Social Network Analysis

Group

Dr. Chun-Hsiang Chan

Department of Geography,
National Taiwan Normal University

Photo credit: midjourney

Outline

- Group Concept
- Cohesion
- Component
- Clique
- k-plex
- k-core
- Modularity
- Louvain’s Method
- Girvan-Newman Method
- Infomap
- k-core Decomposition
- Paper Reading
- References

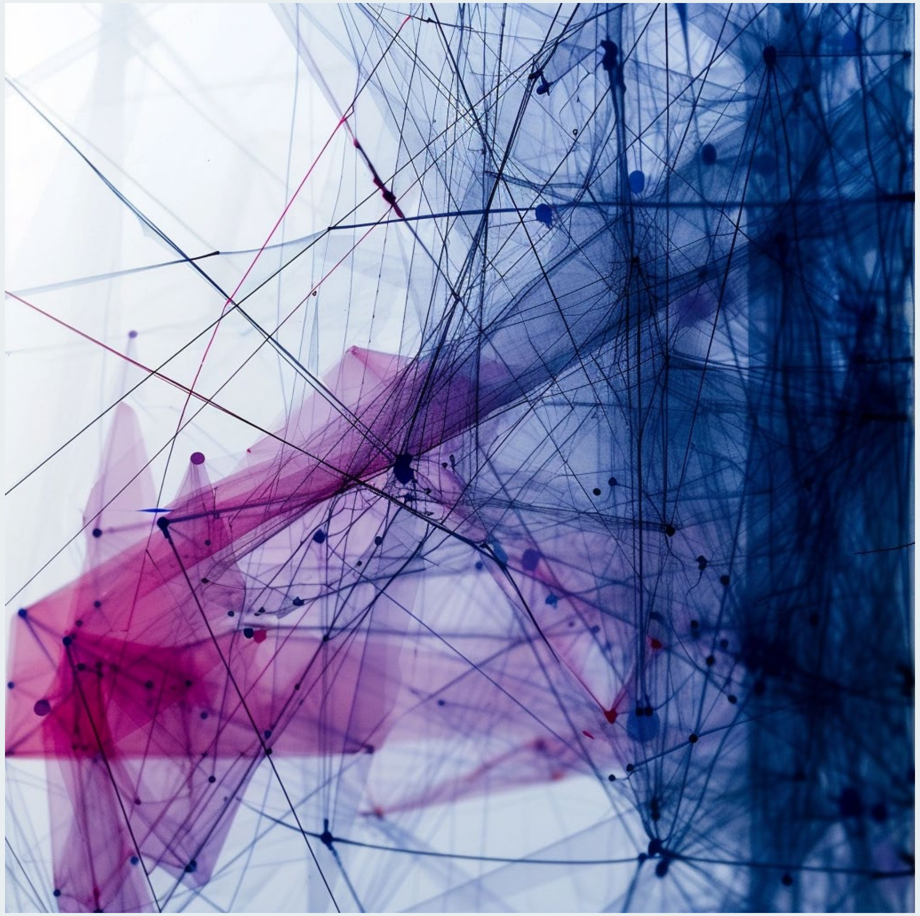
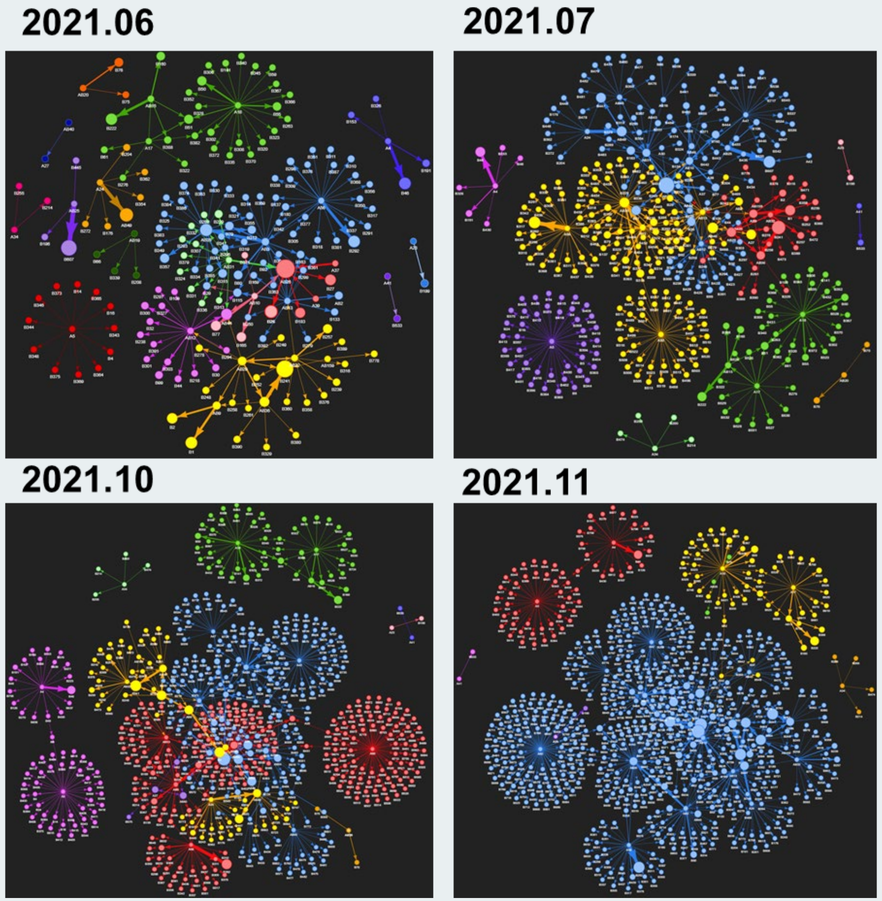


