Plan Overview

A Data Management Plan created using DMPonline

Title: Validity check for C-P structure in usage on identifying influential nodes

Creator: Gyuho Bae

Data Manager: Gyuho Bae

Contributor: Gyuho Bae

Affiliation: University of Strathclyde

Template: University of Strathclyde

Project abstract:

This project scraps various graphs(networks) from KONECT and NETZSCHEDULER network repositories. All networks are treated as undirected, unweighted, and saved in .CSV or sparse network format using Python and Numpy tools.

This project also contains SIR simulated data for a part of scraped networks size under 10** (8). The details of producing SIR simulation data can be found in the authors' preprinted manuscript.

ID: 170420

Start date: 01-01-2023

End date: 01-03-2025

Last modified: 11-02-2025

Validity check for C-P structure in usage on identifying influential nodes

Administrative Data

Creator

Gyuho Bae

Creator Department

Mathematics and Statistics

ID

gbb19209(gyu.bae@strath.ac.uk)

Co-investigator(s)

Philip A. knight and Young-Ho Eom

Co-investigator(s) contact details

University of Strathclyde(G.B and P.A.K), University of Seoul(Y-H. E.)

Project title

Validity Check of Influential Node Groups by Core-periphery Identification Methods

Project Description

This project belongs to part of network science that investigates various diagrams consisting of nodes and links. Here, a network is an abstract representation of systems such as SNS friendship, air transportation, protein interaction, or the social interaction of animals, to name a few. The actors(people, base protein, airports, etc) are represented as nodes, and their interactions are links.

This project, particularly, focuses on networks' substructure, the so-called "core-periphery", which is comprised of the core, a densely connected group of central nodes, and the periphery, subordinately connected to the core while disconnected themselves. For instance, core members are active Twitter users consuming and reproducing information simultaneously, while periphery members usually consume (out-link) or provide (input-link) information. A more physically intuitive example is the physical structure of the Internet or World Airline Network. In the World Airline Network, the core airports reduce the number of connecting flights from one local airport to another local one, providing relatively short stops between them.

Because of these features, the core of the core-periphery structure can be regarded as an important and influential substructure of the network. By producing and investigating this data, we discover how the core's connectivity features, internal connectivity of the core and external connectivity between the core and periphery affect the core's relative influence on the periphery.

We provide the connectivity measures represented as the link density, which indicates the ratio of links inside the core, periphery or between them over the maximal number of links between components(nodes). For instance, if the core has an 'N' number of nodes and an 'L' number of links inside, the link density is calculated as 2L/(N^2-N), where N(N-1)/2 is the number of links every node are connected to each other. One can consider the link density as the occupation of current links over the maximum capacity of the link. In terms of link density, the core-periphery structure, especially, a core-dominant structure, should satisfy p11 >p12 > p22, which indicates the core has the largest link density, intermediate value of it for between the core and periphery, and the smallest link density for the periphery.

There are many methods to Identify the core-periphery structure, which have different basic mechanisms but commonly find the densely connected central node groups from the network. For this reason, identified cores from different algorithms can be similar if the network does have a substructure that is able to be discriminated as the core. Note that the detailed member of the core-periphery partition can be different. However, the correlations of the connectivity measures from different identifications show noticeable positive correlations(Pearson's R over 0.8 with a p-value under 0.0001).

We use SIR epidemic spreading model to estimate the node's influence in the network. The SIR model uses three states of individuals: "susceptible", "infected", and "recovered(removed)". In our setting, we put every node as susceptible except for a single infected node. The number of infected nodes increases as the dynamics proceed step by step(time), while infected nodes can recovered with a certain ratio. In our simulation, recovered nodes remain and do not get infected or change to a susceptible state. The simulation ends where there is no infected node. We estimate the node influence as the number of recovered nodes where the

infection started from that single node. To calculate, we run 30 simulations per node in the network, so the calculation costs 30 x V where V is the number of nodes in the simulated network.
The core's relative influence, then, is the ratio of the average influence of the core and periphery. Note that one should choose the parameter carefully to see differences in the relative influence since wide and fast spreading diminishes difference of influence between the core and periphery. We also run the simulation with different settings to show this feature.

Funder

Question not answered.

Grant reference number

Question not answered.

Project start date

01032023

Project end date

081202024

Date of first version

11022025

Date of last revision

Question not answered.

Related policies

Question not answered.

Existing data

Data Collection

What data will be collected or created?

Network repositories KONECT and NETZSCHEDUELER provide network nodes and links in.CSV or another format. We take those files and proceed it as undirected and unweighted network.

How will the data be collected or created?

Network repositories KONECT and NETZSCHEDUELER provide network nodes and links in .CSV or another format. We take those files and proceed with an undirected and unweighted network.
We use Python Networkx to import & export the data. We identify the core-periphery structure using Python modules provided by previous researchers(link: Gallagher's c-p, & Kojaku's c-p module).

Documentation and Metadata

What documentation or metadata will accompany the data?

All metadata can be accessed at network repositories, however, we put metadata separately in our DMP

That metadata contains what data used for constructing the network, data provider, and related research.

Ethics and Legal Compliance

How will ethical issues relating to data be managed?

Question not answered.

How will copyright and Intellectual property (IPR) issues be managed?

We use the data for purely academic purpose, which means it will be creative common data.

Storage and Backup

How will data be stored, backed up and shared during the research project?

G.B. handles the data personally with an extraneous hard drive. G.B. will compress and upload the data to i drive that is connected to University of Strathclyde for granting public access.

How will access and security to data be managed during the research project?

Question not answered.

Selection and Preservation

Which data should be retained, shared, preserved and destroyed

Most of the simulated results are not necessary for current research, so we keep the collection of the average node influence of the network.

Saved networks as .CSV will remain.

Core-periphery partitions are destroyed after calculating the link density. Only link density and size of the core and periphery

remains.

What is the long-term preservation plan for data?

Question not answered.

Data Sharing

How will the data be shared?

Question not answered.

Are any restrictions on data sharing required?

Question not answered.

Responsibilities and Resources

Who will be responsible for data management?

Question not answered.

What resources will you require to deliver your plan?

Question not answered.

Planned Research Outputs

Journal article - "Validity Check of Influential Node Groups by Core-periphery Identification Methods"

Planned research output details

Title	DOI	Type	Release date	Access level	Repository(ies)	File size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Validity Check of Influential Node Groups by Core- ...		Journal article	2025-02-11	Open	None specified		Creative Commons Attribution Non Commercial No Derivatives 4.0 International	None specified	No	No