嘴砲 (x) 資料詮釋 (o) 的科學系列，之二

Networks Analysis and Visualization

June Lai

Review: Pathway Analysis

First hour:

- Definition: Set-based analysis (non-graphical)
- Underlying statistical hypotheses: Competitive vs. Self-contained
- Common statistical tests: Hypergeometric / T^2 / GSEA (KS)

  Recall *p*-values and multiple testing problem: FDR

Second hour:

- Common databases: GO / KEGG / MSigDB / STRING / Reactome
- Requirements:

  Dataset Gene set (with/without values)

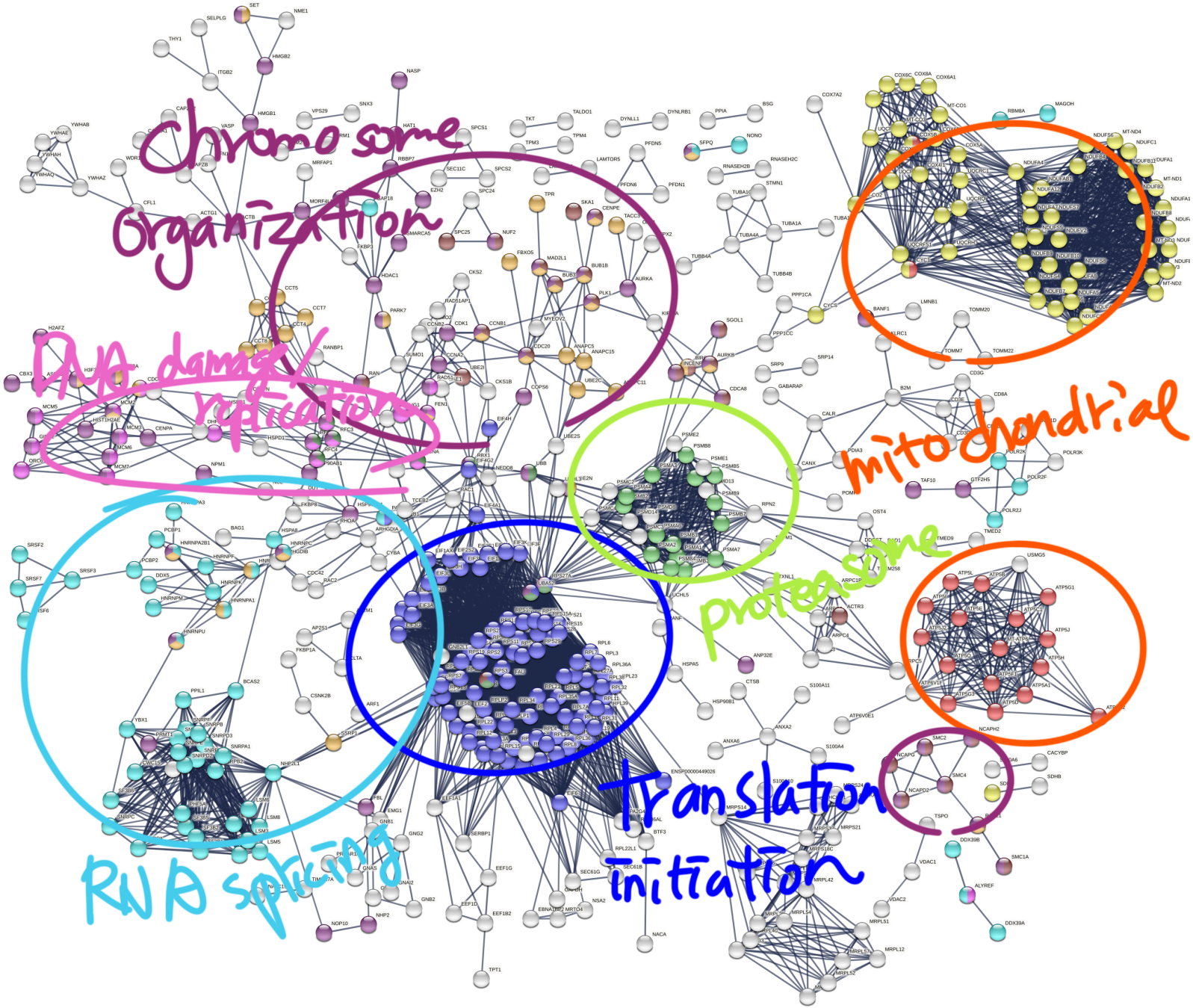
  Knowledge Database (hierarchical or not) & 「脳補之力」& GPT
- Demonstration: STRING / STAGEs / WebGestalt
- Afterword: The **art** of fine-tuning