Photo credit: midjourney

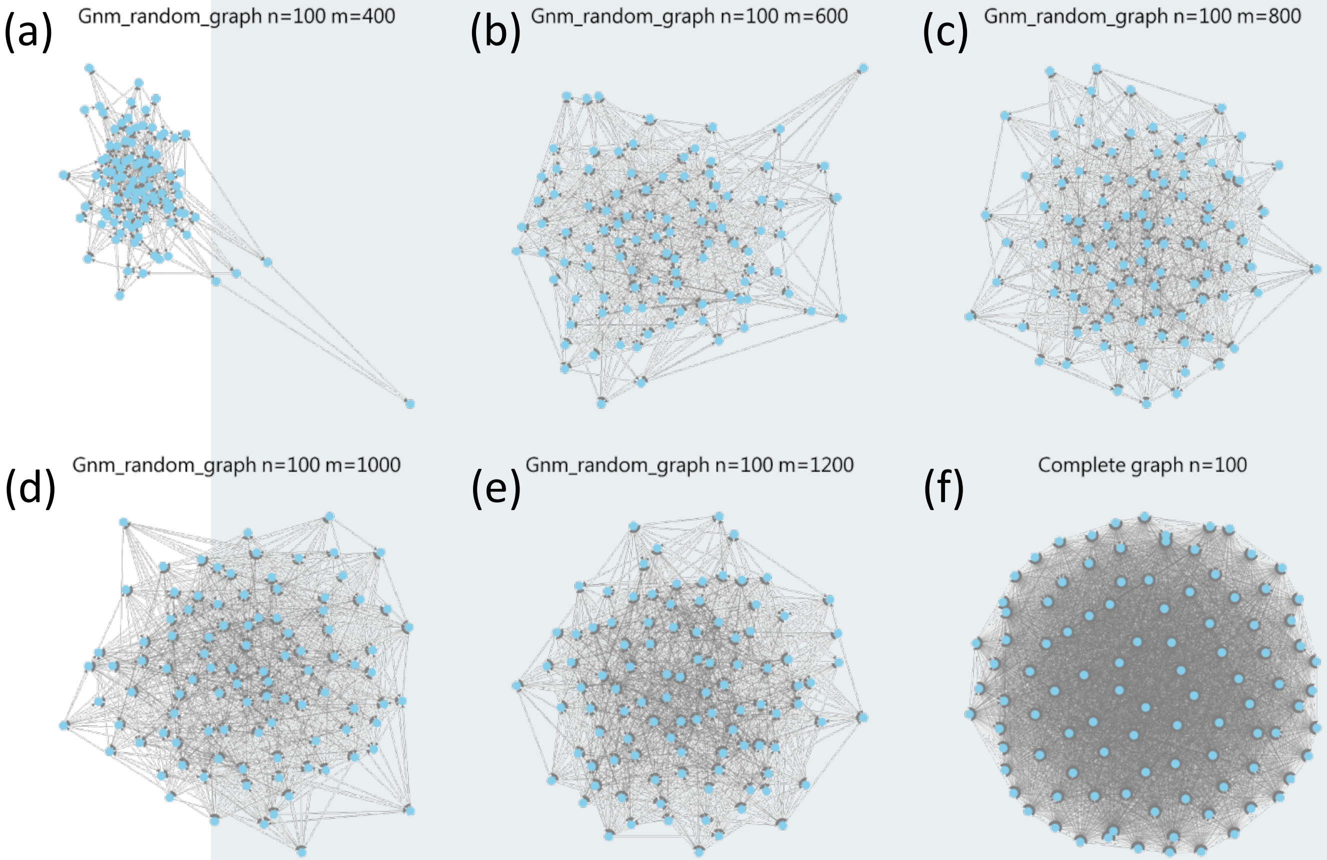
The Group Concept in Network

- **Group** indicates a strong internal cohesive relationship between actors (nodes).
- However, the levels of cohesion in the network could be differentiated into several types of grouping definitions, such as component, clique (n-clique), k-core, and k-plex.
- Moreover, we can extract the community structure in the network, including Louvain and Infomap.



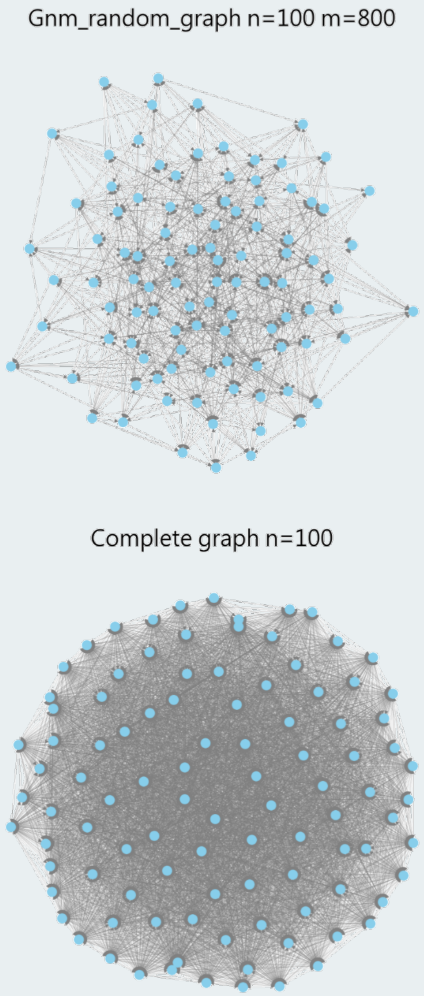
Graph Type

- If all nodes within a graph are connected to each other, then the graph is called a complete graph.
- An example of a complete graph is shown in the (f).



Problem

– If I have a complete network, as shown in the right figure, you will expect the differences between the complete network and the G_{nm} random network.
  – Diameter
  – Geodesic distance



