BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN XUÂN DŨNG

NGHIÊN CỨU CÁC THUẬT TOÁN RÚT GỌN ĐỒ THỊ VÀ ỨNG DỤNG ĐỂ PHÁT HIỆN CỘNG ĐỒNG TRÊN MẠNG XÃ HỘI

LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

HÀ NỘI - 2021

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN XUÂN DŨNG

NGHIÊN CỨU CÁC THUẬT TOÁN RÚT GỌN ĐỒ THỊ VÀ ỨNG DỤNG ĐỂ PHÁT HIỆN CỘNG ĐỒNG TRÊN MẠNG XÃ HỘI

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN
MÃ SỐ: 9.48.01.04

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS.TS Đoàn Văn Ban
2. TS. Đỗ Thị Bích Ngọc

HÀ NỘI - 2021

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong luận án là trung thực và chưa từng được công bố trong bất cứ công trình nào.

TÁC GIẢ

Nguyễn Xuân Dũng

LỜI CẢM ƠN

Qua luận án này tôi xin chân thành cảm ơn PGS.TS Đoàn Văn Ban và TS. Đỗ Thị Bích Ngọc đã tận tình giúp đỡ, động viên, định hướng, hướng dẫn tôi nghiên cứu và hoàn thành luận án này.

Tôi xin chân thành cảm ơn các Thầy, Cô giáo trong Học viện Công nghệ Bưu chính Viễn thông đã tận tình giảng dạy và giúp đỡ tôi trong suốt khóa học. Tôi cũng xin cảm ơn PGS.TS Lê Nhật Thăng - Trưởng Khoa Đào tạo Sau Đại học của Học viện công nghệ bưu chính viễn thông, TS. Nguyễn Duy Phương - Trưởng Khoa Công nghệ thông tin của Học viện công nghệ bưu chính viễn thông và PGS.TS Phạm Thọ Hoàn - Giám đốc Trung tâm Khoa học Tính toán của Trường Đại học Sư phạm Hà Nội đã giúp đỡ tôi trong quá trình thực hiện luận án.

Tác giả chân thành mong nhận được những ý kiến đóng góp từ các Thầy, Cô giáo, các nhà khoa học và bạn bè đồng nghiệp.

Trân trọng cám ơn.

MỤC LỤC

DANH MỤC CÁC CHỮ VIẾT TẮT

TỪ VIẾT TẮT	DẠNG ĐẦY ĐỦ
BIRCH	Balanced iterative regucing and clustering using hierarchies
BFS	Breadth first search
CDAB	Community detection algorithm based on betweenness
DAG	Directed acyclic graph
EBC	Edge betweenness centrality
EAGLE	Agglomerative hierarchical clustering based on maximal clique
ELPA	Edge label propagation algorithm
EMLPA	Balanced multi labed propagation
FBC	Fast algorithm for betweenness centrality
FFS	Forest Fire Sampling
GN	Girvan-Newman
HLPA	Hybrid label propagation algorithm
LREN	Label based reduce equivalence nodes
LPA	Label propagation algorithm
LPAA	Label propagation algorithm on abridged graph
MAA	Majid Arasteh and Alizadeh
NMI	Normal mutual information
OLP	Optimized label propagation
RE	Random Edge Sampling
RNE	Random Node - Edge Sampling
REG	Reduce equivalence graph
SES	Snowball Expansion Sampling
SN	Social network
SNA	Social network analysis
SNAP	Stanford large network dataset collection

DANH MỤC CÁC KÝ HIỆU TOÁN HỌC

KÝ HIỆU	Ý NGHĨA
A_{ij}	Ma trận liền kề
d(x, y)	Khoảng cách giữa đỉnh x và y
G	Đồ thị
V	Tập đỉnh
E	Tập cạnh
D_G	Hệ số cố kết của đồ thị G
$C_D(v)$	Hệ số trung tâm trực tiếp của đỉnh v
deg(v)	Số bậc của đỉnh v
R	Tập số nguyên
$C_{Cl}(v)$	Hệ số trung tâm lân cận của đỉnh v
σ_{vt}	Số đường đi ngắn nhất đi v đến t
$C_B(v)$	Độ đo trung tâm trung gian của đỉnh v
d_i	Bậc của đỉnh i
d_j	Bậc của đỉnh j
$\Gamma(u)$	Tập các đỉnh liền kề với u và kể cả u
DAG_X	Đồ thị định hướng, phi chu trình gốc X
n	Số đỉnh của đồ thị
k	Bậc của đỉnh
L(u)	Nhãn của đỉnh u
L(v)	Nhãn của đỉnh v

DANH MỤC CÁC THUẬT NGỮ

THUẬT NGỮ TIẾNG ANH	THUẬT NGỮ TIẾNG VIỆT
Betweenness centrality	Độ đo trung tâm trung gian
Breadth first search	Duyệt theo chiều rộng
Closeness centrality	Hệ số trung tâm lân cận
Computer vision	Thị giác máy tính
Communication network	Mạng truyền thông
Communities detection	Phát hiện cộng đồng
Community social	Cộng đồng mạng xã hội
Cyclic workflow graph	Quy trình nghiệp vụ theo chu kỳ
Degree centrality	Hệ số trung tâm trực tiếp
Density Cohesion	Hệ số cố kết
Edge sampling	Phát hiện mẫu cạnh
Evolutionary algorithms	Thuật toán tiến hóa
Extremal Optimisation	Tối ưu hóa mở rộng
Graph clustering	Phân cụm theo đồ thị
Graph partitioning	Phân cụm theo đồ thị
Greedy techniques	Tìm kiếm tham lam
Hierarchical Agglomerative Clustering	Phân cụm phân cấp
Identical vertex	Đỉnh đồng nhất
Indexing and retrieval	Lập chỉ mục và hệ thống tìm kiếm
Image restoration	Phục hồi hình ảnh
Information theoretic	Lý thuyết thông tin
Label Propagation Algorithm	Thuật toán lan truyền nhãn
Leaf vertex	Đỉnh treo
Markov chain model-reduction problem	Rút gọn mô hình chuỗi Markov
Modularity Optimisation Based Community Detection Techniques	Thuật toán phát hiện cấu trúc cộng đồng dựa trên tối ưu hóa mô đun
Pair-counting	Tính toán cặp

Partitional clustering	Phân cụm phân hoạch
Sampling from large graphs	Phát hiện mẫu trong các đồ thị lớn
Semantic graph	Đồ thị ngữ nghĩa
Set-matching based	Độ trùng cặp
Side vertex	Đỉnh sườn
Simulated annealing	Mô phỏng luyện kim
Social Networks	Mạng xã hội
Social Network Analysis	Phân tích mạng xã hội
Social Network community	Cộng đồng mạng xã hội
Spectral clustering	Phân cụm theo phổ
Structural conflicts	Xung đột cấu trúc
Structural features	Đặc trưng cấu trúc mạng
Text summarization	Tóm tắt văn bản
Traditional Community Detection Techniques	Thuật toán phát hiện cấu trúc cộng đồng truyền thống
Traversal - based sampling	Phát hiện mẫu dựa trên truyền tải
Vertex sampling	Phát hiện mẫu đỉnh
Workflow management system	Hệ thống quản lý luồng công việc

DANH MỤC HÌNH VẼ

DANH MỤC CÁC BẢNG

MỞ ĐẦU

1. Tính cấp thiết của luận án

Trong vài thập kỷ gần đây, các mạng xã hội (SN - Social Networks) đã trở nên phổ biến và thu hút được sự chú ý của các nhà khoa học thuộc các ngành khác nhau, như xã hội học, dịch tễ học, kinh tế, khoa học máy tính, viễn thông và nhiều ngành khác. Mạng xã hội đang phát triển mạnh mẽ tại khắp mọi nơi, trên mọi quốc gia và trở thành phương tiện quan trọng, không thể thiếu trong cuộc sống để kết nối quan hệ của mọi người trong xã hội. Hiện nay Facebook, Twitter, Youtube, WhatsApp, Instagram, Google+, Linkedin, … là những mạng xã hội phổ biến được nhiều người sử dụng nhất.

Phân tích mạng xã hội (SNA - Social Network Analysis) là một tập hợp các phương pháp thu thập và xử lý dữ liệu, các khái niệm, các lý thuyết nhằm mô tả và phân tích các mối quan hệ giữa các thực thể trong mạng, các quy luật hình thành và biến đổi của những mối quan hệ đó, và nhất là làm sáng tỏ những ảnh hưởng tương quan của các mối quan hệ trong xã hội (hay cấu trúc của mạng) đối với hành vi của các thực thể tham gia. Ví dụ: Phân tích thống kê mạng xã hội, phát hiện cộng đồng trên mạng xã hội, dự đoán liên kết, phân tích vai trò và phân loại các tác nhân trên mạng xã hội, … Trong lĩnh vực phân tích mạng xã hội, việc phân tích và phát hiện các cộng đồng (communities detection) trên mạng xã hội mang nhiều ý nghĩa quan trọng và có nhiều ứng dụng trong các lĩnh vực khác nhau như xã hội học, sinh học, khoa học máy tính, kinh tế, chính trị, …. Cộng đồng mạng xã hội là một nhóm các thực thể trong mạng xã hội có những tính chất tương tự nhau, liên kết chặt chẽ với nhau và cùng đóng một vai trò nhất định. Cộng đồng mạng xã hội là những cấu trúc xã hội được xác định dựa trên những mối quan hệ, có mối quan tâm chung như sở thích, lĩnh vực mà các thành viên của cộng đồng cùng quan tâm, tham gia hay một mục tiêu, dự án chung, vị trí địa lý, hoặc nghề nghiệp. Việc phát hiện và phân tích các cộng đồng mạng xã hội sẽ cung cấp cho chúng ta những thông tin quý giá để hiểu biết và hình dung được những cấu trúc của mạng.