2019 https://www.nature.com/articles/s41596-018-0103-9

2022 https://academic.oup.com/bib/article/23/3/bbac143/6572658

2025 https://www.mdpi.com/2673-6284/14/3/58

https://version-11-5.string-db.org/cgi/network?networkId=bQ32wU7U2Dbk



Outline

First hour:

- Definition: Graphical analysis (Topological approach)
- Common terms in graph theory
- Common network models
- Network centralities

Second hour:

- Clustering
- Demonstration: Cytoscape (local/GUI) / igraph (R/python)
- Visualization: Graphia / Arena3D / Circos
- Afterword: The **art** of fine-tuning

2020 https://www.frontiersin.org/articles/10.3389/fbioe.2020.00034/full

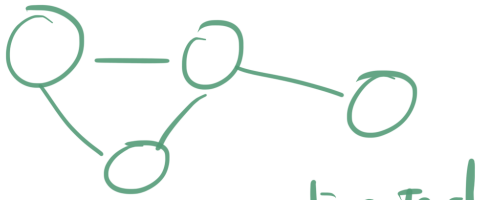
2024 https://link.springer.com/article/10.1007/s13721-024-00467-0

2025 https://www.mdpi.com/2673-6284/14/3/58

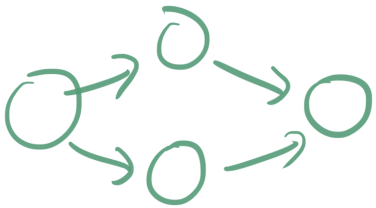
Common terms in graph theory

node (V)	or vertex, entities
edge (E)	relationship, directed or undirected, other edge types (e.g. enhance, inhibit)
graph (G)	V+E, connected, complete, weighted, bipartite
distance (d)	property between two nodes
degree (k)	node property, 有幾隻手/有幾個鄰居, (in-) 被幾隻手戳 / (out-) 伸出幾隻手
density	graph property, $\frac{\lvert E\rvert}{\lvert E_{\max}\rvert} = \frac{2\lvert E\rvert}{\lvert V\rvert(\lvert V\rvert-1)}$, edge 的濃度
clustering coefficient (C)	node property, $\frac{2\lvert e\rvert}{k(k-1)}$, 鄰居間 edge 的濃度
average clustering coefficient	graph property, $\frac{\sum C}{\lvert V\rvert}$

Common terms in graph theory



undirected

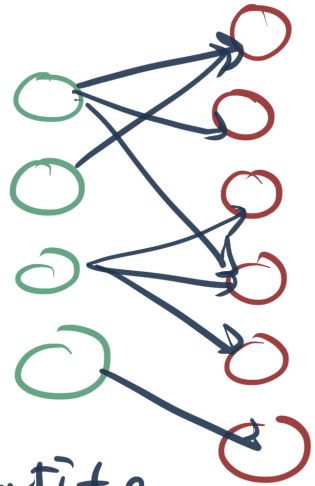


directed

(information flow)

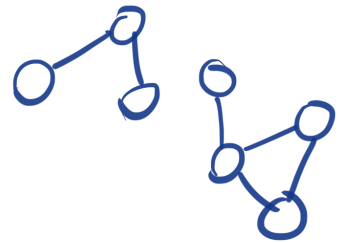
(time-course)

(causal)