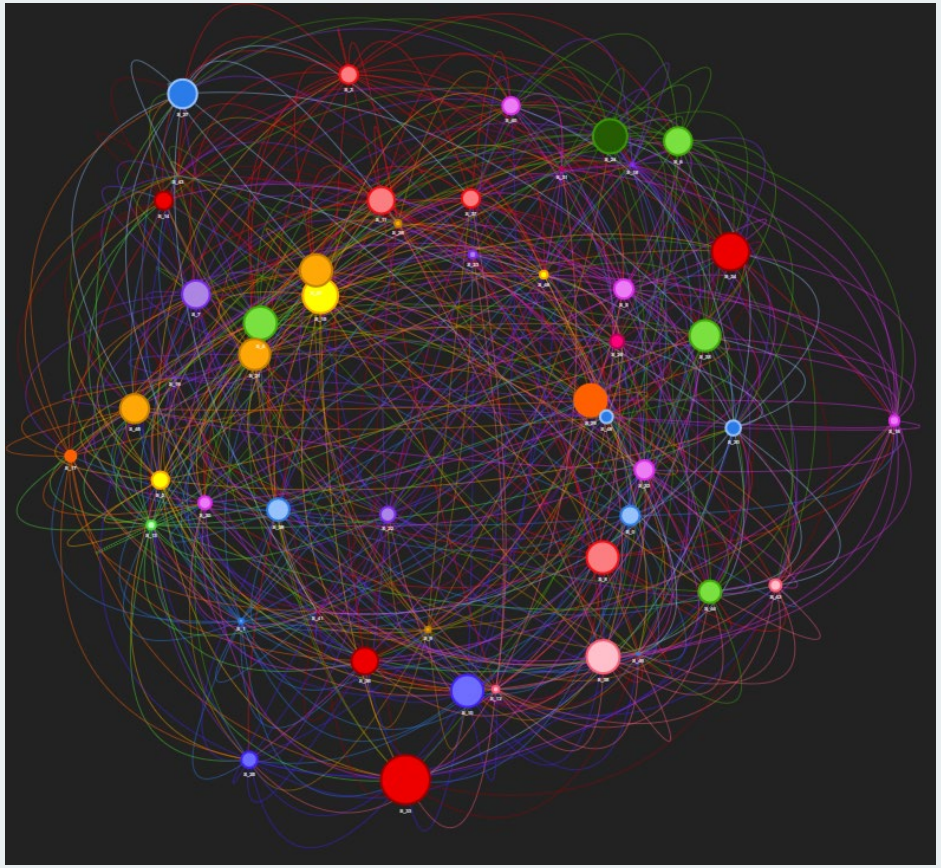
Cohension

– Usually, all members within a cohesive subgroup have a relatively strong, highly dense, and high-frequency relationship with each other.
– In other words, they may have some intrinsic and homogeneous attributes within a group.
– For a cohesive subgroup, it is equipped to be decentralized, distributed, equally positioned, and robust to unilateral action.
– In practice, the members within a group or two group members may have some characteristics that formulate this structure; for instance, they have the same targets or ideas within a group.

Cohension

– **Cohesion** has four characteristics:

1. Mutuality of ties
2. Members within a group have good closeness and reachability
3. A higher number of tie frequency between members within a group
4. A lower number of tie frequency between members from different groups



Subgroup Identification

Here, we give five definitions to differentiate subgroups in a given graph.

– Based on the **connectedness relationship** ➔ component

– Based on **mutual reciprocity** ➔ clique

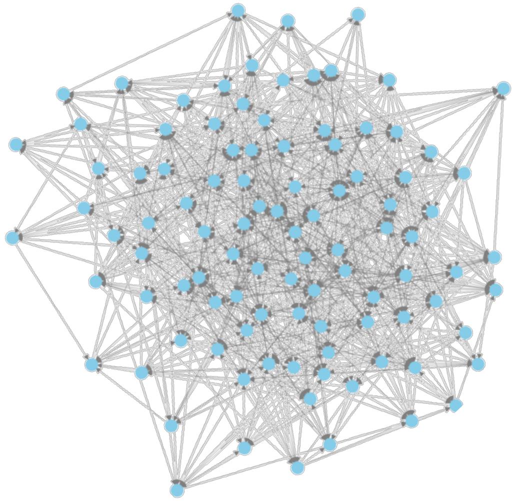
– Based on **reachability** ➔ n-clique

– Based on **the accessibility of nodes** ➔ k-core and k-plex

– Based on **the frequency of connections** ➔ community

Components

- The **component** is a subgroup/ group/ subset/ set of nodes that every node has at least one connection to other nodes within the component.



Weakly and Strongly Connected Components

– **Connected component:**

Sets where all nodes can reach all other nodes via any direction (i.e., monodirectional or bidirectional).

Ex: {A, B, C, D, E} and {F, G}

– **Strongly connected component:**

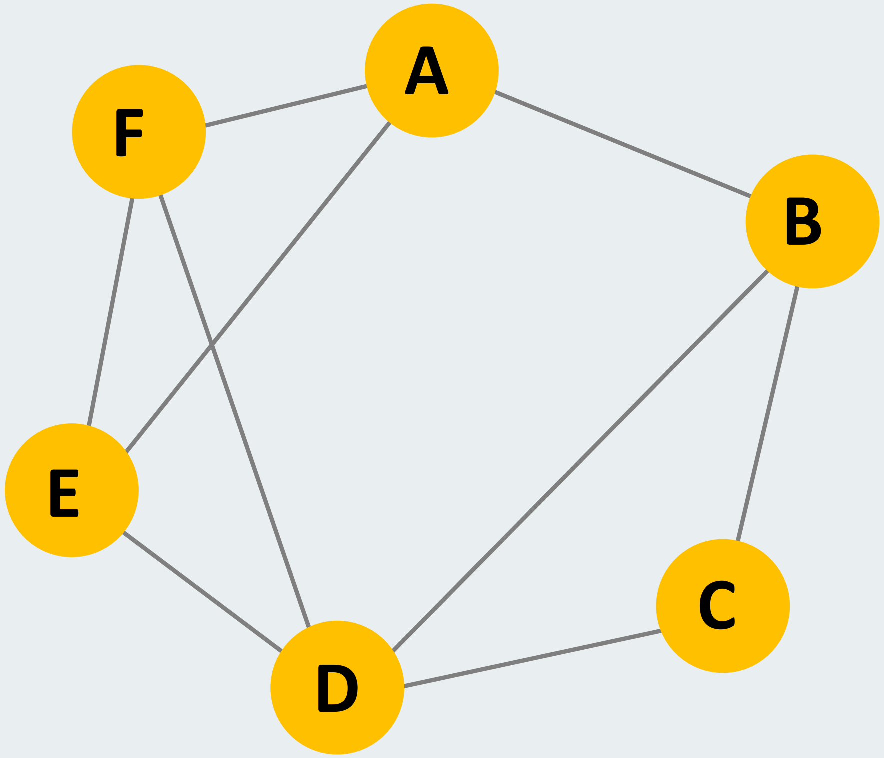
Sets where all nodes can reach all other nodes in both directions, but not necessarily directly.

Ex: {A, B} and {C, D, E}

Cliques

- **Clique** is the strictest definition of grouping, where all nodes within a clique should connect (mutuality).