Phát hiện cộng đồng trên mạng xã hội cũng là một nhiệm vụ quan trọng hàng đầu trong phân tích mạng xã hội. Do tầm quan trọng của các cộng đồng mạng xã hội và khả năng ứng dụng to lớn của chúng trong các lĩnh vực khác nhau đã có nhiều các thuật toán phát hiện cộng đồng trên mạng xã hội đã được đề xuất. Tuy nhiên, hầu hết các thuật toán chưa đạt được hiệu quả trong việc phát hiện cộng đồng trên các mạng xã hội quy mô rất lớn hiện nay. Đồng thời, cùng với sự phát triển mạnh mẽ của công nghệ thông tin thì việc sử dụng các mạng xã hội của chúng ta đang phát triển theo cấp số nhân và hệ quả là quy mô của mạng xã hội phát triển nhanh chóng và trở nên khổng lồ. Điều này dẫn đến việc phát hiện cộng đồng trên các mạng xã hội quy mô rất lớn không thể giải quyết bằng các thuật toán truyền thống do độ phức tạp về thời gian và không gian tính toán. Có nghĩa là, hầu hết các thuật toán hiện có không thể được mở rộng đến kích thước khổng lồ của các mạng xã hội. Để giải quyết được thách thức đặt ra, cần đề xuất các phương pháp giảm kích thước của mạng xã hội để thực hiện phát hiện cộng đồng mạng xã hội hiệu quả đồng thời vẫn phải đảm bảo được các tính chất của cộng đồng mạng xã hội ban đầu là rất ý nghĩa, cần thiết và quan trọng.

Trong những năm gần đây, việc phân tích và phát hiện cộng đồng mạng xã hội là một trong những lĩnh vực nghiên cứu chính trong khai thác, phân tích mạng xã hội. Các thuật toán phát hiện cộng đồng trên mạng xã hội được nhiều người tập trung quan tâm nghiên cứu và phát triển ứng dụng [8], [9], [28], [42], [102], [118], [119], [120], ... Về cơ bản, các thuật toán phát hiện cộng đồng mạng xã hội được chia thành 4 nhóm. Nhóm thuật toán phát hiện cộng đồng truyền thống, nhóm thuật toán phát hiện cộng đồng dựa trên tối ưu hóa độ đo đơn thể, nhóm thuật toán phát hiện cộng đồng dựa vào độ đo trung tâm trung gian, và nhóm thuật toán phát hiện cộng đồng dựa trên nguyên lý lan truyền nhãn. Trong đó, nhóm thuật toán phát hiện cộng đồng truyền thống bao gồm các thuật toán phân cụm đồ thị, phân cụm phân cấp, phân cụm phân hoạch, phân cụm theo phổ [31], [76], [115]. Nhóm thuật toán phát hiện cộng đồng dựa trên tối ưu hóa độ đo đơn thể bao gồm thuật toán tìm kiếm tham lam, mô phỏng luyện kim, tối ưu hoá mở rộng và các thuật toán tiến hoá [15], [78], [91]. Nhóm thuật toán phát hiện cộng đồng dựa vào độ đo trung tâm trung gian bao gồm họ thuật toán

Girvan-Newman theo độ đo trung tâm trung gian của cạnh, phân chia đỉnh [33], [34], [38], [75]. Và cuối cùng là nhóm thuật toán dựa trên nguyên lý lan truyền nhãn bao gồm họ các thuật toán dựa vào nguyên lý lan truyền nhãn [13], [59], [81], [109], [110].

Đồ thị mạng xã hội thường rất phức tạp, có số đỉnh và số cạnh rất lớn, nên công việc phát hiện các cộng đồng đòi hỏi rất nhiều thời gian và cũng là một thách thức rất lớn. Tuy nhiên, các nghiên cứu nêu trên hầu hết tập trung giải quyết bài toán phát hiện cộng đồng trực tiếp trên đồ thị mà rất ít công trình nghiên cứu tính đến việc giảm thiểu không gian đỉnh và cạnh của đồ thị nhưng bảo toàn được các tính chất của đồ thị mạng xã hội ban đầu nhằm mục đích giảm thiểu thời gian phân tích, phát hiện các cộng đồng trên mạng xã hội. Mặt khác, đồ thị mạng xã hội thường có nhiều đỉnh tương đương với nhau theo một số độ đo đã được xác định đặc trưng cho mạng xã hội như: độ đo trung tâm trung gian, hoặc theo nguyên lý lan truyền nhãn, ... Những đỉnh tương đương có cùng độ đo trung tâm trung gian, hay có chung nhãn theo nguyên lý lan truyền nhãn tạo thành các lớp đỉnh tương đương và có thể kết hợp chúng với nhau thành một đỉnh đại diện giúp cho giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội.

Qua phân tích và đánh giá các thuật toán phát hiện các cộng đồng trên mạng xã hội, nghiên cứu sinh đã lựa chọn nghiên cứu các lớp đỉnh tương đương theo độ đo trung tâm trung gian và nguyên lý lan truyền nhãn để rút gọn đồ thị mạng xã hội và từ đó cải tiến các thuật toán phát hiện cộng đồng mạng xã hội hiệu quả trên đồ thị rút gọn nhằm giải quyết hiệu quả bài toán phát hiện cộng đồng trên mạng xã hội có cấu trúc tự do và kích thước rất lớn.

2. Mục tiêu của luận án

Mục tiêu của luận án là nghiên cứu phát triển một số phương pháp phát hiện cộng đồng trên mạng xã hội. Cụ thể:

- Nghiên cứu phát triển và thực nghiệm thuật toán rút gọn đồ thị dựa vào lớp tương đương của các đỉnh trên đồ thị theo độ đo trung tâm trung gian và thuật toán rút gọn đồ thị theo nguyên lý lan truyền nhãn.
- Phát triển thuật toán phát hiện nhanh các cộng đồng trên mạng xã hội sử dụng độ đo trung tâm trung gian và thuật toán phát hiện nhanh các cộng đồng trên mạng xã hội dựa trên tính chất của các lớp đỉnh tương đương theo nguyên lý lan truyền nhãn.

3. Đối tượng nghiên cứu của luận án

- Mạng xã hội, cộng đồng mạng xã hội.
- Các thuật toán rút gọn đồ thị.
- Các lớp đỉnh tương đương theo độ đo trung tâm trung gian và nguyên lý lan truyền nhãn trên đồ thị mạng xã hội.
- Các thuật toán phát hiện cộng đồng mạng xã hội.

4. Phạm vi nghiên cứu của luận án

- Các thuật toán phát hiện cộng đồng mạng xã hội.
- Các lớp đỉnh tương đương theo độ đo trung tâm trung gian trên đồ thị mạng xã hội.
- Các lớp đỉnh tương đương theo nguyên lý lan truyền nhãn trên đồ thị mạng xã hội.
- Các thuật toán rút gọn đồ thị dựa vào các lớp đỉnh tương đương theo độ đo trung tâm trung gian và theo nguyên lý lan truyền nhãn.

5. Phương pháp nghiên cứu của luận án

Phương pháp nghiên cứu của luận án là nghiên cứu lý thuyết và nghiên cứu thực nghiệm.

- *Nghiên cứu lý thuyết:* Nghiên cứu và đánh giá các nguồn tài liệu, công trình liên quan một cách hệ thống, toàn diện bài toán rút gọn đồ thị mạng xã hội và ứng dụng phát hiện cộng đồng trên đồ thị mạng xã hội và các vấn đề còn tồn tại của các nghiên cứu liên quan. Trên cơ sở đó, đề xuất thuật toán rút gọn đồ thị dựa trên các lớp đỉnh tương đương theo một số độ đo trên đồ thị mạng xã

hội và phát triển các thuật toán phát hiện cộng đồng trên đồ thị mạng xã hội rút gọn. Các thuật toán đề xuất, cải tiến được chứng minh chặt chẽ về lý thuyết thông qua các tính chất, hệ quả về sự tương đương của các lớp đỉnh rút gọn.

- *Nghiên cứu thực nghiệm*: Các thuật toán đề xuất được cài đặt, chạy thực nghiệm, so sánh, đánh giá với thuật toán khác trên các bộ dữ liệu mẫu từ kho dữ liệu về mạng xã hội [47], [60] nhằm minh chứng tính hiệu quả của các nghiên cứu về lý thuyết.

