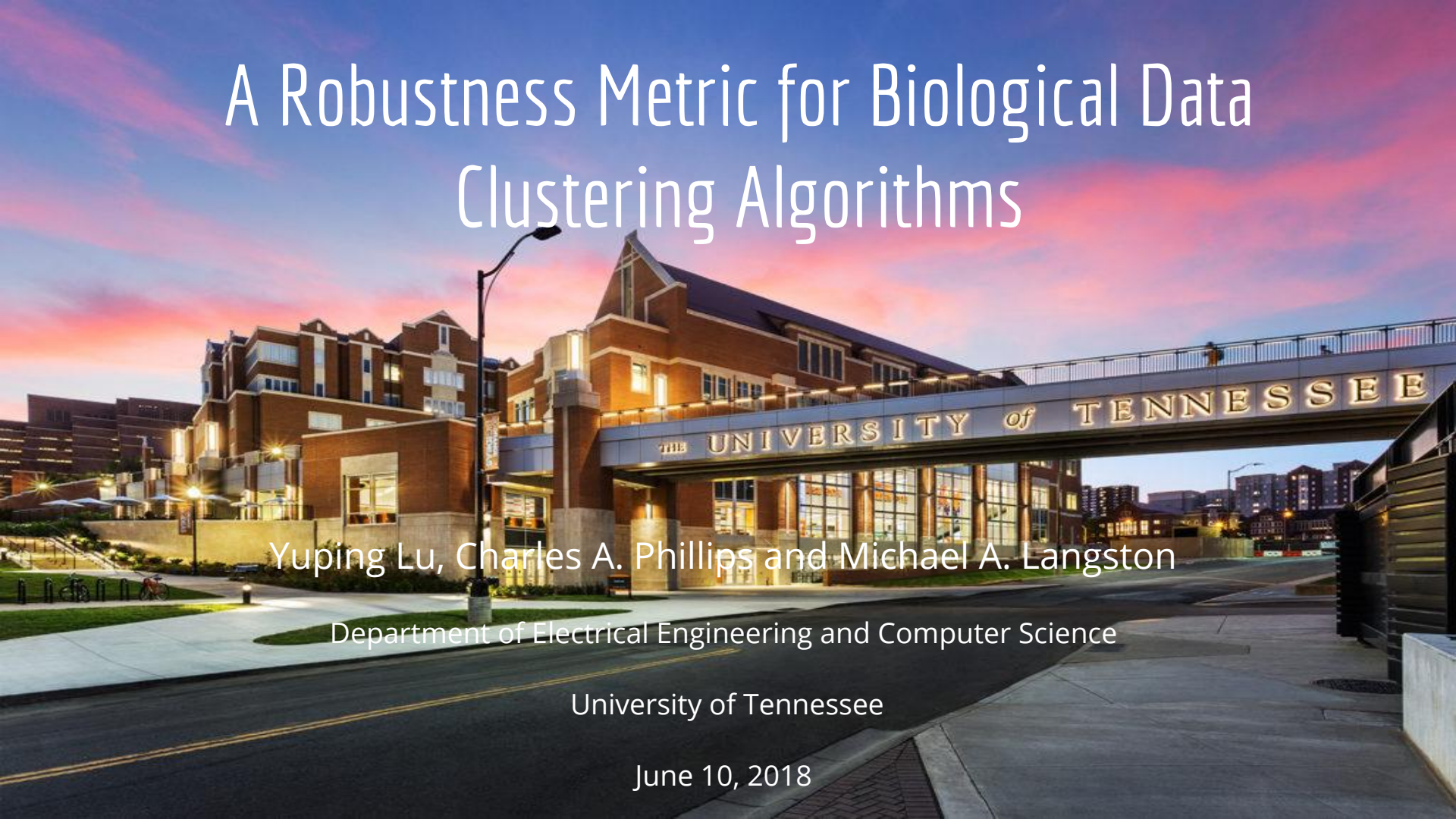
A Robustness Metric for Biological Data Clustering Algorithms



Yuping Lu, Charles A. Phillips and Michael A. Langston

Department of Electrical Engineering and Computer Science

University of Tennessee

June 10, 2018

Clustering Method Basics

Clustering algorithms are generally used to classify a set of objects into subsets using some measure of similarity between each object pair.