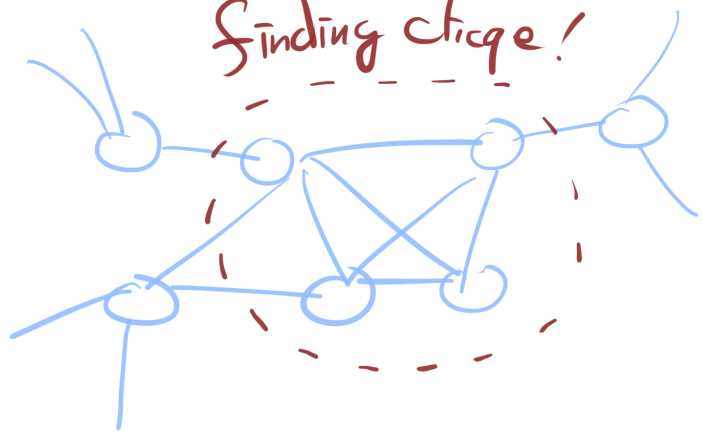
bipartite

weighted



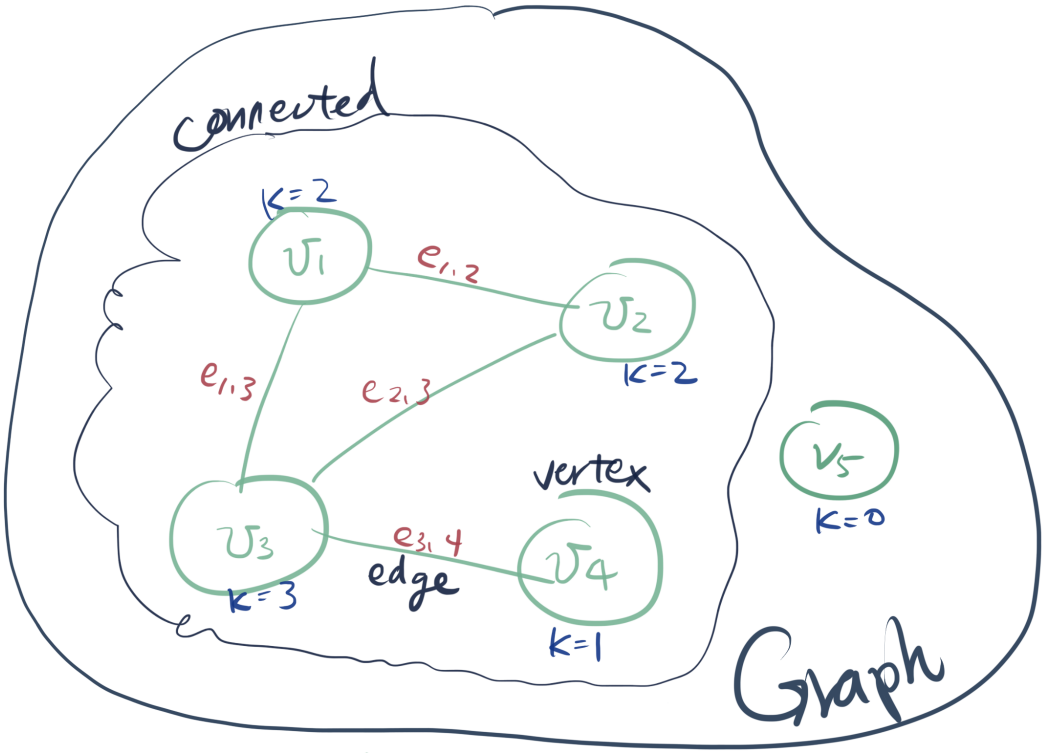
connected component

complete!



finding clique!