6. Các đóng góp chính của luận án

- Đề xuất thuật toán **REG** (**R**educe **E**quivalence **G**raph) rút gọn đồ thị dựa vào lớp tương đương của các đỉnh theo độ đo trung tâm trung gian. Thực hiện các thực nghiệm đánh giá tính hiệu quả và thời gian thực hiện của thuật toán đề xuất so với thuật toán điển hình sử dụng độ đo trung tâm trung gian.
- Đề xuất thuật toán **FBC** (**F**ast algorithm for **B**etweenness **C**entrality) cải tiến thời gian tính độ đo trung tâm trung gian và đề xuất thuật toán **CDAB** (**C**ommunity **D**etection **A**lgorithm based on **B**etweenness centrality) cải tiến thời gian phát hiện các cộng đồng trên đồ thị mạng xã hội rút gọn dựa vào độ đo trung tâm trung gian. Thực hiện các thực nghiệm đánh giá tính hiệu quả và thời gian thực hiện của thuật toán đề xuất **CDAB** so với thuật toán gốc Girvan-Newman (GN) và thuật toán điển hình gần đây.
- Đề xuất thuật toán **LREN** (**L**abel based **R**educe **E**quivalence **N**odes) rút gọn đồ thị dựa vào lớp đỉnh tương đương theo nguyên lý lan truyền nhãn và phát triển thuật toán **LPAA** (**L**abel **P**ropagation **A**lgorithm on **A**bridged graph) cải tiến thời gian phát hiện các cộng đồng dựa vào nguyên lý lan truyền nhãn. Thực hiện các thực nghiệm đánh giá tính hiệu quả và thời gian thực hiện của thuật toán **LPAA** so với thuật toán gốc Label Propagation Algorithm (LPA) và thuật toán điển hình gần đây.

7. Bố cục của luận án

Luận án được tổ chức thành 3 chương, trong đó:

Chương 1. Tổng quan rút gọn đồ thị và phát hiện cộng đồng trên mạng xã hội

Nội dung chính của chương 1 là trình bày tổng quan về mạng xã hội, cộng đồng mạng xã hội và các phân tích, đánh giá về các thuật toán rút gọn đồ thị, thuật toán phát hiện cộng đồng trên mạng xã hội và các ứng dụng trong các lĩnh vực khác nhau. Một số các độ đo được giới thiệu để sử dụng đánh giá tính hiệu quả của thuật toán rút gọn đồ thị và thuật toán phát hiện cộng đồng trên mạng xã hội.

Chương 2. Thuật toán rút gọn đồ thị mạng xã hội dựa vào độ đo trung tâm trung gian và nguyên lý lan truyền nhãn.

Chương 2 nghiên cứu các tính chất của lớp đỉnh tương đương dựa vào độ đo trung tâm trung gian, đề xuất thuật toán **REG** rút gọn đồ thị dựa trên thay thế các lớp đỉnh tương đương theo độ đo trung tâm trung gian, đề xuất này nhằm mục tiêu giảm thiểu không gian tính toán của đồ thị, từ đó giảm thiểu độ phức tạp tính toán của bài toán so với các phương pháp trước đây. Đồng thời trong chương này cũng nghiên cứu các tính chất của lớp đỉnh tương đương dựa vào nguyên lý lan truyền nhãn, từ đó đề xuất thuật toán **LREN** rút gọn đồ thị dựa trên thay thế các lớp đỉnh tương đương.

Các thực nghiệm khẳng định hiệu quả của thuật toán đề xuất trong bài toán rút gọn đồ thị mạng xã hội. Nội dung trình bày trong chương được công bố trong [CT1], [CT3], [CT4].

Chương 3. Áp dụng thuật toán rút gọn đồ thị để phát hiện cộng đồng trên mạng xã hội.

Chương 3 đề xuất thuật toán **FBC** cải tiến thời gian tính độ đo trung tâm trung gian trên đồ thị mạng xã hội. Đề xuất này nhằm mục tiêu giảm thiểu thời gian tính toán độ đo khoảng cách trên đồ thị mạng xã hội phục vụ cho thuật toán đề xuất phát hiện cấu trúc cộng đồng **CDAB** trên đồ thị mạng xã hội rút gọn. Đồng thời trong chương này cũng đề xuất thuật toán **LPAA** phát hiện các cộng đồng trên đồ thị mạng xã hội rút gọn. Đề xuất này nhằm mục tiêu giảm thiểu thời gian tính toán cho thuật toán phát hiện các cộng đồng trên đồ thị mạng xã hội rút gọn.

Các thực nghiệm khẳng định hiệu quả của thuật toán đề xuất trong bài toán phát hiện cộng đồng mạng xã hội. Nội dung trình bày trong chương được công bố trong [CT2], [CT3].

Cuối cùng là kết luận và các hướng phát triển tiếp theo.

CHƯƠNG 1. TỔNG QUAN RÚT GỌN ĐỒ THỊ VÀ PHÁT HIỆN CỘNG ĐỒNG TRÊN MẠNG XÃ HỘI

Chương này giới thiệu tổng quan về mạng xã hội, cộng đồng trên mạng xã hội, các thuật toán phát hiện cộng đồng mạng xã hội và các thuật toán rút gọn đồ thị cho nhiều ứng dụng khác nhau. Trong nội dung chương cũng thực hiện phân tích, đánh giá rõ những mặt hạn chế, tồn tại của mỗi phương pháp từ đó xác định hướng phát triển thuật toán rút gọn đồ thị và ứng dụng để cải tiến thuật toán phát hiện cộng đồng trên mạng xã hội. Cuối chương trình bày một số độ đo phổ biến được sử dụng để đánh giá hiệu quả của các thuật toán rút gọn đồ thị và thuật toán phát hiện cộng đồng trên mạng xã hội.

1.1. Mạng xã hội

Mạng xã hội là một cấu trúc xã hội được tạo ra từ các thực thể, các tác nhân hoặc các tổ chức được liên kết, kết nối bởi một hoặc nhiều quan hệ với nhau [8], [42], [102]. Theo Fortunato và các cộng sự [31] mạng xã hội là một tập hợp các thực thể được kết nối với nhau bằng một tập hợp các mối quan hệ, liên kết, như quan hệ bạn bè, gia đình, cộng sự hay trao đổi thông tin, … Các mối quan hệ giữa các thực thể có thể mang nhiều nội dung khác nhau từ sự tương trợ, trao đổi thông tin cho đến việc trao đổi hàng hóa, dịch vụ, … Mạng xã hội cung cấp nhiều cách khác nhau để các tổ chức thu thập thông tin, cạnh tranh với nhau trong việc thiết lập giá kinh doanh hoặc chính sách, … Mạng xã hội thường có những đặc tính như sau [9], [34], [68], [102]:

- *Dựa vào người dùng* (User-based): Trước khi các mạng xã hội như Facebook, Twitter, MySpace, … phổ biến trở thành chuẩn mực, các trang web dựa trên nội dung được cập nhật bởi người dùng và được người sử dụng truy cập trên mạng Internet để đọc, tham khảo thông tin. Các mạng xã hội trực tuyến được xây dựng và định hướng bởi chính người dùng. Người dùng thực hiện các cuộc hội thoại và các nội dung trao đổi với nhau trên mạng. Hướng của nội dung đó được xác định bởi bất kỳ ai tham gia vào cuộc thảo luận. Vì vậy, mạng xã hội trở nên rất

hấp dẫn, thu hút bởi tính năng tương tác nhiều hơn đối với người dùng Internet thông thường.

- *Tương tác* (Interactive): Một đặc điểm khác của các mạng xã hội hiện đại là các thực thể thường xuyên tương tác thông qua các mối liên kết. Điều này có nghĩa là một mạng xã hội không chỉ là một bộ sưu tập các phòng chat, diễn đàn, …, trang web như Facebook mà còn chứa các ứng dụng chơi trò chơi, quảng cáo, bán hàng online, tin tức, … Các mạng xã hội ngày nay đang phát triển nhanh chóng và được người dùng lựa chọn nhiều hơn so với truyền hình bởi vì nó không chỉ là giải trí, học tập, trao đổi công việc mà đó còn là cách thức để mọi người kết nối, tương tác với nhau.
- *Hướng đến cộng đồng* (Community-driven): Mạng xã hội được xây dựng và phát triển từ các khái niệm về cộng đồng. Điều này có nghĩa là các cộng đồng hoặc các nhóm xã hội trên toàn thế giới được thành lập dựa trên thực tế là các thành viên có những sở thích, những quan điểm chung, ...
- *Các mối quan hệ* (Relationships): Không giống như các trang web trong quá khứ, các mạng xã hội phát triển mạnh về các mối quan hệ. Càng có nhiều mối quan hệ trong mạng, các thực thể càng thiết lập được vai trò trung tâm của mạng đó. Mối quan hệ giữa các thực thể như mối quan hệ hai người có thể là bạn bè hoặc không quen biết nhau. Tồn tại tính địa phương, mối quan hệ giữa các thực thể có xu hướng tạo thành các cụm (cộng đồng). Mạng xã hội cung cấp tiềm năng rất lớn về tương tác và giao tiếp giữa rất nhiều các thành viên trong mạng ở khắp mọi nơi, không phụ thuộc vào không gian địa lý. Đồng thời tạo môi trường cho việc tương tác và chia sẻ thông tin giữa các thành viên trong mạng như người thân, đồng nghiệp, gia đình, bạn bè, người hâm mộ, … [68].
- *Cảm xúc về nội dung* (Emotion over content): Một đặc điểm độc đáo khác của mạng xã hội là yếu tố cảm xúc. Mặc dù các trang web trong quá khứ tập trung chủ yếu vào việc cung cấp thông tin cho người truy cập, nhưng mạng xã hội ngày nay thực sự mang đến cho người dùng sự an toàn về mặt cảm xúc và cảm giác rằng dù có chuyện gì xảy ra, bạn bè của họ vẫn ở trong tầm kiểm soát.

