Applications of Community Detection

Ian Chen

University of Illinois Urbana-Champaign

August 15, 2026

Clustering

- Given $G = (V, E)$, partition V into dissimilar groups of like nodes.
- Given $A \in \mathbb{R}^{n \times n}_{\geq 0}$ distance matrix, partition n
- Given $X = \{ \mathbf{x}_1, \ldots, \mathbf{x}_n \}$ feature vectors on n objects, partition X
  - Nodes are authors, features are the set of published papers $\implies$ co-authorship network
  - Nodes are products, features are whether customers bought them $\implies$ co-purchasing network
  - Nodes are roads, features are locations $\implies$ road network
  - Nodes are pixels, features are position, intensity $\implies$ image segmentation
  - Nodes are cells, features are gene expression $\implies$ cell-type identification

Clustering

- Given $G = (V, E)$, partition V into dissimilar groups of like nodes.
- Given $A \in \mathbb{R}^{n \times n}_{\geq 0}$ distance matrix, partition n
- Given $X = \{ \mathbf{x}_1, \ldots, \mathbf{x}_n \}$ feature vectors on n objects, partition X
  - Nodes are authors, features are the set of published papers $\implies$ co-authorship network
  - Nodes are products, features are whether customers bought them $\implies$ co-purchasing network
  - Nodes are roads, features are locations $\implies$ road network
  - Nodes are pixels, features are position, intensity $\implies$ image segmentation
  - Nodes are cells, features are gene expression $\implies$ cell-type identification

From the view of a biologist

DISCLAIMER: I AM NOT A BIOLOGIST

From the view of a biologist

- The lifecycle of the data:
  1. Single cells are given unique identifiers
  2. Short gene snippets is collected, amplified, and de-duplicated into library
  3. Sequencing & alignment into gene databases
  4. Collect into count matrices; `cells` $\mapsto$ `genes`
- What can go wrong?

From the view of a biologist (cont.)

- An incomplete list of what can go wrong:
  1. Barcodes not mapping injectively from cells.
  2. Barcodes may be from dying, broken, or doublet cells.
  3. Cells may be at different times in their life cycle.
  4. Different batches may have different confounders (temperature, ...).
  5. Different samples may get over- or under-sampled.
  6. Scaling challenge from massive genomes.
- Data is noisy $\implies$ the network is unreliable (missing or incorrectly weighted edges)

What happens upstream

- Quality Control
  - Reduces overcounting and undercounting genes
- Normalization
  - Reduces experimental noise (batching, inflation, loss of signal)
- Dimension reduction
  - PCA, t-SNE, UMAP, auto-encoders, CIDR
  - Improve computational efficiency and reduces noise
- k-Nearest neighbors
  - Forming a sparse graph; keeping only strong connections in adjacency
  - Improves computational efficiency and reduces noise

What happens downstream

- Cluster annotation: finding gene signatures for each cluster
  - Whether the clusters agree upon biological markers / metadata (external information).
  - Question: what resolution of clusters are meaningful? Are “T-cells” cell types, or “$CD4^+$” and “$CD88^+$” T-cells different types?
  - Comparison to reference databases – experimentally derived.
- Compositional analysis: changes wrt background (environment)
  - e.g. proportion of enterocytes changes for patients with salmonella infections.
- Trajectory analysis and inference: dynamic gene expression and differentiation
- Many more: cell-interactions, metastable states, ...

Re-evaluating clustering

- What types of clusters do we want?
  - Small, large?
  - Separated, cohesive?
- What types of methods do we want?
  - Should we tell it how many clusters to form?
  - What about a minimum or maximum size?
  - Methods that give us a confidence level, so that statistical testing can be performed?
- What makes modularity optimization so salient? Is it just first-mover advantage?

References I

[1] Malte D Luecken and Fabian J Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Molecular Systems Biology* 15.6 (June 2019). ISSN: 1744-4292. DOI: 10.15252/msb.20188746. URL: http://dx.doi.org/10.15252/msb.20188746.

[2] Shixiong Zhang et al. “Review of single-cell RNA-seq data clustering for cell-type identification and characterization”. In: *RNA* 29.5 (Feb. 2023), pp. 517–530. ISSN: 1469-9001. DOI: 10.1261/rna.078965.121. URL: http://dx.doi.org/10.1261/rna.078965.121.