Common terms in graph theory



— undirected

→ directed

⊣ inhibit

⇝ enhance

$V = \{v_1, v_2, v_3, v_4, v_5\}$

$E = \{e_{1,2}, e_{2,3}, e_{1,3}, e_{3,4}\}$

$$\text{density} = \frac{|E|}{|E_{max}|} = \frac{2 \times 4}{5 \times 4} = 0,4$$

clustering coefficient

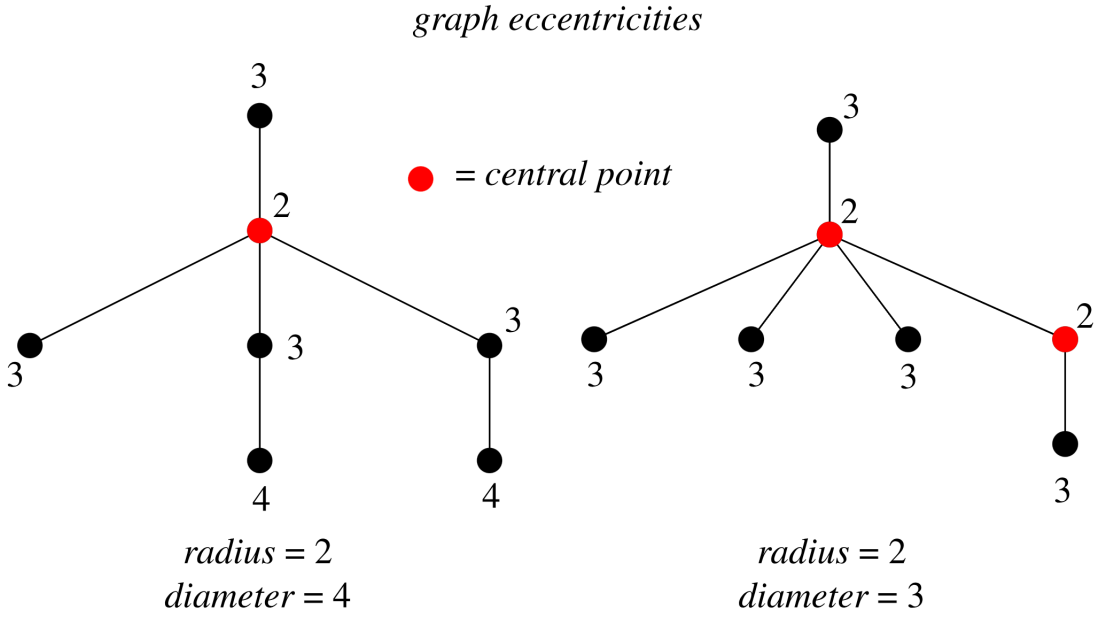
$$cc_1 = \frac{2 \times 1}{2 \times 1} = 1 = cc_2$$

$$cc_3 = \frac{2 \times 1}{3 \times 2} = \frac{1}{3} = 0,33$$

$$cc_4 = cc_5 = 0$$

$$\text{average} = \frac{1+1+0,33}{5} = 0,466$$

Other network centralities



eccentricity node property, 最遙遠的距離, $\frac{1}{\max d}$

diameter 整張圖「最長的」最遠距離

radius 整張圖「最短的」最遠距離

characteristic path length 整張圖「平均的」最遠距離

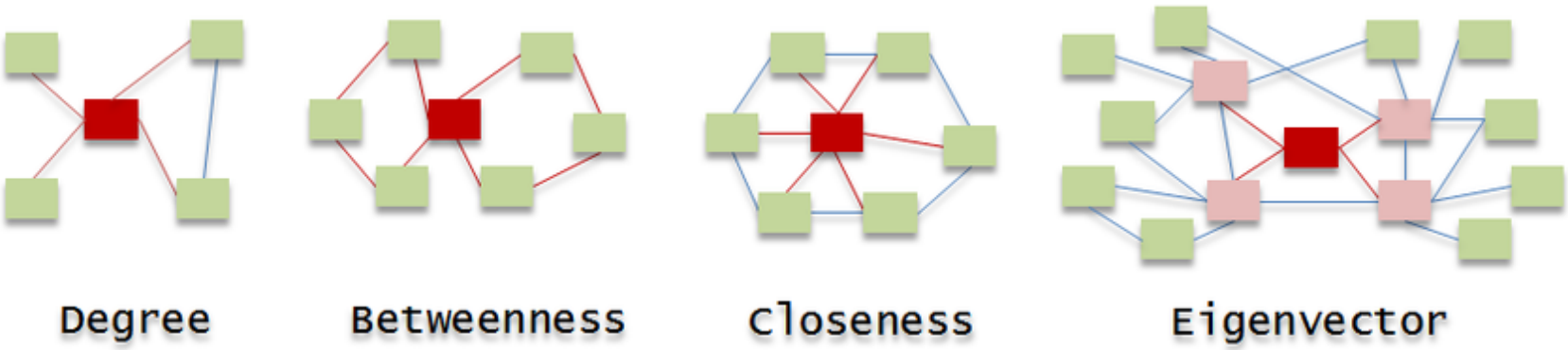
---

figure credit → https://mathworld.wolfram.com/GraphRadius.html

Other network centralities

closeness graph property, 世界小不小 (?), $\frac{1}{\sum d}$

betweenness node property, 高乘載強度 XD, $\frac{\#\text{有多少條經過此點}}{\#\text{任兩點最短路徑}}$