Measurement of cluster quality:

1) known classification schemes, e.g. domain-specific knowledge such as ontological enrichment.
2) theoretical standards, e.g. modularity, clustering coefficient, silhouette coefficient and etc.

New Comparison Metric

Clustering algorithms typically have one or more adjustable settings.

Such a setting may represent, for example, a preset variable, a parameter of interest, or various sorts of initial assignments.

A question of interest then is this: to what degree do the clusters produced vary as setting values change?

Robustness Intuition

If a pair of items appear together once, will they appear together consistently?

Robustness Definition

$$R = t / (dr)$$

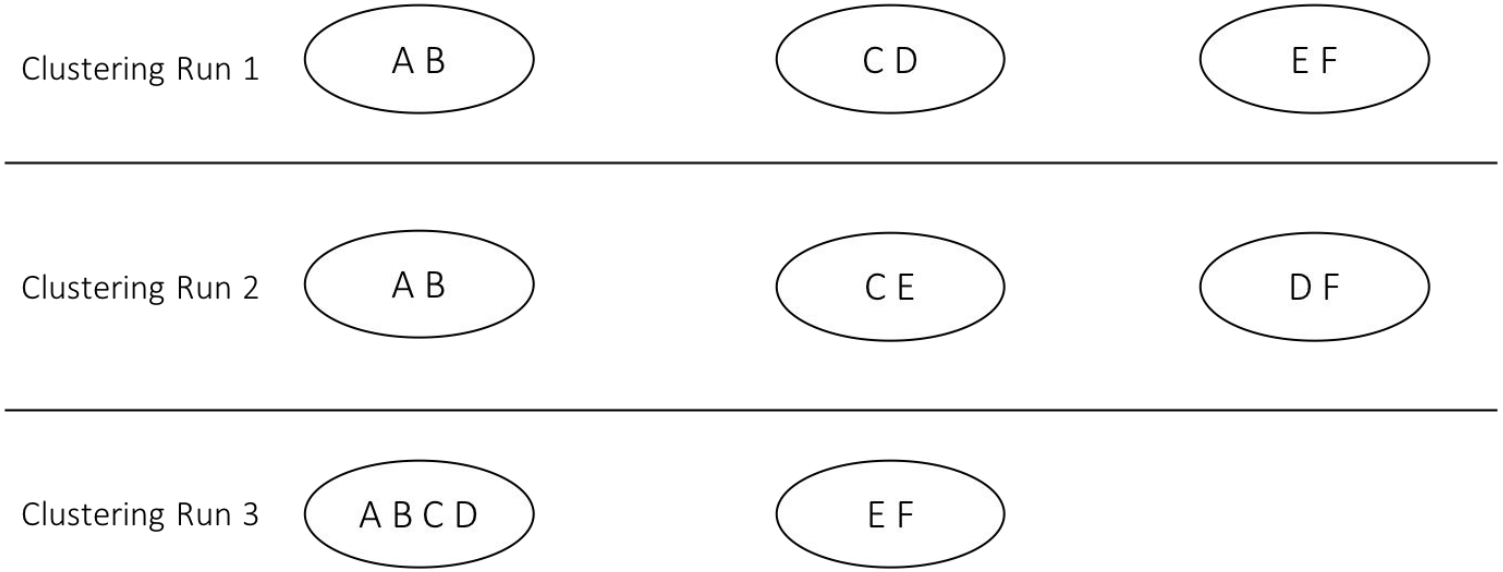
t - the total number of (not necessarily distinct) pairs of objects that appear together in some cluster summed over all runs.

d - the number of distinct pairs of objects that appear together in some cluster produced by some run.

r - the number of times the clustering algorithm was run, each run using a different value for some setting of interest.

Robustness lies in the interval **(0, 1]**.

An Example



(A,B): 3/3; (A,C): 1/3; (A,D): 1/3; (B,C): 1/3; (B,D): 1/3; (C,D): 2/3; (C,E): 1/3; (D,F): 1/3; and (E,F): 2/3. Robustness = 0.481