Hiện nay, mạng xã hội đang phát triển nhanh chóng, với số lượng người dùng và số lượng các mối quan hệ giữa các thành viên trong mạng rất lớn. Từ đó, yêu cầu khách quan đặt ra đòi hỏi phải có những phương pháp nghiên cứu và kỹ thuật phân tích mạng xã hội phù hợp.

1.2. Một số hệ đo quan trọng trên đồ thị mạng xã hội

Phân tích mạng xã hội (Social Network Analysis) [8], [9], [28], [42], [102], [105] dựa vào lý thuyết đồ thị là một tập hợp các phương pháp lựa chọn mẫu, thu thập và xử lý dữ liệu, phân tích các khái niệm, sử dụng lý thuyết đồ thị để mô tả và phân tích các mối quan hệ giữa các thực thể, các tác nhân trong mạng, xác nhận các quy luật hình thành và biến đổi của những mối quan hệ đó, và nhất là làm sáng tỏ những ảnh hưởng của các mối quan hệ xã hội (hay cấu trúc của mạng) đối với hành vi của các tác nhân. Mục tiêu chính của phân tích mạng xã hội là:

- *Xác định những thực thể, tác nhân quan trọng nhất trong mạng xã hội*: Độ đo trung tâm (centrality) là một độ đo điển hình để xác định tầm quan trọng của một tác nhân trong mạng, đồng thời giúp chúng ta hiểu được tầm ảnh hưởng và quyền lực của một cá nhân trong xã hội.
- *Phát hiện các cộng đồng trên mạng xã hội*: Một số thực thể trong mạng xã hội có liên kết chặt chẽ với nhau tạo thành từng cụm, và giữa các cụm đó được nối với nhau chỉ bằng một số ít cạnh khác. Nhiệm vụ xác định các cộng đồng mạng xã hội được thực hiện thông qua nghiên cứu cấu trúc mạng xã hội và cấu trúc liên kết giữa các thực thể trên mạng xã hội.

Mục này trình bày khái niệm đồ thị mạng xã hội và một số hệ đo quan trọng được sử dụng phổ biến trên đồ thị mạng xã hội. Mạng xã hội thường được mô hình hóa, trực quan hóa và biểu diễn dưới dạng một đồ thị, chỉ giữ lại các thành viên và mối quan hệ giữa các thành viên trên mạng có tồn tại hay không. Thông thường đồ thị mạng xã hội là đồ thị vô hướng, ví dụ như đồ thị mạng bạn bè trên mạng xã hội Facebook, … Nhưng chúng cũng có thể là đồ thị có hướng như đồ thị mạng xã hội những người theo dõi nhau (followers) trên mạng xã hội Twitter hoặc Google +.

Định nghĩa 1.1. Đồ thị mạng xã hội là đồ thị G = (V, E), trong đó V là tập các đỉnh (nút) và E là tập các cạnh (cung). Tập V biểu diễn cho các thành viên (tác nhân) của mạng xã hội, còn tập E thể hiện mối quan hệ xã hội giữa các thành viên với nhau.

Dựa vào lý thuyết đồ thị, cấu trúc mạng xã hội cũng có thể được biểu diễn thông qua ma trận liền kề $A = (A_{ij}) \in R^{n \times n}$, với n = |V|, R = {0, 1} và $A_{ij} = 1$ nếu hai đỉnh i và j có cạnh nối giữa chúng (có liên kết - quan hệ trực tiếp với nhau), ngược lại thì $A_{ij} = 0$.

Để áp dụng được kỹ thuật khai phá dữ liệu trong phân tích mạng xã hội, thì trước tiên phải định nghĩa được độ đo khoảng cách (distance measure) giữa các đỉnh, cạnh của đồ thị. Khi các cạnh của đồ thị được gắn nhãn thì các nhãn này có thể được sử dụng như là độ đo khoảng cách, tùy thuộc vào những gì mà chúng đại diện. Nhưng khi các cạnh không có nhãn, như đồ thị "bạn bè" thì cần phải định nghĩa độ đo khoảng cách giữa các đỉnh.

Trước tiên ta quy ước, những đỉnh gần nhau (closed) nếu chúng có cạnh nối trực tiếp giữa chúng, ngược lại là những đỉnh xa nhau (distant). Khoảng cách giữa đỉnh x và $y \in V$, ký hiệu là d(x, y), có thể định nghĩa d(x, y) theo hai cách:

- d(x, y) = 0 nếu $(x, y) \in E$, ngược lại thì d(x, y) = 1.
- Hoặc d(x, y) = 1 nếu có cạnh nối giữa chúng, và bằng ∞ khi chúng xa nhau, không có cạnh nối giữa chúng.

Tuy nhiên, cả hai trường hợp trên đều không phải là định nghĩa độ đo khoảng cách thực sự (metric), bởi chúng không thỏa mãn bất đẳng thức tam giác. Dễ nhận thấy, nếu có cạnh nối A với B và cạnh nối B với C, thì không có gì đảm bảo có cạnh nối A với C.

Có nhiều độ đo (measures) khác nhau được sử dụng để phân loại, phân tích, đánh giá đồ thị mạng xã hội. Chúng thường được sử dụng bởi các nhà nghiên cứu để phân tích các đặc điểm của mạng xã hội cần được xem xét. Các phép đo quan trọng nhất được xác định phần lớn đều dựa trên lý thuyết đồ thị. Tasleem Arif [8] sử dụng các hệ số cố kết mạng và hệ số trung tâm vector đặc trưng [79], [87], [94] để phân tích, đánh giá mạng xã hội. Freeman [32] đề xuất một tập hợp các độ đo (measures)

xác định độ đo trung tâm của các đỉnh, cạnh trên đồ thị, như độ đo trung tâm trực tiếp theo bậc của đỉnh, độ đo trung tâm lân cận và độ đo trung tâm trung gian (betweenness centrality) được sử dụng rất nhiều trong phân tích mạng xã hội và phát hiện các cộng đồng trên mạng xã hội.

1.2.1. Hệ số cố kết của mạng

Trong phân tích mạng xã hội có rất nhiều hệ số để so sánh các mạng xã hội với nhau, một trong những hệ số quan trọng nhất đó là hệ số cố kết (density cohesion) [8]. Khi hệ số cố kết của mạng càng lớn, mức độ gắn kết, sự chặt chẽ của các mối quan hệ giữa các thực thể, tác nhân trong mạng càng lớn, và do đó, sự tương trợ, hỗ trợ, … giữa các tác nhân cũng càng nhiều, càng hiệu quả hơn, sự điều tiết của mạng đối với hành vi của tác nhân cũng mạnh mẽ hơn và ngược lại.

Định nghĩa 1.2. Tính cố kết của mạng lưới là tỷ lệ giữa tổng các mối liên hệ thực tế trong mạng và tổng các mối quan hệ lý thuyết của nó (tức là tổng các mối quan hệ có thể có của mạng). Hệ số cố kết của đồ thị G, được tính như sau:

$$D_G = \frac{2\,k}{n\,(n-1)} \qquad (1.1)$$

Trong đó, k là tổng các mối liên hệ thực tế của mạng, k = |E| và n = |V|. Giá trị của hệ số này trong khoảng từ 0 đến 1. Khi giá trị này càng gần tới 1 thì tính cố kết của mạng lưới càng mạnh và do đó sự tương trợ, sự trao đổi thông tin, … giữa các thành viên trong mạng được diễn ra càng tốt và ngược lại. Theo Scott [95], hệ số cố kết của mạng lưới phụ thuộc vào số lượng tác nhân của nó, tức là khi càng có nhiều các tác nhân thì hệ số cố kết của nó càng nhỏ và ngược lại. Đối với những đồ thị đầy đủ (clique) thì hệ số cố kết là tuyệt đối, tức là $D_G = 1$.

1.2.2. Các hệ số đo tính trung tâm của tác nhân

Bên cạnh việc đo lường hệ số cố kết của cả mạng, trong phân tích mạng xã hội các nhà nghiên cứu thường xuyên sử dụng độ đo trung tâm (centrality) [11] để xác định vị trí của từng tác nhân trong mạng, bởi dù mạng có tính cố kết cao nhưng không phải mọi tác nhân đều có vị trí hay quyền lực như nhau trong mạng xã hội. Để đo lường được sự hơn kém giữa các tác nhân trong mạng, thường phải thông qua một số

đặc trưng cấu trúc mạng (structural features) như các đặc trưng về độ đo trung tâm trực tiếp theo bậc (degree), độ lân cận, hay độ gần nhau (closeness), và nhất là độ đo trung tâm trung gian (betweenness centrality) [8], [10], [29], [30], [32], [36], [48], [73], [74], [84], [98], [101].

- *Hệ số trung tâm trực tiếp (degree centrality)*

Hệ số này giúp chúng ta đo lường được số lượng của các mối quan hệ trực tiếp của một tác nhân nào đó (bậc của đỉnh trong đồ thị) với các thành viên khác trong mạng xã hội.

Định nghĩa 1.3. Hệ số trung tâm trực tiếp C_D của tác nhân (đỉnh) v trên đồ thị G, được tính theo bậc của nó, nghĩa là:

$$C_D(v) = deg(v) \quad (1.2)$$

Trong đó, deg(v) là số bậc của đỉnh v.