reference → https://en.wikipedia.org/wiki/Centrality

figure credit → https://www.researchgate.net/publication/296194726

Common network models

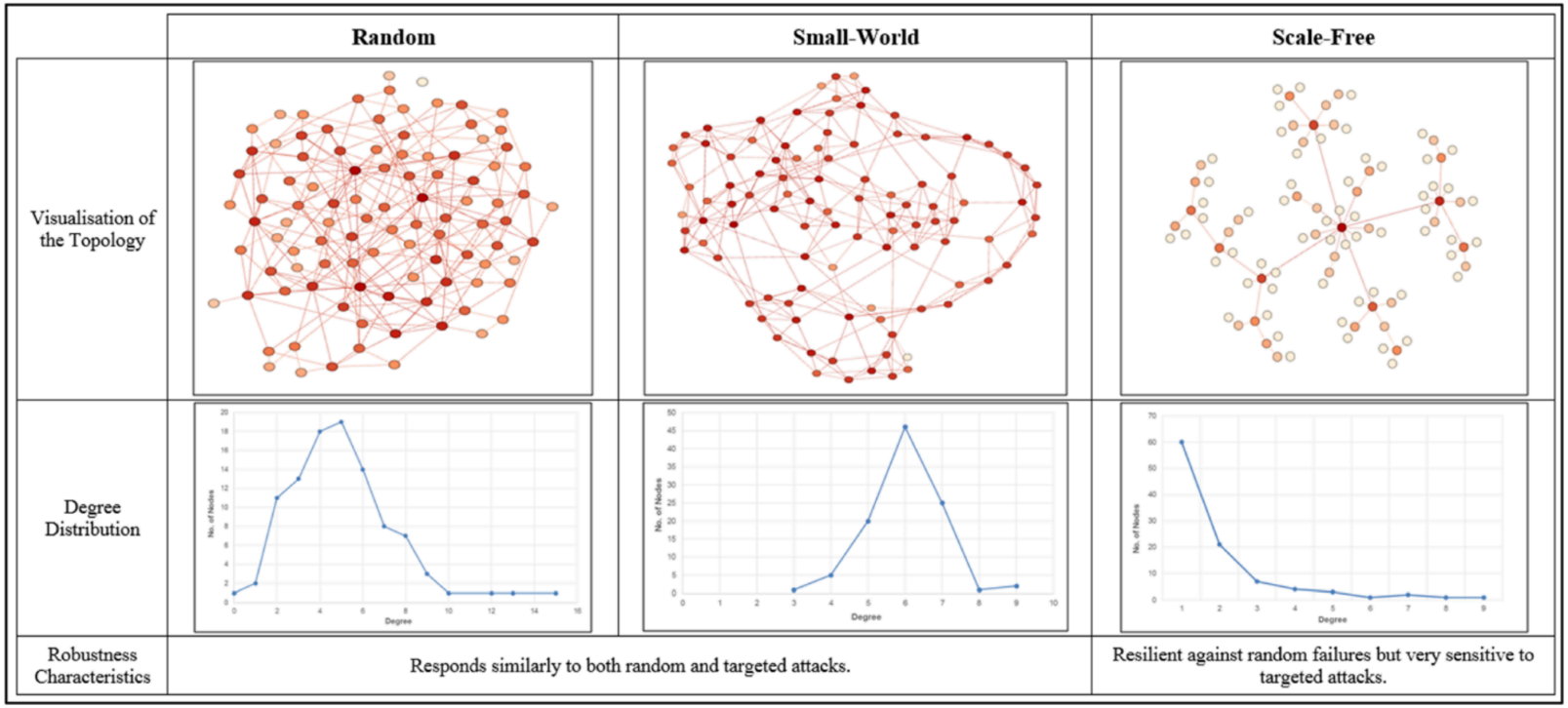


Fig. 1 Comparison of random, small-world and scale free networks. Topological structure of benchmark network models. Random and Small-world network topologies do not include hub nodes. In contrast, scale-free topologies are characterised by the presence of small number of highly connected hub nodes and a high number of feebly connected nodes. Presence of distinct hubs in scale-free networks make them more vulnerable to targeted attacks, compared to random and small-world networks

figure credit → https://appliednetsci.springeropen.com/articles/10.1007/s41109-017-0053-0