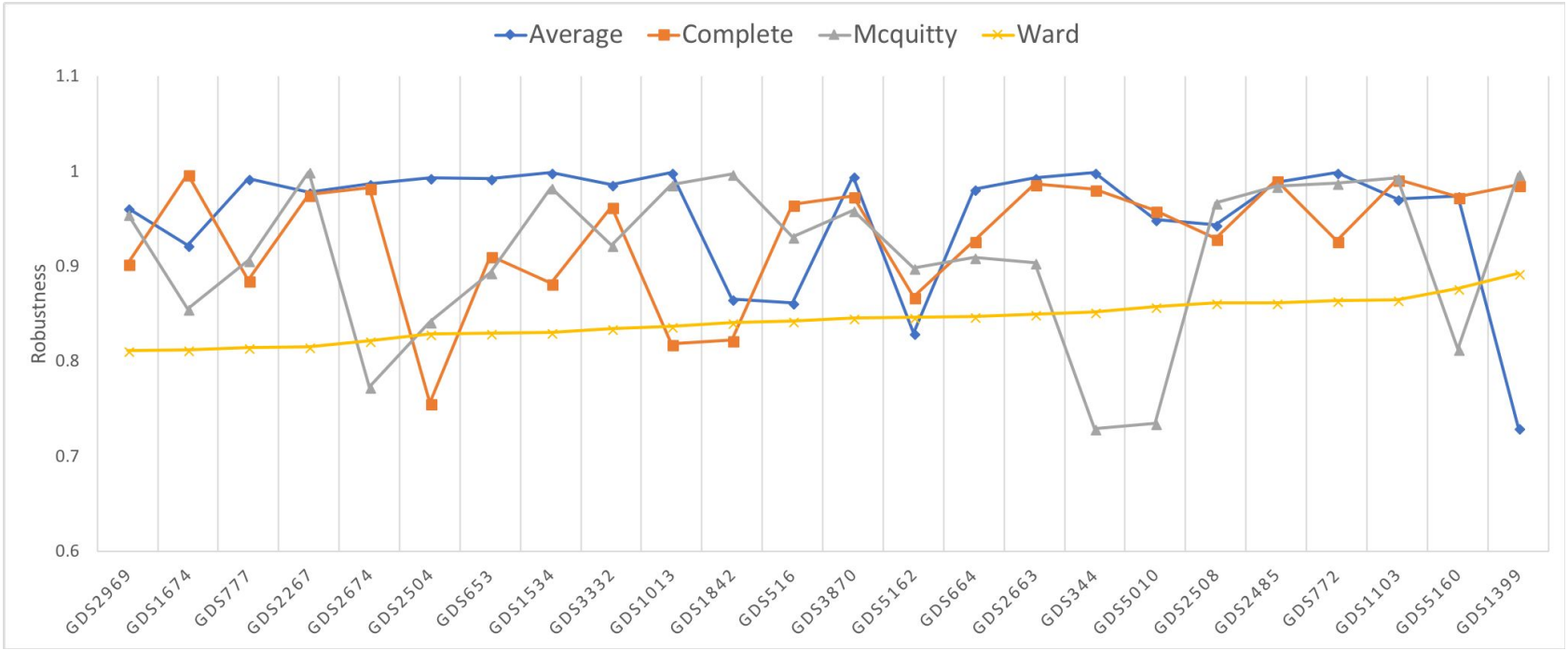
Clustering methods tested for robustness

Hierarchical	**Setting**	**Graph-based**	**Setting**
Average	Number of clusters	CLICK	Cluster homogeneity
Complete	Number of clusters	NNN	Min neighborhood size
Mcquitty	Number of clusters	Paraclique	Starting clique
Ward	Number of clusters	WGCNA	Power
Partitioning	**Setting**	**Neural network**	**Setting**
K-means	Number of clusters	SOM	Grid type/size
QT Clust	Max cluster diameter		