Bậc của đỉnh thường là độ đo hiệu quả cao phản ánh tầm quan trọng hoặc tầm ảnh hưởng (influence, importance) của một tác nhân (đỉnh). Trong nhiều mạng xã hội, những người có nhiều liên kết với người khác trong mạng thì luôn có xu hướng là có tầm ảnh hưởng lớn hơn và được nhiều người theo dõi hơn. Ví dụ: diễn viên điện ảnh nổi tiếng, ngôi sao ca nhạc, chính trị gia nổi tiếng, …

Các độ đo trung tâm thường được sử dụng cho cả các mạng đối xứng (đồ thị tương ứng là vô hướng) và mạng phi đối xứng (đồ thị tương ứng là có hướng).

Giả sử $K \in R^n$ là vector bậc của các đỉnh và $I \in R^n$ là vector đơn vị (tất cả thành phần là 1), R là tập số nguyên. Khi đó:

$$K = AI \quad (1.3)$$

Nếu mạng là đồ thị G có hướng, thì người ta thường định nghĩa hai độ đo trung tâm: theo bậc vào K^{in} (in degree) và bậc ra K^{out} (out degree). Bậc vào của một đỉnh là số các cạnh hướng tới đỉnh đó còn bậc ra là số cạnh đi tới những đỉnh khác. Những đỉnh có bậc vào K^{in} cao hơn sẽ có độ đo trung tâm cao hơn. Những đỉnh có bậc ra K^{out} cao hơn sẽ có mức độ uy tín (prestigious) cao hơn.

Định nghĩa 1.4. Độ đo trung tâm theo bậc vào/ ra: Giả sử $A \in \{0, 1\}^{n\times n}$ là ma trận liền kề của đồ thị định hướng và K^{in}, $K^{out} \in R^n$ là các vectors bậc vào, ra tương ứng. Khi đó

$K^{out} = A^T I$ (Tổng các cột của A); (1.4)

$K^{in} = AI$ (Tổng các hàng của A). (1.5)

Độ đo trung tâm theo bậc là độ đo đơn giản nhất trong số các độ đo trung tâm của mạng [32].

- *Hệ số trung tâm lân cận (closeness centrality)*

Hạn chế của hệ số trung tâm trực tiếp là chỉ tính các mối quan hệ trực tiếp của tác nhân. Một tác nhân lân cận với các tác nhân khác trong mạng không chỉ bao gồm những quan hệ trực tiếp mà còn có nhiều quan hệ gián tiếp. Tính lân cận cũng là một trong những tiêu chí quan trọng thể hiện vị thế của tác nhân trong mạng.

Một cách trực quan, hai tập các thực thể là gần nhau nếu chúng là lân cận của nhau. Trong lý thuyết đồ thị [108], độ gần nhau là độ đo trung tâm của các đỉnh trong một đồ thị. Những đỉnh che bóng các đỉnh khác (những đỉnh có khuynh hướng có những khoảng cách trắc địa ngắn nhất) (short geodesic distances) tới những đỉnh khác sẽ có độ gần nhau nhiều hơn. Độ gần nhau là độ đo trung tâm khá phức tạp. Nó được định nghĩa theo những khoảng cách trắc địa, là số các đường đi ngắn nhất giữa đỉnh v và những đỉnh khác mà nó có đường đi tới.

Định nghĩa 1.5. Hệ số trung tâm lân cận C_{Cl} (gọi tắt là độ lân cận, độ gần nhau) của đỉnh v được định nghĩa như sau:

$C_{Cl}(v) = \sum_{t\in V\setminus v} \sigma_{vt} /(n-1))$ (1.6)

Trong đó, σ_{vt} là số đường đi ngắn nhất đi v đến t. Độ gần nhau được xem như là độ dài mà luồng thông tin có thể trải qua từ một đỉnh cho trước tới những đỉnh khác trên mạng. Một số người định nghĩa độ gần nhau khác có thể tỷ lệ thuận hoặc nghịch nhau về số lượng, nhưng về cách mà lượng thông tin truyền thông trên mạng là như nhau.

Định nghĩa 1.6. Độ gần nhau $C_{Cl}(v)$ của đỉnh v được định nghĩa là tỷ lệ nghịch với tổng các khoảng cách trắc địa tới tất cả các đỉnh của V:

$$C_{Cl}(v) = 1 / \sum_{t \in V \setminus v} \sigma_{vt} \quad (1.7)$$

Độ gần nhau được nhiều người sử dụng để phân tích mạng xã hội [37], [52], [54], [76], [78].

- *Hệ số trung tâm trung gian (betweenness centrality)*

Độ đo trung tâm của mạng hay đồ thị được xác định theo hai cách phổ biến. Cách thứ nhất dựa vào lý thuyết đồ thị, trung tâm của đồ thị được đánh giá theo bậc, hay độ lân cận của các đỉnh trong đồ thị. Những đỉnh có bậc cực đại có thể được xem như là tâm điểm của đồ thị. Đo theo cách này thì việc ứng dụng sẽ bị hạn chế, bởi nó chỉ áp dụng được cho những bài toán như: thiết kế truyền thông với mục đích nhằm đạt được hiệu quả truyền thông cực đại. Cách thứ hai là dựa vào ưu thế trội (domination) của các đỉnh. Một thực thể (đỉnh) có ưu thế trội là thực thể có thể điều khiển sự truyền thông trên mạng (đồ thị).

Theo quan điểm của Freeman [32], một tác nhân nào đó trong mạng có thể ít gắn kết với các thành viên khác trong mạng (tức hệ số trung tâm trực tiếp thấp, bậc của đỉnh không cao), cũng không "gần gũi" lắm với mọi thành viên trong mạng (tức hệ số trung tâm lân cận thấp), nhưng lại là "cầu nối" (bridge), là "trung gian" cần thiết trong mọi cuộc trao đổi trong mạng. Nếu một tác nhân đóng vai trò trung gian càng lớn trong mạng lưới, tác nhân đó sẽ càng ở vị trí thuận lợi trong việc "kiểm soát" các giao dịch, các thông tin trong mạng, tác nhân đó cũng tác động đến mạng một cách dễ dàng bằng cách thanh lọc hoặc "lái" thông tin lưu chuyển trong mạng theo hướng có lợi cho mình nếu muốn; đồng thời tác nhân đó cũng đứng ở vị trí tốt nhất để thúc đẩy sự phối hợp giữa các thành viên khác trong mạng. Freeman [32] đã đề xuất độ đo trung tâm trung gian *(betweenness centrality)* của một đối tượng trong mạng xã hội, là số các cá thể có thể trao đổi với nhau thông qua đối tượng đó. Những đỉnh (cá thể) xuất hiện trên nhiều đường đi ngắn nhất giữa các đỉnh có độ đo trung tâm trung gian cao hơn những đỉnh không nằm trên những đường đi ngắn nhất đó.

Chúng ta xét một đồ thị đơn liên thông G = (V, E) (những đỉnh độc lập không cần xét). Xét cặp đỉnh $\{v_i, v_j\}$ bất kỳ, không phân biệt thứ tự đỉnh đầu, đỉnh cuối. Giữa chúng có thể có một hoặc nhiều đường đi. Nếu có đường đi giữa chúng thì độ

dài đường đi là bằng số cạnh (tổng trọng số trên các cạnh đối với đồ thị có trọng số) trên đường đi đó. Trong số các đường đi đó sẽ có một số đường đi ngắn nhất. Nếu (v_i, v_j), $(v_j, v_i) \in E$, thì đường đi ngắn nhất sẽ có độ dài là 1. Trường hợp đường đi ngắn nhất có độ dài (tổng số cạnh trên đường đi) lớn hơn 1 thì chắc chắn phải có ít nhất một đỉnh khác nằm trên đường đi ngắn nhất nối giữa v_i với v_j và những đỉnh này có tiềm năng để điều khiển sự liên thông hay truyền thông giữa các đỉnh v_i, v_j.

Cho đồ thị G = (V, E) có n đỉnh, độ đo trung tâm trung gian $C_B(v)$ của đỉnh v được xác định như sau:

- Với mỗi cặp đỉnh (s, t), tính tất cả các đường đi ngắn nhất nối giữa chúng - σ_{st};
- Với mỗi cặp đỉnh (s, t), tính phân số giữa những đường đi ngắn nhất $\sigma_{st}(v)$ có đi qua v và số các đường đi ngắn nhất từ s tới t là $\sigma_{st}(v)/\sigma_{st}$;
- Tính tổng các phân số của tất cả các cặp đỉnh (s, t).

Ta ký hiệu σ_{st} là số đường đi ngắn nhất đi từ s tới t, và $\sigma_{st}(v)$ là số đường đi ngắn nhất đi từ s tới t và có đi qua v.

Định nghĩa 1.7. Độ đo trung tâm trung gian kí hiệu là $C_B(v)$ của đỉnh v được xác định như sau:

$$C_B(v) = \sum_{s \neq t \neq v} \sigma_{st}(v) / \sigma_{st} \quad (1.8)$$

Hệ số này cũng có giá trị trong khoảng từ 0 đến 1. Khi một tác nhân nào đó có độ đo trung tâm trung gian càng gần đến 1 thì số lượng quan hệ giữa các tác nhân khác phải "thông qua" tác nhân này càng nhiều và do đó ảnh hưởng của tác nhân này trên mạng cũng càng lớn.