- How many cliques exist in the graph?
- Ex: {A, D, E}, {B, C, D}, {D, E, F}



Cliques

– The dilemma and challenge of clique in the real-world situation

1. **Too strict definition:** The clique will disappear depending on the strict definition, causing one node or edge loss.
2. **Sparse network:** Most real-world networks are sparse networks; in other words, it is difficult to formulate a clique within the network. If they exist, these cliques are very small.
3. **Limitation in data collection:** During the data collection, if you have set the number of connected nodes, it will also limit the large number of nodes within a clique.
4. **Equal physical meaning:** We cannot differentiate the core or peripheral group from cliques.

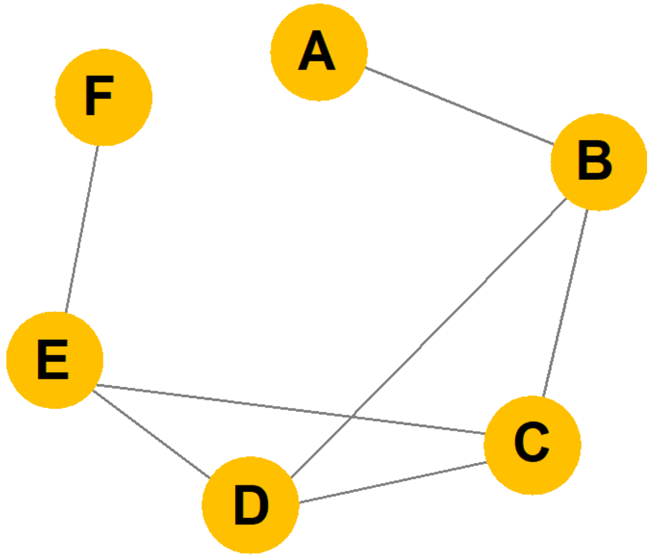
Cliques

– **How to solve this problem?**

- Reachability or diameter to relax the definition of cliques
    - Some important organizations or groups require a key person (intermediary) to link all members in real-world situations.
    - The members within the group do not have to connect directly with each other.
    - This parameter is called “n-cliques.”
- Using degree to define a subgroup
    - K-core and k-plex

N-cliques

– **N-cliques** is a subgroup defined by any two nodes with a shortest distance less than n.

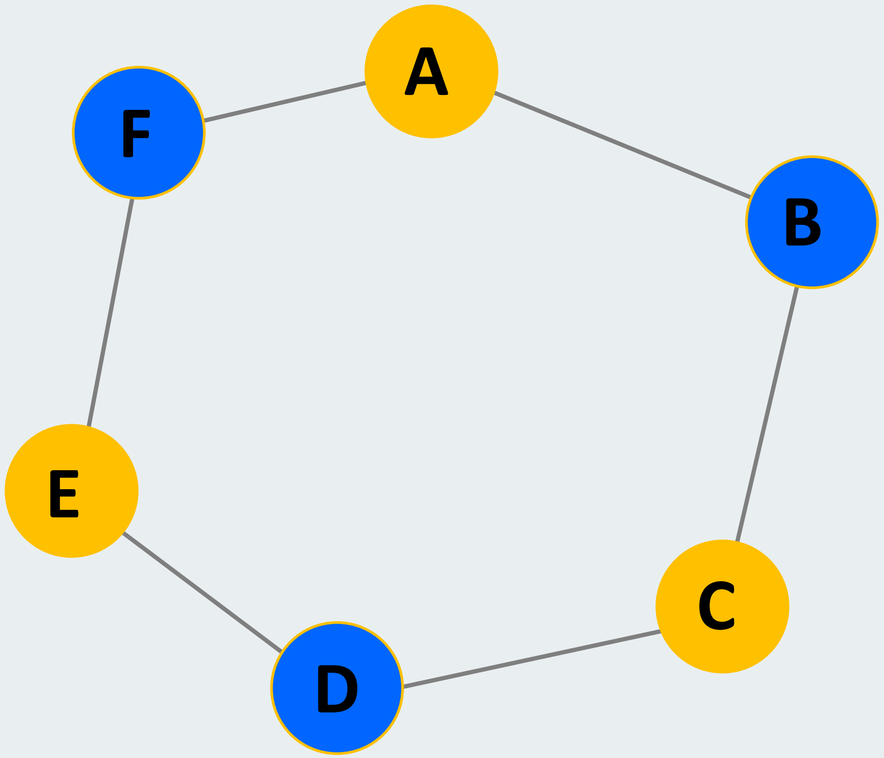


1-clique: {B, C, D}, {C, D, E}
2-clique: {B, C, D, E}
3-clique: ???

Problems with N-cliques

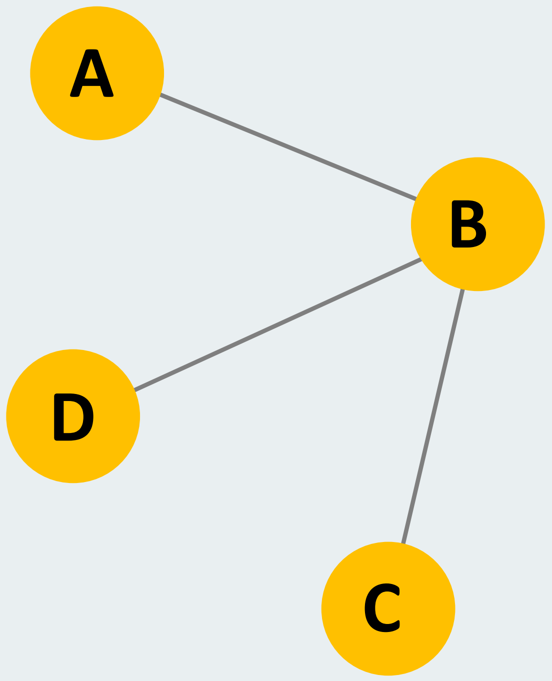
– Sometimes, n-cliques do not represent a group …

– Here, we give an example: there are two 2-clique in the following graph; however, blue nodes belong to the same 2-clique, but none of them are adjacent.