\(0w0)/ Intermission 中場休息 \(0w0)/

Outline

First hour:

- Definition: Graphical analysis
- Common terms in graph theory
- Common network models
- Network centralities

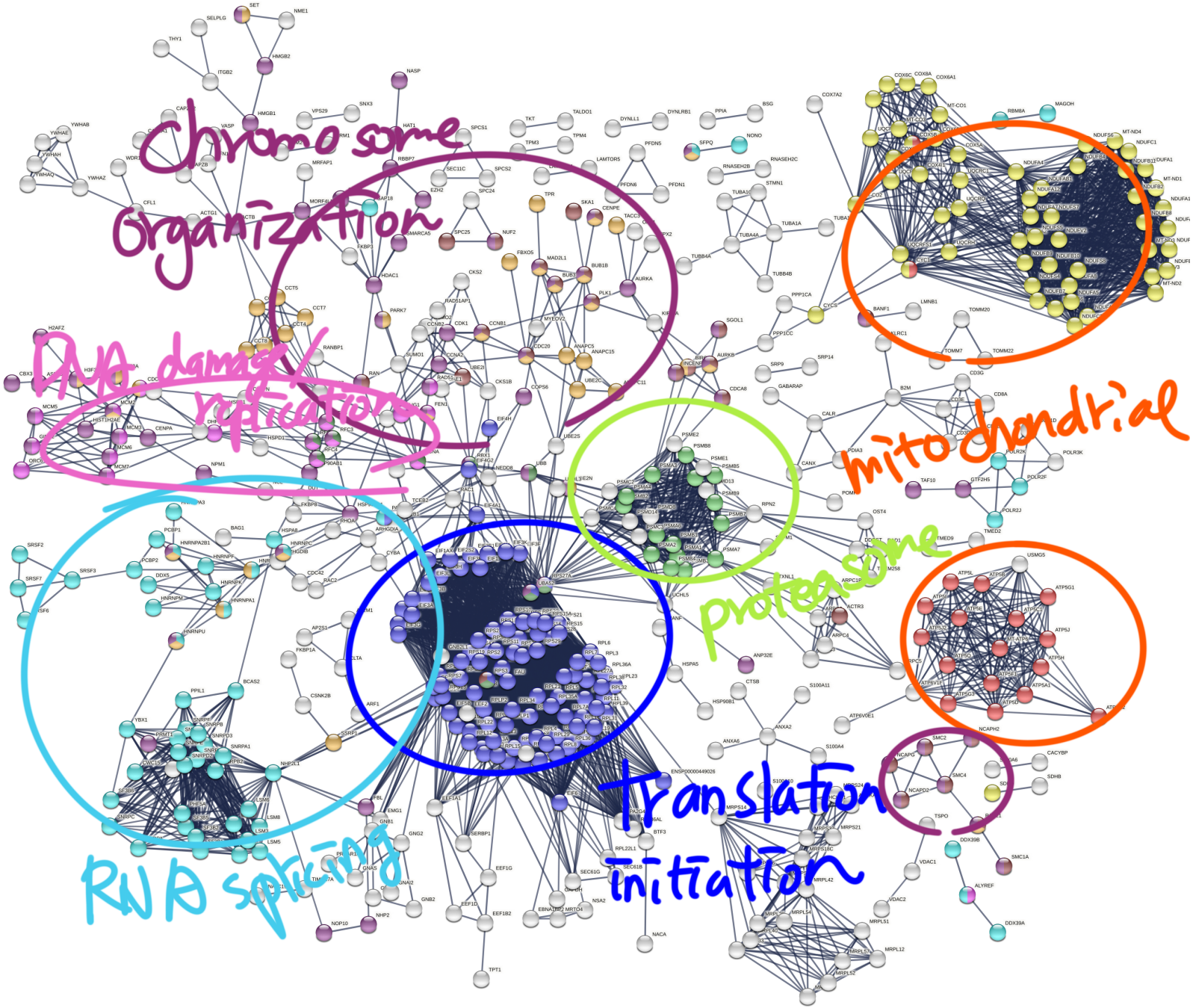
Second hour:

- Clustering: Key $\rightarrow$ definition of "distance"
- Demonstration: Cytoscape (local/GUI) / igraph (R/python)
- Visualization: Graphia / Arena3D / Circos
- Afterword: The **art** of fine-tuning

---

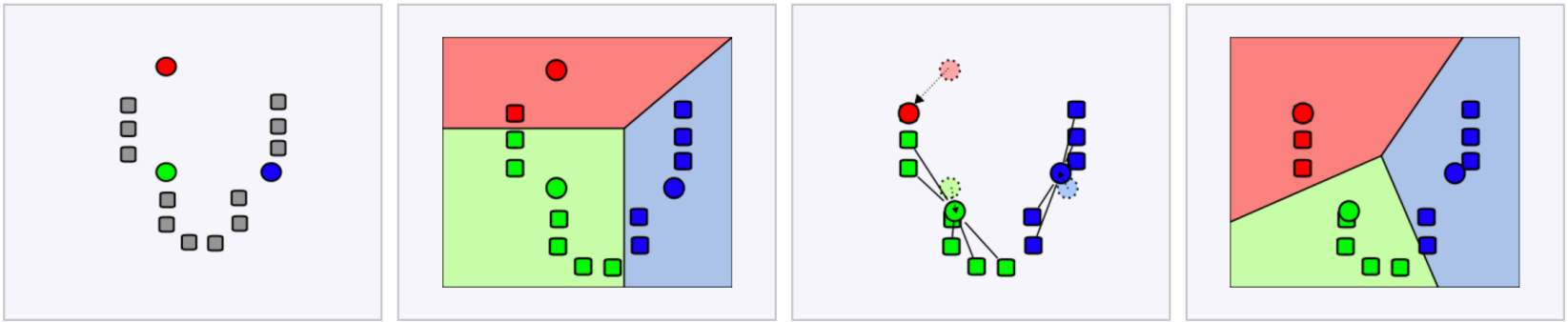
reference $\rightarrow$
https://www.frontiersin.org/articles/10.3389/fbioe.2020.00034/full

https://version-11-5.string-db.org/cgi/network?networkId=bQ32wU7U2Dbk



k-means clustering

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

---

figure credit → https://en.wikipedia.org/wiki/K-means_clustering
iterating → https://en.wikipedia.org/wiki/File:K-means_convergence.gif

Demonstration / Visualization

STRING v12 (web)

- Clustering demo
- Export → Cytoscape → further analysis

Cytoscape v3.10.3 → v3.10.4 (local/GUI)

- Visualization, Layouts, can be manually adjusted
- Common network statistics / enrichment analysis
- APP store!!!
- R/js interfaces (R package: RCy3)

Cytoscape manual → https://manual.cytoscape.org/en/stable/
Cytoscape APP store → https://apps.cytoscape.org/
Cytoscape R package → https://cytoscape.org/RCy3/

Demonstration / Visualization

igraph (R/python)

- Comprehensive graphical statistics and management
- Visualization, various layouts, but the output can hardly be manually adjusted

Arena3D (web shiny app/cytoscape app)

- Multi-layered 3D for multi-omics or time-course

PAVER (R/web shiny app)

- **P**athway **A**nalysis **V**isualization with **E**mbedding **R**epresentations

---

igraph manual → https://igraph.org/r/doc/
Arena3D → https://pavlopoulos-lab-services.org/shiny/app/arena3d
PAVER → https://cdrl.shinyapps.io/PAVER/

我私心的嘴砲 best practice XD

1. 先畫個全景圖看看資料粘的好不好 (STRING/Cytoscape)
   - knowledge-based, databases required
   - data-driven, sample size required

   資料太少 or 太散 → 考慮延伸 network
2. trimming: node/edge/small subgraph
   - 測試各種參數組合
3. 丟上 STRING 看看 (no more than 3000 entities)
   - 嘗試各種 score sources 跟 cut-offs
   - Best set size is around 300 to 1000
4. clustering 看看結果是不是大約跟研究問題方向一致
5. 重複步驟 2-4 直到滿意
6. enrichment analysis (注意 null background)
7. 挑自己想要的結果畫圖 XD