Chúng ta nhận thấy, độ đo trung tâm trung gian của đỉnh v đạt được giá trị cực đại khi mọi đỉnh khác trong G đều có cạnh nối với đỉnh v và đỉnh v nằm trên tất cả các đường đi ngắn nhất có độ dài lớn hơn 1. Những đồ thị như thế sẽ có dạng hình sao (star) hoặc hình bánh xe (wheel) [8], [32].

Tuy nhiên, việc sử dụng những độ đo này chỉ phù hợp cho những mạng, trong đó khái niệm độ đo trung tâm trung gian (betweenness centrality) được xem là quan trọng trong khả năng ảnh hưởng tới quá trình xử lý sự liên kết giữa các đỉnh.

Leavitt [32] đã nghiên cứu mối quan hệ giữa điểm trung tâm và mức độ thỏa mãn (satisfaction) của con người cùng tham gia giải quyết các vấn đề đặt ra. Mỗi người tham gia có một mẩu (một phần) tin cần thiết để giải một bài toán. Mỗi người chỉ có thể trao đổi với một số người được chỉ định khác và bài toán được giải khi tất cả các mẩu tin được chia sẻ giữa những người cùng tham gia giải bài toán đó. Leavitt đã tính toán độ đo trung tâm của đỉnh theo hàm độ dài của đường đi hoặc khoảng cách giữa các đỉnh trên đồ thị. Qua đó xác định được mối quan hệ nguyên thủy giữa độ dài đường đi với mức độ thỏa mãn của những người tham gia giải quyết công việc. Thông qua trao đổi với những người khác, mỗi người có thể nhận được những thông tin quan trọng để giải quyết bài toán và họ cảm thấy thỏa mãn, nghĩa là độ đo trung tâm trung gian lớn hơn thì mức độ thỏa mãn sẽ lớn hơn. Đối với những mạng đã được Leavitt nghiên cứu, độ dài đường đi và độ đo trung tâm trung gian của các cạnh có mối liên hệ chặt chẽ với nhau. Cạnh nối giữa các nhóm (cộng đồng) có ảnh hưởng rất lớn đến dòng chảy của thông tin giữa các tác nhân (đỉnh) khác nhau, đặc biệt là trong trường hợp thông tin lưu truyền trong mạng chủ yếu theo con đường ngắn nhất [32].

Girvan-Newman [37] đề xuất định nghĩa độ đo "trung tâm trung gian" (betweenness centrality) của cạnh (a, b) là số các cặp đỉnh x và y mà cạnh (a, b) nằm trên đường đi ngắn nhất nối giữa x và y. Thuật toán điển hình, phổ biến nhất trong các thuật toán sử dụng độ đo này để phát hiện cộng đồng trên mạng xã hội là thuật toán Girvan-Newman.

Brandes [19] đã đề xuất thuật toán tính độ đo trung tâm trung gian theo kỹ thuật tích lũy phụ thuộc (dependency accumulation technique). Thuật toán này được thực hiện qua 2 bước:

Bước 1. Tính độ dài và số đường đi ngắn nhất giữa các cặp đỉnh

Bước 2. Tính tổng tất cả các phụ thuộc cặp đỉnh.

Đối với đồ thị liên thông, vô hướng và không trọng số G = (V, E) thì thuật toán Brandes có độ phức tạp thời gian tính toán là O(mn) với $m = |E|, n = |V|$.

Trong những năm gần đây, độ đo trung tâm trung gian đã được rất nhiều các nhà khoa học quan tâm, nghiên cứu để phân tích, phát hiện các cộng đồng trên mạng

xã hội [22], [33], [37], [40], [76], [78], [84]. Việc tính toán độ đo trung tâm trung gian và độ gần nhau của tất cả các đỉnh trên đồ thị liên quan đến việc tính số các đường đi ngắn nhất giữa các cặp đỉnh trên đồ thị. Như chúng ta đã biết, bài toán tìm đường đi ngắn nhất trên đồ thị là bài toán NP-đầy đủ. Hơn nữa, đồ thị mạng xã hội có kích thước ngày một lớn do số lượng người tham gia (đỉnh) cũng như các mối quan hệ (cạnh) giữa những người tham gia trên mạng xã hội ngày càng gia tăng nhanh chóng. Do vậy, nhiều thuật toán tính toán độ đo trung tâm trung gian và phát hiện cộng đồng mạng xã hội không thực sự mang lại hiệu quả cao.

Trong Chương 2 nghiên cứu sinh tập trung nghiên cứu các tính chất tương đương của các lớp đỉnh dựa vào độ đo trung tâm trung gian từ đó thực hiện đề xuất thuật toán rút gọn đồ thị mạng xã hội nhưng vẫn bảo toàn giá trị độ đo trung tâm trung gian. Đồng thời phát triển một số thuật toán cải tiến thời gian phát hiện cộng đồng trên mạng xã hội.

1.3. Bài toán phát hiện cộng đồng mạng xã hội

Phát hiện cộng đồng trên mạng xã hội là một trong những lĩnh vực nghiên cứu quan trọng và nổi bật hàng đầu trong phân tích mạng xã hội. Phát hiện cộng đồng trên mạng xã hội có tầm quan trọng lớn trong xã hội học, sinh học và khoa học máy tính, ... Phát hiện cộng đồng trên mạng xã hội gặp thách thức lớn đặc biệt sự phức tạp tính toán bị chi phối bởi hai yếu tố chính. Yếu tố đầu tiên phải kể đến là kích thước của mạng xã hội rất lớn như mạng xã hội Facebook đã đạt đến hàng tỷ người dùng. Vì vậy, cần có giải pháp thích hợp để giảm kích thước của đồ thị mạng xã hội ban đầu theo một cách thức có thể quản lý và kiểm soát được. Do đó mà chi phí tính toán giảm, thời gian tính toán giảm nhưng không làm giảm chất lượng của giải pháp hay tính chất của mạng xã hội ban đầu. Yếu tố thứ hai liên quan đến bản chất của mạng xã hội là động, cấu trúc của mạng biến đổi, phát triển không ngừng theo thời gian. Chính những thách thức này đã thu hút được một số lượng lớn các nhà khoa học quan tâm nghiên cứu liên tục trong những năm qua.

1.3.1. Cộng đồng mạng xã hội

Cộng đồng mạng xã hội (social network community) là một tập hợp các tác nhân tương tác với nhau thông qua các phương tiện truyền thông cụ thể, có khả năng vượt qua những ranh giới địa lý và chính trị để theo đuổi lợi ích hay mục tiêu chung. Cộng đồng mạng xã hội là một nhóm các thực thể có những tính chất tương tự nhau, liên kết chặt chẽ với nhau hơn và cùng đóng một vai trò nhất định trong mạng xã hội. Cộng đồng mạng xã hội còn được định nghĩa là những cấu trúc xã hội rất mạnh mẽ và được liên kết với nhau dựa trên những mối quan hệ, quan tâm chung. Nó có thể là một sở thích, một lĩnh vực mà các thành viên trong cộng đồng cùng quan tâm, hay một mục tiêu, dự án chung, theo vị trí địa lý, hoặc nghiệp vụ. Người tham gia vào cộng đồng, vì quan tâm đến lợi ích chung này mà gắn kết các thành viên cộng đồng với nhau. Một cộng đồng được mô tả như một nhóm các đỉnh (đồ thị con) cùng chia sẻ các thuộc tính chung hoặc đóng vai trò tương tự trong một mạng. Một cộng đồng có thể được định nghĩa là một tập hợp các đỉnh có mật độ liên kết giữa các đỉnh cao và mật độ liên kết thấp với phần còn lại của mạng. Nói cách khác một cộng đồng được tạo ra bởi một tập hợp các tác nhân có tương tác thường xuyên với nhau hơn so với những cá nhân khác ngoài cộng đồng. Do vậy, nó thường là một tập hợp như: bạn bè, đồng nghiệp, nhóm những người có cùng sở thích, cùng chuyên môn, mối quan tâm, … [22], [31], [40], [52], [53], [76], [77], [78], [99], [106], [107].

Trong lý thuyết đồ thị, chúng ta có thể định nghĩa cộng đồng một cách hình thức như sau:

Định nghĩa 1.8. Cho trước đồ thị G = (V, E), với V là tập các đỉnh, E là tập các cạnh. Các cộng đồng là tập các đồ thị con của G, C = {G_1, G_2, …, G_k}, với G_i = (V_i, E_i), i = 1, 2, …, k sao cho:

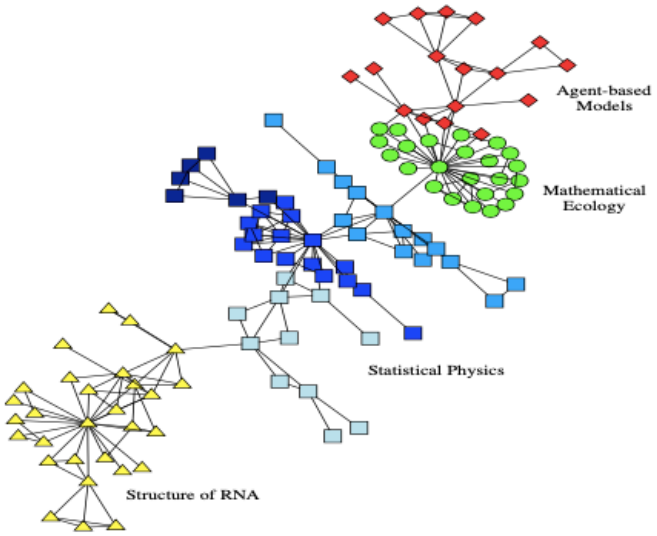
(i) $\forall i \neq j = 1, 2, \ldots, k$, $V_i \cap V_j = \phi$, các cộng đồng rời nhau

(ii) $\bigcup_{i=1}^{k} V_i = V$ và $\bigcup_{i=1}^{k} E_i \subseteq E$, cộng đồng là các đồ thị con của G

(iii) Các đỉnh trong cùng một cộng đồng có liên kết (cạnh nối) với nhau nhiều hơn số liên kết với các đỉnh ở những cộng đồng khác, nghĩa là: $|E_i| > |E_{i,j}|$, với $E_{i,j} = \{(u, v) \in E - (E_i \cup E_i), u \in V_i, v \in V_j$ và $i \neq j = 1, 2, \ldots, k\}$.