Problems with N-cliques

- The n-cliques exist as a robustness and vulnerable problem …
- Here, we give an example on the right-hand side. This is a 2-clique. As you can observe, this clique is vulnerable because it will not be a 2-clique since removing the B node.
- On the contrary, if it is a robust structure, it will not be affected by removing one or two nodes.



Using Degree to Define a Subgroup

– Some scientists mentioned that we could utilize reachability to define the subgroup, which emphasizes a compact group constructed with direct or indirect connections.

– Other scientists thought only direct connections could be considered in some network process, indicating the adjacent relationship rather than distance or path.

– Hence, here, we use the adjacency concept to understand the social process of direct contact.

– Parameters: k-core and k-plex

K-plex

- **k-plex** contains g_s nodes, where each node has at least $(g_s - k)$ adjacent nodes within a subgroup.

$$d_s(i) \geq (g_s - k)\ for\ all\ n_s \in N_s$$

where $d_s(i)$ is the degree of a node within the subgroup g_s

- For example, 2-plexes with n=7 will find all groups of size seven in which each person is connected to at least five (n-k) others in the group.

K-plex

- When $k = 1$, 1-plex equals clique because each node can connect at most $g_s - 1$ adjacent node.
- When k is larger, each node within the subgroup can be adjacent to more non-adjacent nodes (loose the condition while k is larger).
- k-plex is more robust than clique because the adjacent relationship of each k-plex node has several nodes. It does not break down, while we remove a one or few nodes.

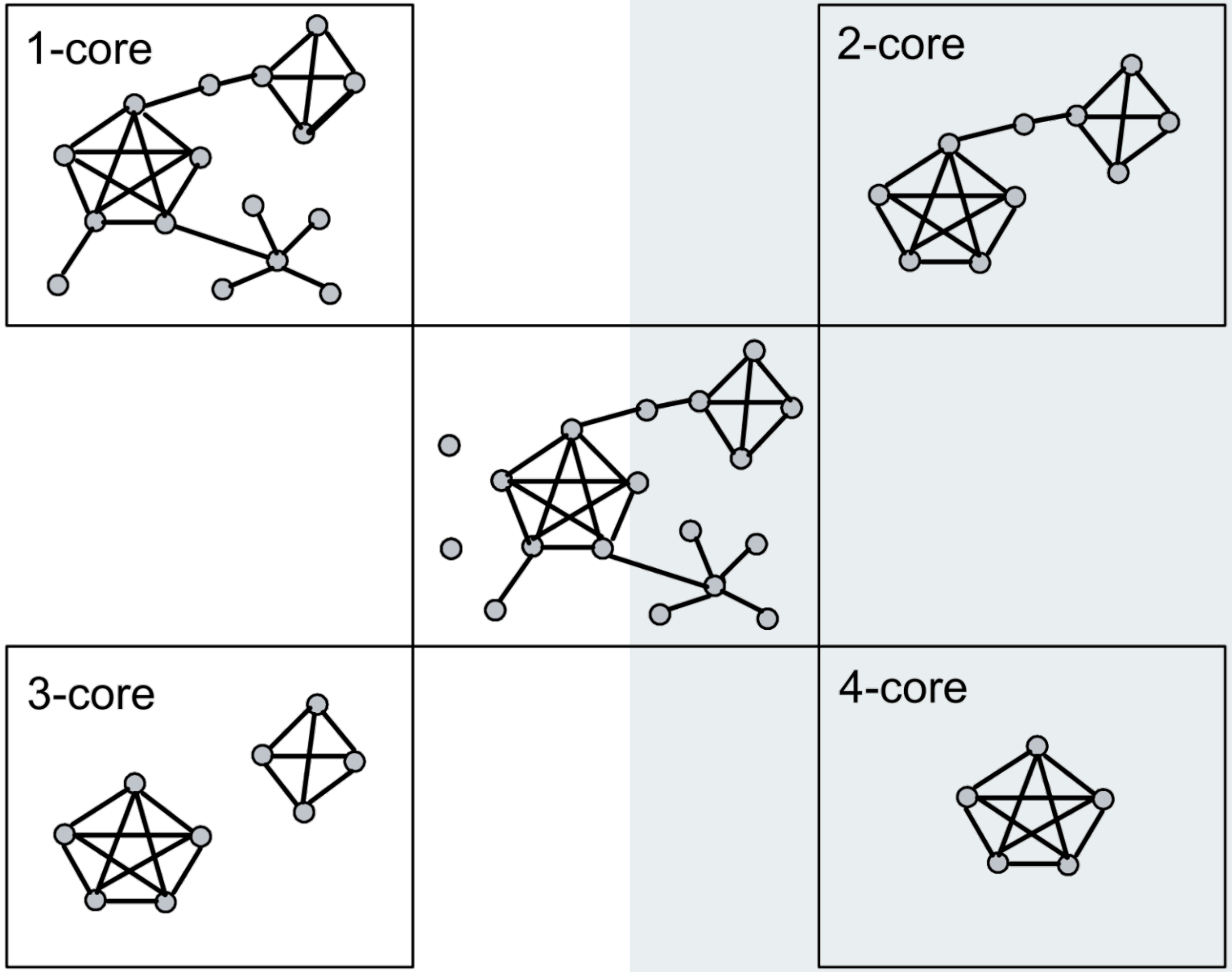
K-core

– K-core is a subgraph in which each point is adjacent to at least k other points.

– For each subgroup, k-core will be defined as the following condition.

$$d_s(i) \geq k, \forall n_i \in N_s$$

K-core



Zhang, G. Q., Zhang, G. Q., Yang, Q. F., Cheng, S. Q., & Zhou, T. (2008). Evolution of the Internet and its cores. *New Journal of Physics*, *10*(12), 123027.