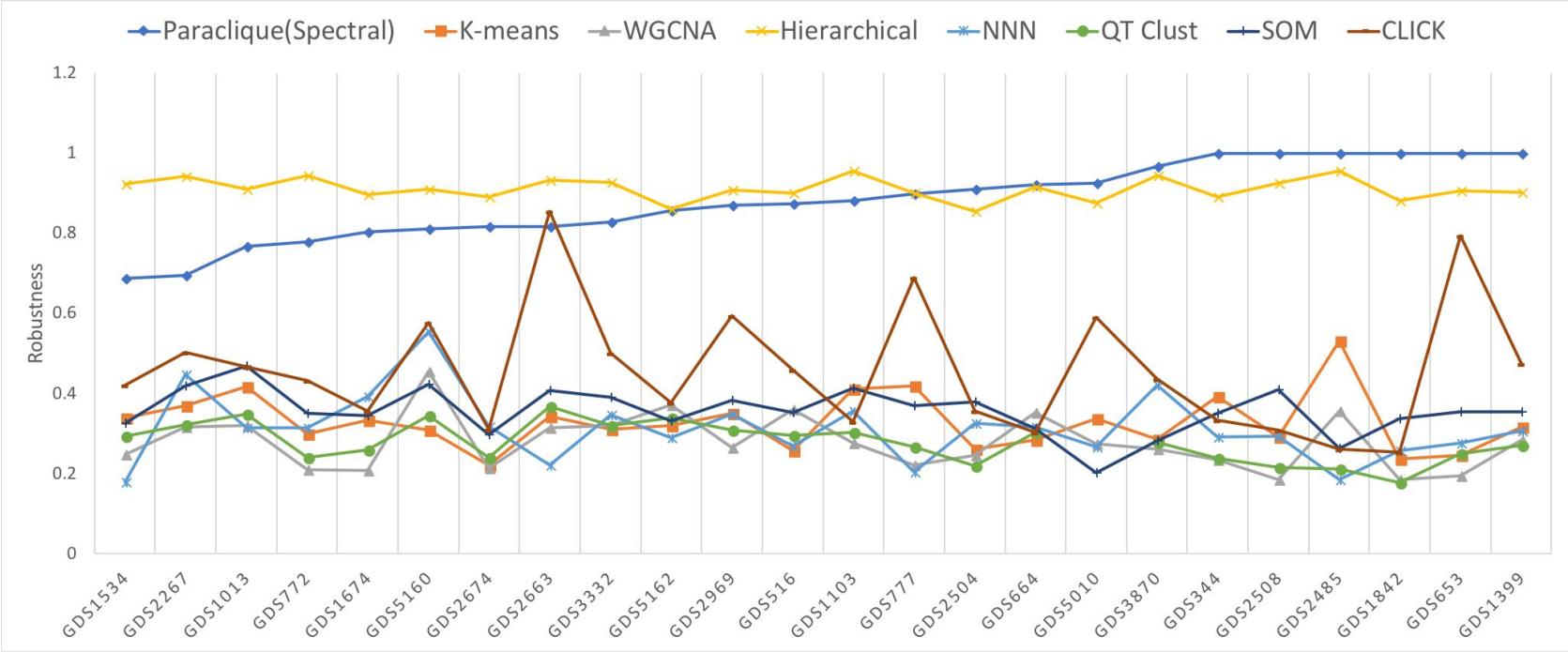
Data

24 gene co-expression datasets from Gene Expression Omnibus (GEO), including the species *Drosophila melanogaster*, *Escherichia coli*, *Mus musculus* , *Penicillium chrysogenum and Saccharomyces cerevisiae*.

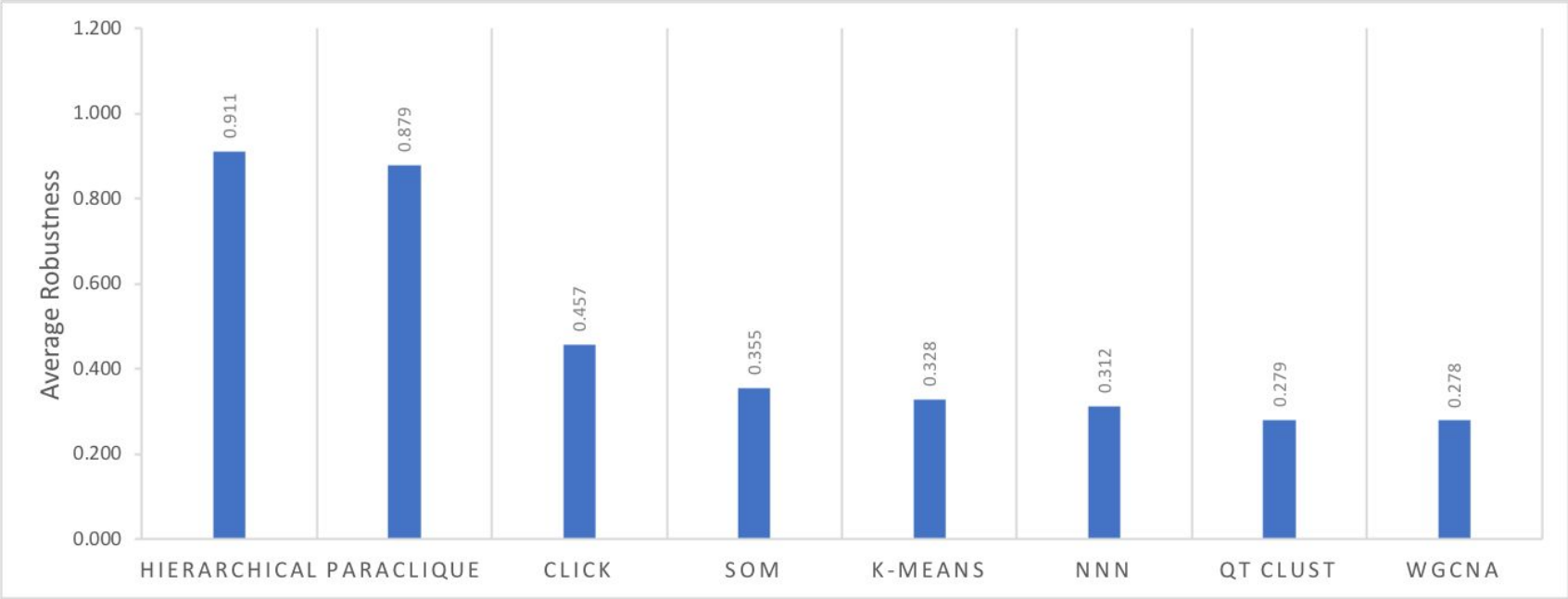
Robustness of four hierarchical algorithms on 24 transcriptomic datasets



