIAMOND HILL RD
5	100 DIAMOND HILL ROAD
6	123 SPRINGFIELD AVENUE
7	123 SPRINFGIELD AVE



2 Development of key concepts: Single Cells

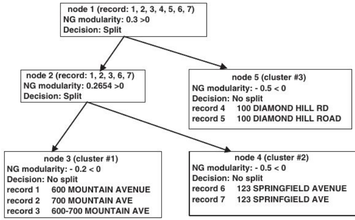
Normalized Cuts and Image Segmentation (Jianbo Shi et al. 2000)

- Similarity between pixels are calculated by color change in euclidean space ***A***.
- Define Normalized Cuts, Graph Laplacian Matrix ***L***
- Bring up a brilliant idea that it is possible to optimally biparition a picture by the second smallest eigenvector of matrix ***L***

Efficient Spectral Neighborhood Blocking for Entity Resolution (Liangcai et al 2011)

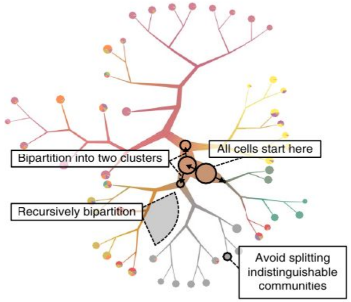
- Use q-gram, TF-IDF and cosine similarity to describe the similarities ***A*** among records
- Build “Graph” Laplacian Matrix ***L*** based on ***A***
- Fast way to calculate the second eigenvector in sparse matrix
- Introduce Newman-Girvan modularity from social network research to decide when to stop splitting.

Record#	Address
1	600 MOUNTAIN AVENUE
2	700 MOUNTAIN AVE
3	600-700 MOUNTAIN AVE
4	100 DIAMOND HILL RD
5	100 DIAMOND HILL ROAD
6	123 SPRINGFIELD AVENUE
7	123 SPRINFGIELD AVE



TooManyCells identifies and visualizes relationships of single-cell clades (Gregory W. Schwartz et al. 2020)

- TF-IDF and cosine similarity to describe the similarities ***A*** among cells.
- Implement Liangcai’s clustering algorithm
- Add more downstream analysis tools: preprocessing options, normalization methods, visualization, Differentially Expression, Diversity Analysis and Cluster Purity...



3 How it works

Suppose you have a similarity matrix **A** where A(i, j) represents the similarity between item i and j.
Then we can calculate the degree matrix **D** = diag(**A1**), there d(i) is sum_{j} A(i, j).
Define graph laplace matrix:

$$\mathcal{L}(\mathbf{A}) = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

It is shown that the second smallest eigenvector can be used to optimally bipartition the dataset.

3 How it works

Suppose **B1** (m x n) is a SC-seq UMI read matrix. Use **TF-IDF** to normalize the counts.

$$\mathbf{B}_2 = \log(m/d_j)\mathbf{B}_1(i, j)$$ (If the degree is high, then we add a penalty to the frequency.)

Suppose you have a similarity matrix **A** where A(i, j) represents the similarity between item i and j.
Then we can calculate the degree matrix **D** = diag(**A1**), there d(i) is sum_{j} A(i, j).
Define graph laplace matrix:

$$\mathcal{L}(\mathbf{A}) = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

It is shown that the second smallest eigenvector can be used to optimally bipartition the dataset.

3 How it works

Suppose **B1** (m x n)is a SC-seq UMI read matrix. Use **TF-IDF** to normalize the counts **B2**. Use Cosine Similarity to represent the distance between two cells.

$$\mathbf{A}(i,j) = \frac{\sum_{k=1}^{n} \mathbf{B}_2(i,k)\mathbf{B}_2(j,k)}{\sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(i,k)}\sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(j,k)}}$$

Suppose you have a similarity matrix **A** where A(i, j) represents the similarity between item i and j.
Then we can calculate the degree matrix **D** = diag(**A1**), there d(i) is sum_{j} A(i, j).
Define graph laplace matrix:

$$\mathcal{L}(\mathbf{A}) = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

It is shown that the second smallest eigenvector can be used to optimally bipartition the dataset.

3 How it works

Suppose **B1** (m x n)is a SC-seq UMI read matrix. Use **TF-IDF** to normalize the counts **B2**. Use Cosine Similarity to represent the distance between two cells.

$$\mathbf{A}(i,j) = \frac{\sum_{k=1}^{n} \mathbf{B}_2(i,k)\mathbf{B}_2(j,k)}{\sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(i,k)}\sqrt{\sum_{k=1}^{n} \mathbf{B}_2^2(j,k)}}$$

Suppose you have a similarity matrix **A** where A(i, j) represents the similarity between item i and j. Then we can calculate the degree matrix **D** = diag(**A1**), there d(i) is sum_{j} A(i, j).
Define graph laplace matrix:

$$\mathcal{L}(\mathbf{A}) = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

It is shown that the second smallest eigenvector can be used to optimally bipartition the dataset.

Use Newman-Girvan Modularity as stopping criteria.

$$Q(C_1, C_2) = \sum_{k=1}^{2}\left(\frac{O_{kk}}{L} - \left(\frac{L_k}{L}\right)^2\right)$$

Q > 0 denotes non-random communities
Q < 0 demonstrates communities randomly found

4 TooManyCells on Traf6 dataset

After integration, apply tooManyCells algorithm on 20-dim latent subspace.

But there are too many subsets in binary tree and we should consider pruning the branches.

4 TooManyCells on Traf6 dataset

MAD * 5		
Cluster	Size	WT%
10	255	67.06
11	159	72.32
7	203	65.51
8	218	65.51
13	208	52.40
14	218	52.40
15	231	58.44
16	251	62.15
22	231	55.41
21	2674	49.92
18	240	58.75
19	236	59.32