Differences between k-plex and k-core

- k-plex defines that each node can have at most k missing edges
  - The smaller the k, the stricter
  - k = 1 ➔ the strictest: clique
- k-core defines that each node can have at least k edges
  - The larger the k, the stricter

Modularity

- **Modularity** is a measure of network or graph structure that measures the strength of the division of a network into modules (also called groups, clusters, or communities).
- Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

Community Detection

– **Our objective:**

– Find optimal clusters

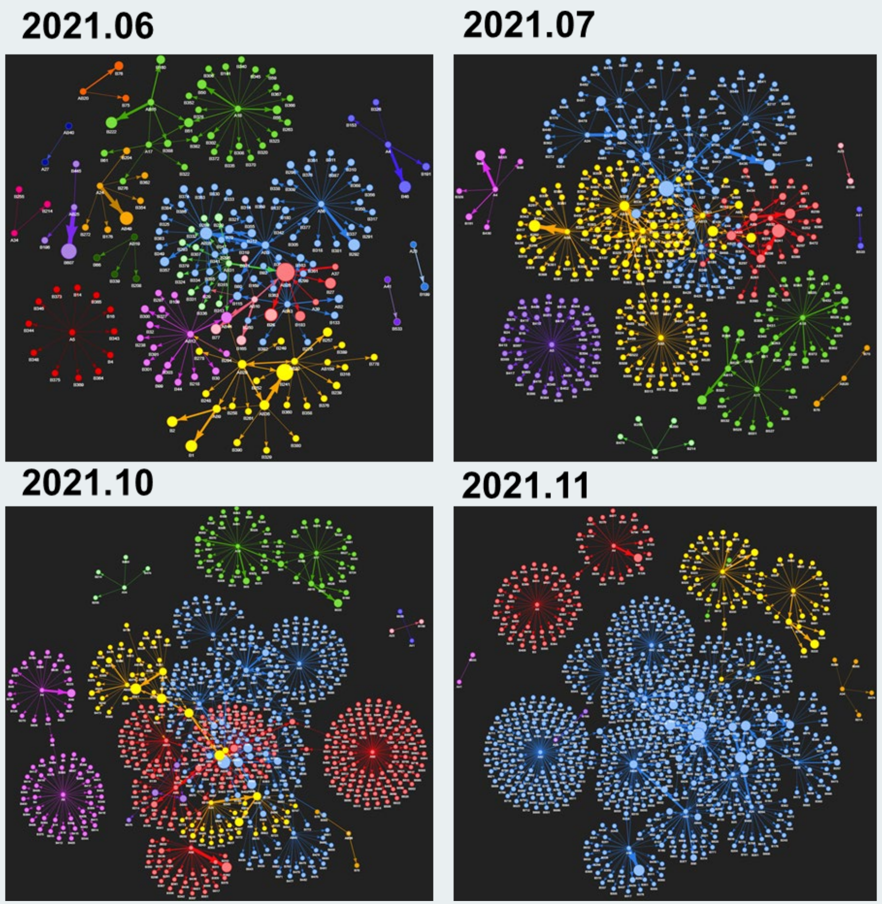
  – Intra-cluster: high frequent connections

  – Inter-cluster: low frequent connections

– Notice:

  – Weighted or non-weighted

  – Directed or undirected



Louvain's Method

– Given a weighted graph, the modularity is defined as…

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

where
A_{ij} represents the edge weight between nodes i and j
k_i and k_j are the sum of the edges attached to nodes i and j, respectively.
m is the sum of all of the edge weights in the graph
c_i and c_j are the communities of the nodes
δ is Kronecker delta function ($\delta(x, y) = 1$ if $x = y$, 0 otherwise)

Louvain's Method

– Based on the above equation, the modularity of a community c can be calculated as

$$Q_c = \frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2$$

Σ_{in} is the sum of edge weights between nodes within the community c (each edge is considered twice).

Σ_{tot} is the sum of all edge weights for nodes within the community (including edges which link to other communities).

Girvan-Newman Method

RESEARCH ARTICLE | APPLIED MATHEMATICS

Modularity and community structure in networks

M. E. J. Newman Authors Info & Affiliations

Edited by Brian Skyrms, University of California, Irvine, CA, and approved April 19, 2006

June 6, 2006 | 103 (23) 8577-8582 | https://doi.org/10.1073/pnas.0601602103

Girvan-Newman Method

– Girvan-Newman's method could produce un-overlapping community results to detect the performance of community detection results.
– The percentage of within-group ties to between-group ties is calculated.

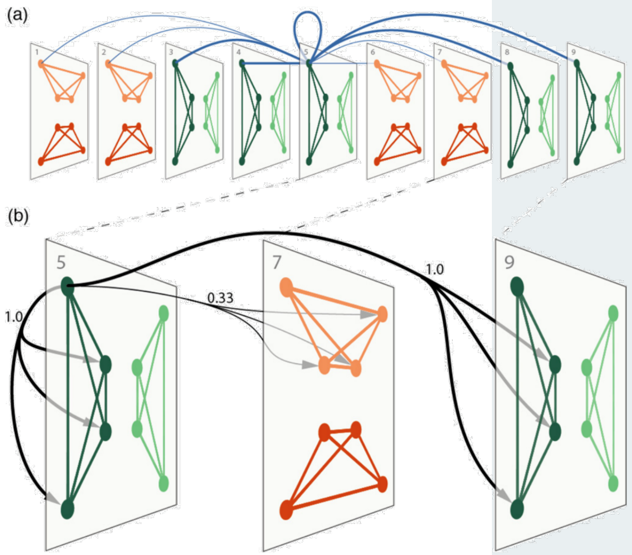
$$modularity = Q = \sum_i (e_{ii} - a^2) = Tre - ||e^2||$$

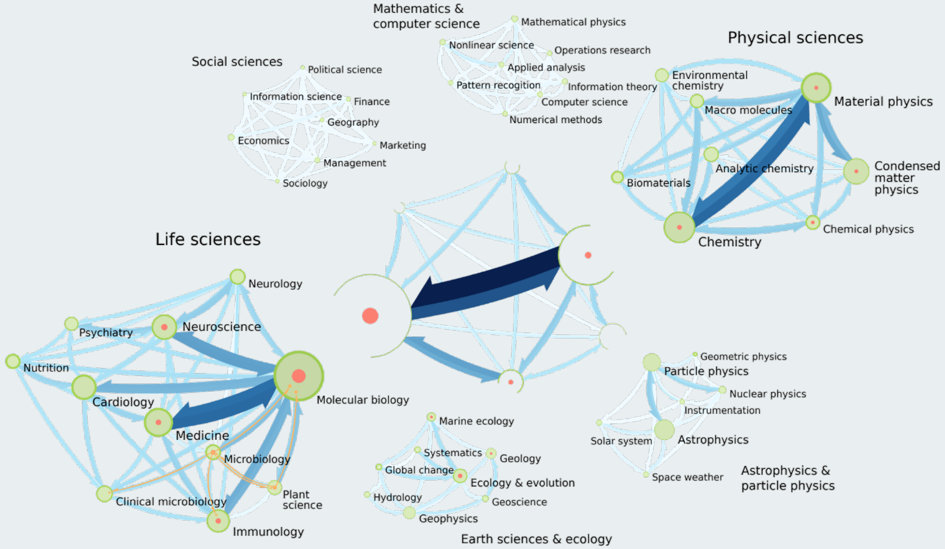
– High modularity ➔ more edges within the module that you expect by chance
– e_{ii} is the probability edge is in module i.
– a^2 is the probability a random edge would fall into module i.
– Tre indicates the trace of the matrix (the sum of the diagonal elements)
– $||e^2||$ refers sum of all the elements in the matrix
– A Q of 90% indicates almost all links are to members within the groups identified by Girvan-Newman, whereas one of 10% indicates that few are.