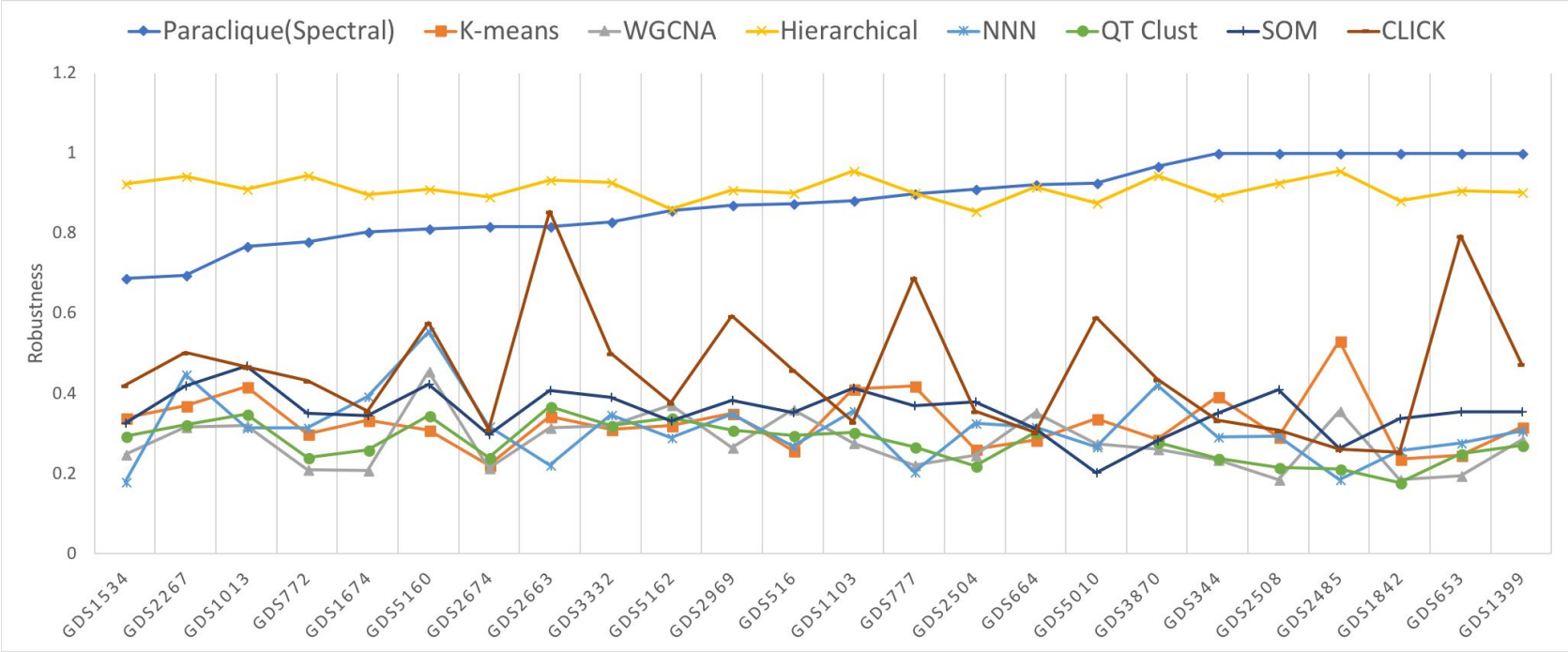
Robustness of all algorithms tested on 24 transcriptomic datasets