Tuy nhiên, trong bài toán phát hiện cộng đồng trên mạng xã hội, phần lớn chúng ta chỉ quan tâm tới việc xác định các tập đỉnh (tác nhân) V_i, i = 1, 2, …, đại diện cho cộng đồng mạng xã hội.

Ví dụ 1.1. Các cộng đồng trong mạng lưới các cộng tác nghiên cứu của các nhà khoa học làm việc tại Viện Santa Fe.



Hình 1.1. Cộng đồng trong mạng lưới các nhà khoa học làm việc tại viện Sante Fe [31]

Hình 1.1 biểu diễn các thành phần kết nối lớn nhất trong mạng lưới các cộng tác nghiên cứu của các nhà khoa học làm việc tại Viện Santa Fe (SFI). Đồ thị bao gồm 118 đỉnh đại diện cho các nhà khoa học làm việc tại SFI và các cộng tác viên của họ. Các cạnh được liên kết giữa các nhà khoa học khi họ đã công bố cùng với nhau ít nhất một bài báo. Ở mạng này ta quan sát được một số cộng đồng, mỗi cộng đồng biểu hiện cho những tác giả đã cùng nhau công bố một hay nhiều bài báo khoa học. Mặt khác ta cũng thấy giữa các cộng đồng trong mạng trên chỉ có một số ít mối liên kết. Các đỉnh cùng màu là cùng một cộng đồng theo các lĩnh vực nghiên cứu của SFI.

Trên mạng xã hội, việc trích xuất và phát hiện được những cộng đồng là rất hữu ích vì nó giúp chúng ta nghiên cứu được cấu trúc tổng thể của mạng. Khái niệm của khám phá, phát hiện cộng đồng tương tự như phân phân cụm đồ thị nhưng có một số khác biệt như trong phân cụm đồ thị, số lượng nhóm và kích thước của chúng đã được

biết trước, nhưng trong trường hợp phát hiện ra cộng đồng, chúng ta không biết về số lượng cộng đồng trong mạng và cộng đồng cũng có thể không cùng kích cỡ với nhau.

Một số ứng dụng chính của bài toán phát hiện cộng đồng trên mạng xã hội [3], [4], [25] là:

- Phát hiện cộng đồng có thể được sử dụng trong tư vấn thông tin và xác định được những cộng đồng có cùng một số quan tâm, sở thích tương tự. Ví dụ: Trong lĩnh vực kinh doanh, việc phát hiện và xác định được các cộng đồng, nhóm khách hàng có chung sở thích, mối quan tâm trong một mạng lưới có thể giúp ta xây dựng được hệ thống chăm sóc khách hàng, hệ thống tư vấn với các chính sách kinh doanh đạt hiệu quả hơn.
- Cộng đồng cũng sẽ giúp chúng ta hiểu cấu trúc của mạng xã hội, làm rõ các thuộc tính và chức năng của mạng xã hội.
- Phát hiện các cộng đồng để hiểu hành vi của mạng xã hội trong quy mô lớn vì nó sẽ làm rõ các quá trình chia sẻ thông tin và truyền bá thông tin.
- Các phương pháp phát hiện cộng đồng có lợi thế lớn trong việc định tuyến nhận thức trong xã hội và ngăn chặn thông tin độc hại trên mạng xã hội. Ví dụ: Việc phát hiện và xác định được các thông tin độc hại chính là cơ sở để đưa ra các quyết định khuyến cáo nâng cao cảnh giác, chủ động phòng chống, góp phần giữ vững an ninh, trật tự an toàn xã hội.
- Mạng xã hội loài người thể hiện cộng đồng mạnh mẽ. Một mạng lưới có cộng đồng mạnh bao gồm các cộng đồng, các cộng đồng này có nhiều kết nối trong đó và ít kết nối giữa các cộng đồng. Ví dụ: Cộng đồng không chỉ ảnh hưởng đến sự lây lan của bệnh truyền nhiễm trong cộng đồng nhưng cũng bảo vệ mạng khỏi dịch bệnh quy mô lớn.
- Trong hệ sinh học và hệ chăm sóc sức khỏe, có nhiều thuật toán phát hiện cộng đồng được phát triển cho các mạng xã hội cũng có thể được mở rộng thành công cho các mạng sinh học. Ví dụ: Phát hiện cộng đồng được sử dụng trong Calderone để so sánh các mạng tương tác Alzheimer và Parkinson.

1.3.2. Các thuật toán phát hiện cộng đồng mạng xã hội

Mục tiêu của bài toán phát hiện cộng đồng mạng xã hội là từ các mạng xã hội cho trước, phát hiện được các cộng đồng nằm trong đó và tìm hiểu về mối liên hệ bên trong các cộng đồng cũng như giữa các cộng đồng với nhau, mối liên hệ đó có ảnh hưởng thế nào đến toàn mạng xã hội. Một tập hợp các đỉnh trên đồ thị được coi là một cộng đồng nếu mật độ cạnh giữa các đỉnh bên trong nó cao hơn so với mật độ của các cạnh giữa đỉnh của nó và những đỉnh khác bên ngoài.

Phát hiện cộng đồng nhằm mục đích nhóm các đỉnh liên kết mạnh theo các mối quan hệ giữa chúng để tạo thành các đồ thị con từ đồ thị ban đầu. Các mạng xã hội thường được biểu diễn dưới dạng đồ thị đơn nên việc phát hiện cộng đồng trên mạng xã hội dựa trên cơ sở lý thuyết đồ thị còn được gọi là bài toán phân cụm đồ thị.

Bài toán: Phát hiện các cộng đồng trong mạng xã hội.

Đầu vào: Đồ thị mạng xã hội G = (V, E) gồm tập V có các đỉnh: v_1, v_2,…, v_n và tập E các cạnh E = $\{(v_i,v_j)\}$.

Đầu ra: Tập các cộng đồng mạng xã hội C.

Trong nhiều thập kỷ qua, số lượng các giải pháp phát hiện cộng đồng trên mạng xã hội đã được nghiên cứu là rất nhiều, thường xuyên và liên tục [3], [12], [17], [21], [22], [24], [37], [39] [44], [45], [49], [52], [59], [66], [67], [69], [70], [72], [77], [80], [104], [109], [116], [117]. Về cơ bản, các thuật toán này được chia thành 4 nhóm thuật toán chính đó là nhóm thuật toán phát hiện cộng đồng truyền thống, nhóm thuật toán phát hiện cộng đồng dựa trên tối ưu hóa độ đo đơn thể, nhóm thuật toán phát hiện cộng đồng dựa vào độ đo trung tâm trung gian và nhóm thuật toán phát hiện cộng đồng dựa trên lan truyền nhãn. Dưới đây sẽ trình bày chi tiết về các nhóm thuật toán này.

1.3.2.1. Nhóm thuật toán phát hiện cộng đồng truyền thống

Nhóm thuật toán phát hiện cộng đồng truyền thống bao gồm các thuật toán: Phân cụm đồ thị, phân cụm phân cấp, phân cụm phân hoạch, phân cụm theo phổ và thuật toán phân chia.

- **Thuật toán phân cụm đồ thị (Graph clustering)**

Thuật toán phân cụm đồ thị thực hiện chia đồ thị ban đầu thành các đồ thị con (cụm) có kích thước được xác định trước và số cạnh trong một cụm nhiều hơn so với số cạnh giữa các cụm hay một phân vùng và được đánh giá là tốt nếu số cạnh trung gian giữa phân vùng đó với phân vùng khác là ít [31]. Giải thuật phân cụm tìm ra các cụm đỉnh bằng cách sử dụng độ đo giữa các cặp đỉnh. Minh họa điển hình về thuật toán phân cụm theo đồ thị là thuật toán Kernighan - Lin của Kernighan và Lin [50]. Nhược điểm chính của thuật toán Kernighan - Lin là phải chỉ định trước kích thước của hai cộng đồng. Tuy nhiên, ngay cả khi hạn chế này có thể khắc phục thì thuật toán Kernighan - Lin vẫn còn hạn chế giống như các thuật toán phân chia khác là nó chỉ chia mạng thành hai nhóm chứ không phải là một số nhóm tùy ý. Vì những lý do kể trên, các phương pháp phân vùng đồ thị hay phân cụm đồ thị không thực sự hiệu quả để phân tích dữ liệu mạng lớn.