Infomap (mapequation)

– A multi-layer (temporal) community detection method that could process weighted/ unweighted directed/ undirected networks, bipartite networks, and multi-layer networks.

Aslak, U., Rosvall, M., & Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. *Physical Review E*, *97*(6), 062312.





Persson, C., Bohlin, L., Edler, D., & Rosvall, M. (2016). Maps of sparse Markov chains efficiently reveal community structure in network flows with memory. *arXiv preprint arXiv:1606.08328*.

Infomap (mapequation)

- For a module partition M of n nodes $\alpha = 1, 2, \ldots, n$ into m modules $i = 1, 2, \ldots, m$, we define this lower bound on code length to be $L(M)$.

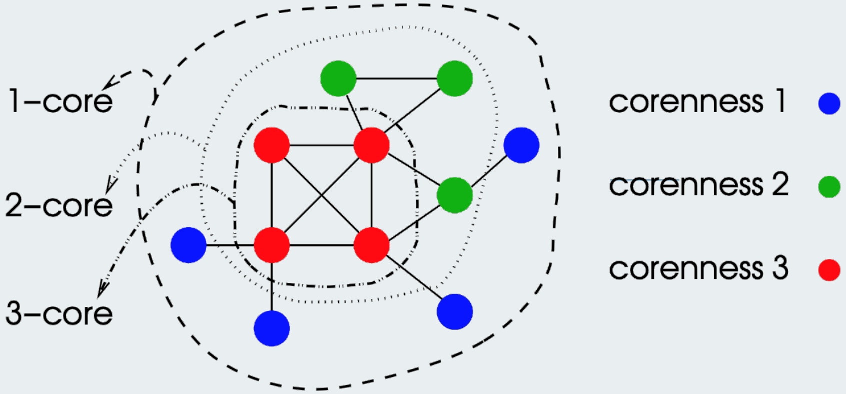
$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{\circlearrowright}^{i} H\left(\mathcal{P}^{i}\right)$$

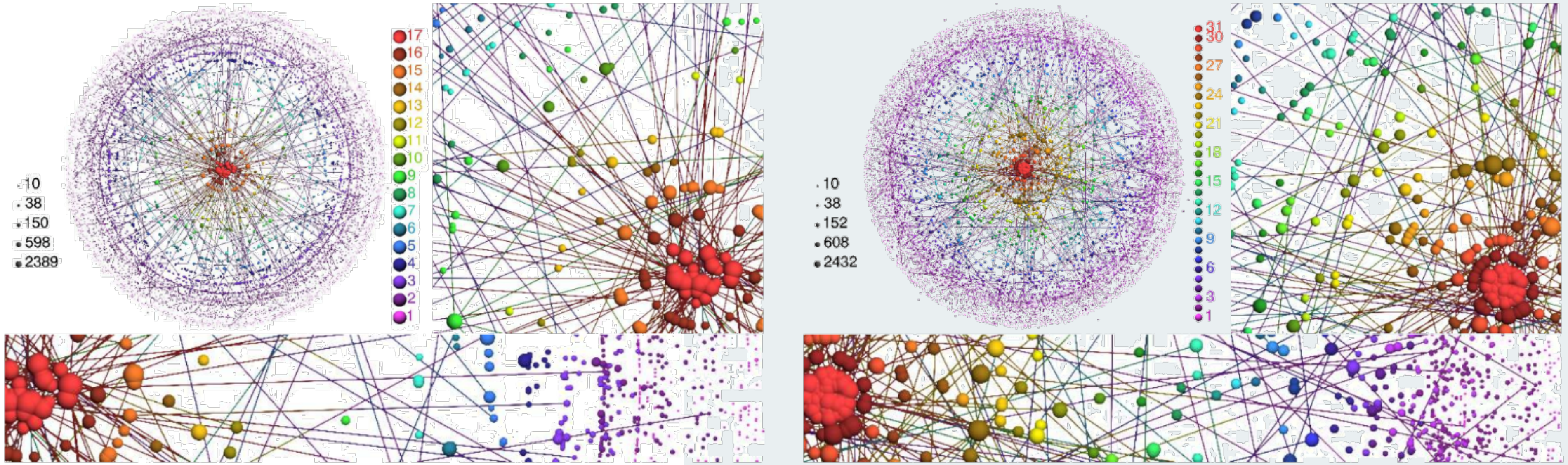
- $H(\mathcal{Q})$ is the frequency-weighted average length of codewords in the index codebook and $H\left(\mathcal{P}^{i}\right)$ is frequency-weighted average length of codewords in module codebook i. Further, the entropy terms are weighted by the rate at which the codebooks are used. With $q_{i\curvearrowright}$ for the probability to exit module i, the index codebook is used at a rate $q_{\curvearrowright} = \sum_{i=1}^{m} q_{i\curvearrowright}$, the probability that the random walker switches modules on any given step. With p_{α} for the probability to visit node α, module codebook i is used at a rate $p_{\circlearrowright}^{i} = \sum_{\alpha \in i} p_{\alpha} + q_{i\curvearrowright}$, the fraction of time the random walk spends in module i plus the probability that it exits the module and the exit message is used.

k-core Decomposition

K-core decomposition could differentiate the core–periphery structure of a given graph.

Think about the importance of core–periphery structure in your field!