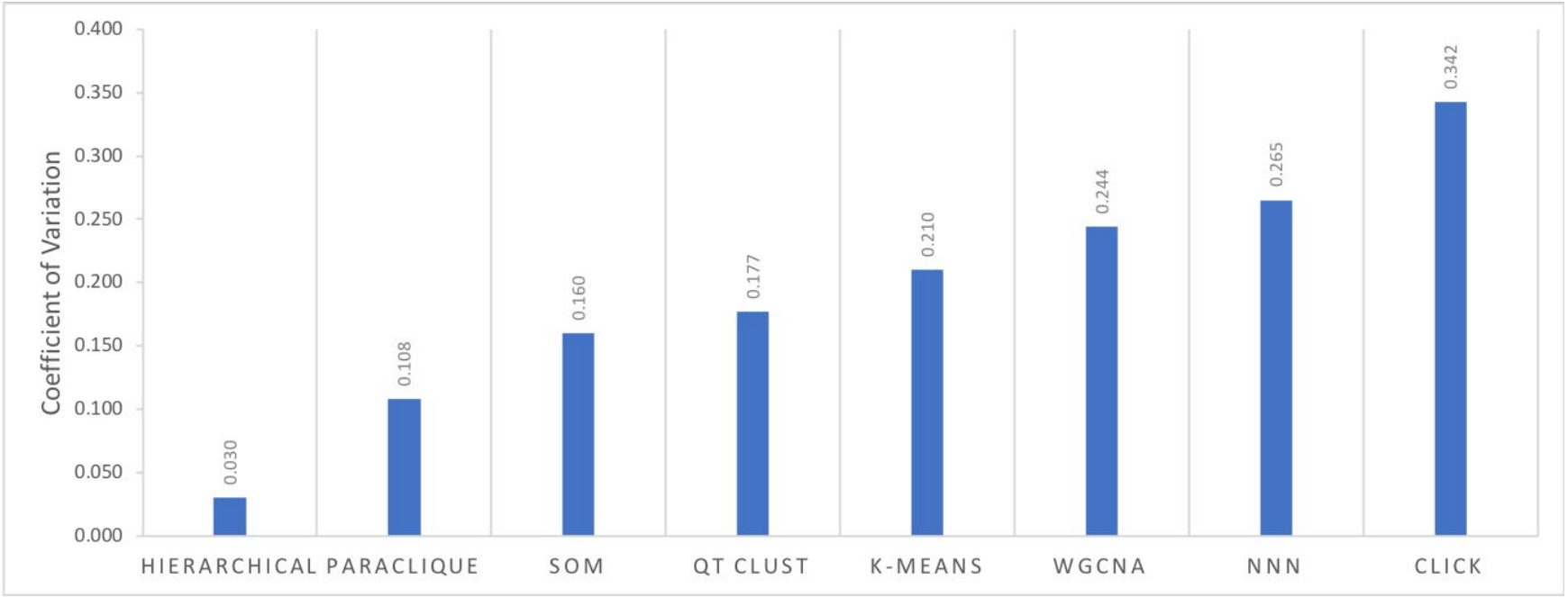
Average robustness of each algorithm



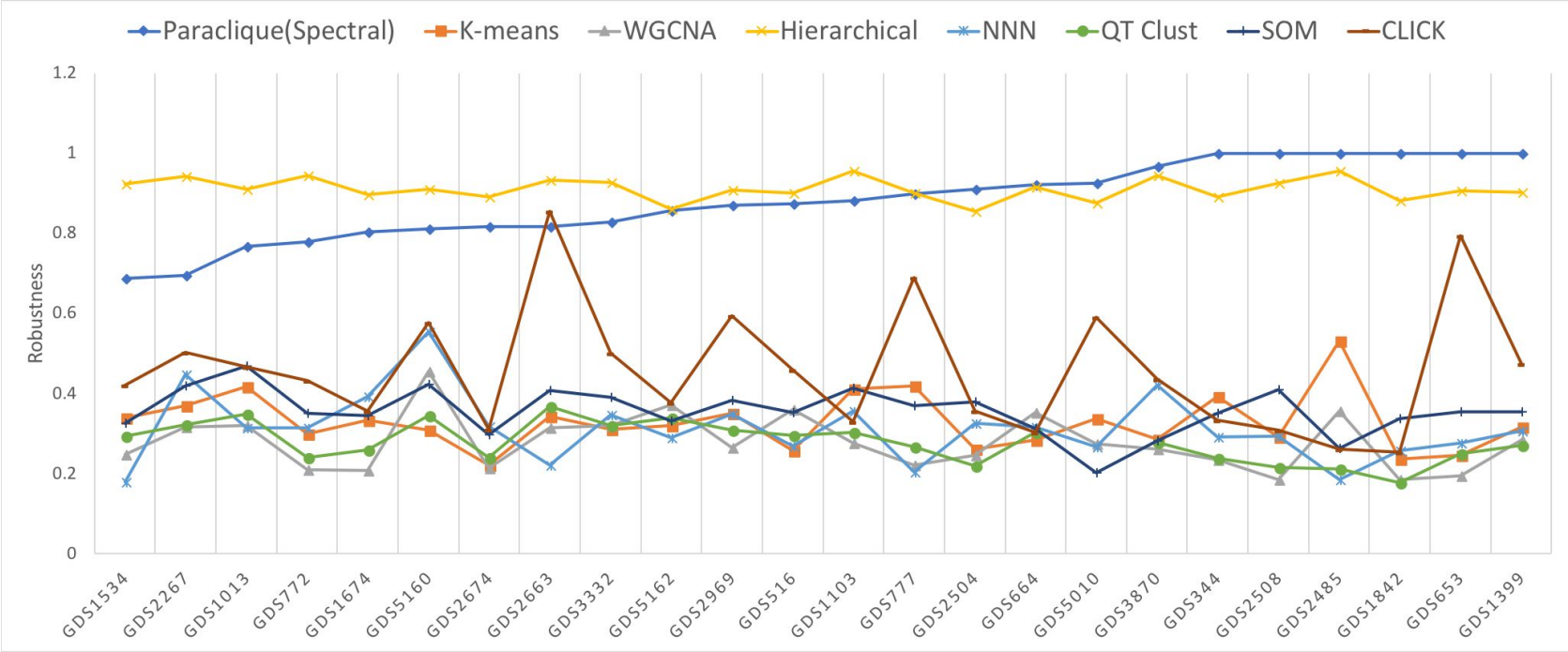
Robustness of all algorithms tested on 24 transcriptomic datasets



Coefficient of variation of each algorithm



Robustness of all algorithms tested on 24 transcriptomic datasets



Discussion

- Hierarchical methods display the highest overall robustness.

- WGCNA uses soft-power to construct its network, however, the topology of each weighted network changes with different powers, so that item pairs are not at all stable.

- For k-means, items often shift to different clusters as the number of clusters changes.

Discussion

- It is the similar situation for SOM, QT Clust, CLICK and NNN.

- For paraclique, the high robustness with different starting cliques is likely due in part to the fact that many of these cliques have significant overlap, at least on transcriptomic data.

Conclusions

- We have introduced a new metric, robustness, in an effort to provide the research community with an intuitive and informative measure of the stability and predictability of a clustering algorithm's behavior.

- Hierarchical methods generally exhibited the highest robustness on most datasets. Of the more complex non-hierarchical strategies, the paraclique algorithm yielded consistently higher robustness than other algorithms tested, approaching and even surpassing hierarchical methods on several datasets.

Directions for Future Research

- ➔ Clique has the propensity to produce overlapping clusters on biological data (genes, for example, are very often pleiotropic, and thus likely to belong to multiple clusters).

- ➔ We are studying an alternative notion of robustness that might be applied to virtually any non-overlapping clustering algorithm.

The Authors

Yuping(Allan) Lu

PhD student at UTK
Research Assistant at ORNL

Charles A. Phillips

Postdoc at UTK

Michael A. Langston

Professor at UTK

Thanks. Questions?