Paper Reading

International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism Ankara, Turkey, 3-4 Dec, 2018

Application Areas of Community Detection: A Review

Arzum Karataş
Department of Computer Engineering
Izmir Institute of Technology
İzmir, Turkey
arzumkaratas@iyte.edu.tr

Serap Şahin
Department of Computer Engineering
Izmir Institute of Technology
İzmir, Turkey
serapsahin@iyte.edu.tr

***Abstract*— In the realm of today's real world, information systems are represented by complex networks. Complex networks contain a community structure inherently. Community is a set of members strongly connected within members and loosely connected with the rest of the network. Community detection is the task of revealing inherent community structure. Since the networks can be either static or dynamic, community detection can be done on both static and dynamic networks as well. In this study, we have talked about taxonomy of community detection methods with their shortages. Then we examine and categorize application areas of community detection in the realm of nature of complex networks (i.e., static or dynamic) by including sub areas of criminology such as fraud detection, criminal identification, criminal activity detection and bot detection. This paper provides a hot review and quick start for researchers and**

functional similarities among the members of the network [3]. Therefore, detecting community structure provides us meaningful insights about the network structure and its organization principle.

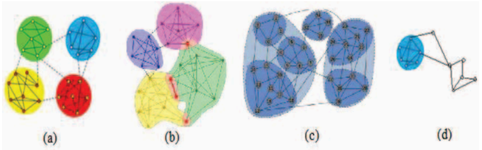


Figure 1. An example of illustrating different types of communities: (a) disjoint, (b) overlapping, (c) hierarchical and (d) local communities

Karataş, A., & Şahin, S. (2018, December). Application areas of community detection: A review. In 2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT) (pp. 65-70). IEEE.

Questions:

1. How do you think that community detection could apply into these areas?
2. If you want to leverage community detection into your area, and what are the importance and meanings of community in your graph?

References

- Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.
- Tsvetovat, M., & Kouznetsov, A. (2011). Social Network Analysis for Startups: Finding connections on the social web. "O'Reilly Media, Inc.".
- Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods.
- Aslak, U., Rosvall, M., & Lehmann, S. (2018). Constrained information flows in temporal networks reveal intermittent communities. Physical Review E, 97(6), 062312.
- Persson, C., Bohlin, L., Edler, D., & Rosvall, M. (2016). Maps of sparse Markov chains efficiently reveal community structure in network flows with memory. arXiv preprint arXiv:1606.08328.
- Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., & Vespignani, A. (2005). k-core decomposition: A tool for the visualization of large scale networks. arXiv preprint cs/0504107.

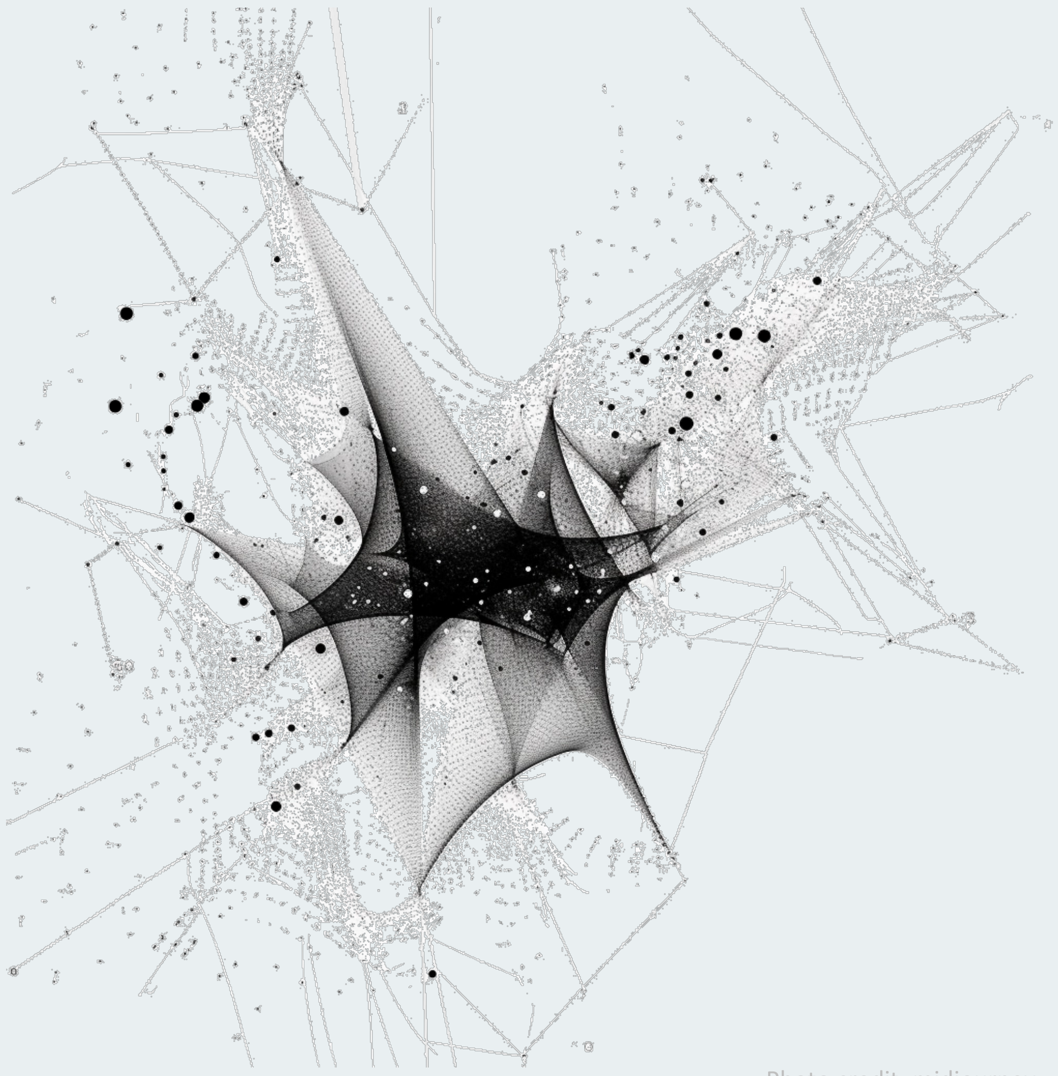


Photo credit: midjourney

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Website: toodou.github.io

Photo credit: midjourney