- **Thuật toán phân cụm phân cấp (Hierarchical agglomerative clustering)**

Đồ thị có thể chứa cấu trúc phân cấp, mỗi cộng đồng có thể là một tập hợp các cộng đồng nhỏ ở các cấp độ khác nhau [31]. Trong các trường hợp như vậy, kỹ thuật phân cụm phân cấp [92] thường được sử dụng để xác định cộng đồng nhiều cấp của đồ thị. Kỹ thuật phân cụm phân cấp dựa trên đo độ tương tự của đỉnh. Chúng không cần xác định trước kích thước và số lượng các cộng đồng. Nhưng chất lượng phát hiện cộng đồng không cao do việc lựa chọn độ đo tương tự của đỉnh. Thuật toán sắp xếp dữ liệu ban đầu thành cấu trúc có dạng hình cây, cây này được xây dựng theo kỹ thuật đệ quy theo hai phương pháp Bottom-up và Top-down.

- Thuật toán phân cụm phân cấp điển hình sử dụng chiến lược phân cụm Top-down là thuật toán BIRCH (Balanced Iterative Regucing and Clustering using Hierarchies) [109].

Đầu vào: Cơ sở dữ liệu gồm n đối tượng, ngưỡng T cho trước.
Đầu ra: k cụm dữ liệu

Bước 1. Thuật toán duyệt tất cả các đối tượng trong cơ sở dữ liệu và khởi tạo một cấu trúc cây. Một đối tượng được chèn vào đỉnh lá gần nhất tạo thành cụm con. Nếu đường kính của cụm con này lớn hơn ngưỡng T thì đỉnh lá được tách. Khi một đối tượng thích hợp được chèn vào đỉnh lá, tất cả các đỉnh trỏ tới gốc của cây được cập nhật với các thông tin cần thiết.

Bước 2. Nếu cây hiện thời không có đủ bộ nhớ thì tiến hành xây dựng một cây nhỏ hơn bằng cách điều khiển bởi ngưỡng T, khi tăng ngưỡng T thì đồng thời sẽ nhập một số cụm con thành cụm lớn, làm cho cây nhỏ hơn.

Bước 3. Thực hiện phân cụm, các đỉnh lá của cây lưu giữ các đại lượng thống kê của các cụm con. Thuật toán sử dụng các đại lượng thống kê này để áp dụng một số kỹ thuật phân cụm như k-means.

Bước 4. Phân phối lại các đối tượng dữ liệu bằng cách dùng các đối tượng trọng tâm cho các cụm đã được khám phá từ Bước 3.

Thuật toán BIRCH gặp một số hạn chế như: chất lượng phân cụm không cao do việc lựa chọn ngưỡng T ban đầu ảnh hưởng rất lớn tới chất lượng phân cụm.

Huawei Shen và các cộng sự đề xuất thuật toán EAGLE (agglomerativ**E** hierarchic**A**l clusterin**G** based on maxima**L** cliqu**E**) [44] để phát hiện những cộng đồng cả gối nhau lẫn phân cấp. Thuật toán đề cập đến tập các cliques cực đại và thực hiện theo kỹ thuật gộp nhóm (tích tụ dần). Clique C [57] của đồ thị G là tập các đỉnh của một đồ thị con đầy đủ của G, sao cho giữa hai đỉnh bất kỳ của C đều có cạnh nối giữa chúng. k-Clique C của đồ thị G là tập các đỉnh của một đồ thị con của G, sao cho đường đi ngắn nhất (đường trắc địa) giữa hai đỉnh v, w bất kỳ đều có độ dài $1 \leq d(v, w) \leq k$.

Một cộng đồng được xem như là một tập các đỉnh, trong đó chúng liên kết với nhau nhiều hơn phần còn lại của mạng. Điều này chỉ ra rằng cộng đồng có mật độ liên kết tương đối cao. Nói chung, mật độ liên kết của một clique là cao nhất trong tất cả các tập đỉnh con của mạng (đồ thị). Những cộng đồng có mật độ liên kết cao thường chứa một clique lớn, được xem như là lõi (core) của cộng đồng. Dựa vào quan sát này, thuật toán EAGLE được phát triển để phân cụm phân cấp và gộp nhóm để nghiên

cứu cộng đồng. Sự chồng lấp (gối nhau) giữa những cliques cực đại (không phải là tập con của clique nào khác) đảm bảo rằng có sự giao nhau giữa các cộng đồng và cấu trúc phân cấp những cộng đồng này được phát hiện bởi quá trình phân cụm phân cấp tích tụ dần. Do vậy, mở rộng độ đo đơn thể (modularity) của phương pháp phân hoạch mạng, chúng ta sử dụng Q_c để đánh giá chất lượng các thuật toán phát hiện cộng đồng của mạng.

Trong thuật toán EAGLE, trước tiên chúng ta cần tìm tất cả các cliques trong mạng. Lưu ý rằng, không phải tất cả các cliques cực đại tìm được đều được sử dụng để phát hiện những cộng đồng gối nhau và phân cấp trong mạng. Những clique cực đại mà đỉnh của chúng là đỉnh của những clique cực đại lớn hơn, được gọi là clique cực đại thứ cấp (subordinate maximal cliques).

- **Thuật toán phân cụm phân hoạch (Partitional clustering)**

Phân cụm phân hoạch [26], [31], [35] phân chia đồ thị ban đầu thành một số cụm không chồng chéo được xác định trước. Mục tiêu là chia đồ thị thành các cụm để tối ưu hóa hàm mục tiêu dựa trên đo độ khác nhau giữa các đỉnh. Một số hàm mục tiêu được sử dụng là K-mean, K-clustering và K-center. Minh họa điển hình về kỹ thuật phân cụm phân hoạch bao gồm phân cụm K-means [69] và phân cụm K-means mờ [14]. Thuật toán K-Means có ưu điểm là đơn giản, dễ cài đặt và có thể ứng dụng với tập dữ liệu lớn. Tuy nhiên, thuật toán K-Means có hạn chế là hiệu quả của thuật toán phụ thuộc vào việc lựa chọn số nhóm K (phải xác định trước) và độ phức tạp thực hiện vòng lặp tính toán khoảng cách lớn khi số cụm K và dữ liệu phân cụm lớn, ngoài ra còn hạn chế về kích thước cụm, mật độ cụm, hình dạng cụm, việc khởi tạo các tâm cụm được lựa chọn ngẫu nhiên ảnh hưởng lớn tới kết quả phân cụm.

Hạn chế đối với các thuật toán phân cụm phân hoạch là các thuật toán phân cụm phân hoạch thường phụ thuộc vào khoảng cách cơ bản giữa các điểm để lựa chọn các điểm dữ liệu nào có quan hệ là gần nhau với mỗi điểm khác và các điểm dữ liệu nào không có quan hệ hoặc có quan hệ là xa nhau với các điểm khác. Phương pháp này không thể xử lý được các cụm có hình dạng đặc biệt hoặc các cụm có mật độ điểm

dày đặc, đa chiều. Các thuật toán phân cụm phân hoạch có độ phức tạp tính toán cao khi thực hiện việc xác định nghiệm tối ưu toàn cục cho bài toán phân cụm dữ liệu.

- **Thuật toán phân cụm theo phổ (Spectral clustering)**

Phân cụm theo phổ bao gồm tất cả các kỹ thuật sử dụng ma trận véc tơ đặc trưng để phân chia dữ liệu dựa trên sự tương đồng về cặp giữa chúng [1], [26], [31] điển hình là các thuật toán: phân cụm theo phổ không chuẩn hóa, phân cụm theo phổ chuẩn hóa theo Shi và Malik [97] và Jordan và Weiss [2]. Điểm khác nhau giữa các thuật toán trên là chúng sử dụng các đồ thị Laplace khác nhau. Trong các thuật toán thủ thuật chính là thay đổi cách biểu diễn của các điểm dữ liệu trừu tượng.

Những vấn đề tồn tại khi sử dụng các thuật toán phát hiện cộng đồng truyền thống:

- Một lượng thông tin bị mất trong quá trình phân cụm dẫn đến chất lượng thuật toán phát hiện cộng đồng có độ chính xác thường không cao.

- Nhóm các phương pháp này chỉ tập trung vào các liên kết, kết nối và cấu trúc của đồ thị mạng xã hội mà không xem xét, chú ý đến các tương tác của người sử dụng mạng xã hội và ảnh hưởng của người dùng trên toàn mạng xã hội.

1.3.2.2. Nhóm thuật toán phát hiện cộng đồng dựa trên tối ưu hoá độ đo đơn thể

Độ đo đơn thể Q (Modularity Q) [14], [76], [77] được sử dụng để đánh giá chất lượng thuật toán phát hiện cộng đồng, độ đo đơn thể Q có giá trị càng lớn thể hiện độ chính xác của thuật toán càng cao, chất lượng việc phát hiện cộng đồng được đánh giá là tốt.

Nhóm thuật toán này gồm: thuật toán tìm kiếm tham lam, mô phỏng luyện kim, tối ưu hoá mở rộng và các thuật toán tiến hoá.

- **Thuật toán tìm kiếm tham lam (Greedy techniques)**

Thuật toán tìm kiếm tham lam của Newman [23], [78] là thuật toán đầu tiên được đề xuất nhằm tối ưu hóa độ đo đơn thể. Đây là một kỹ thuật tích tụ, trong đó ban đầu mỗi đỉnh thuộc về một mô đun riêng biệt, sau đó chúng được hợp nhất lặp lại dựa trên mức độ tăng các mô đun. Thuật toán có độ phức tạp thời gian tính toán

là $O(n^3)$, với n là số đỉnh của đồ thị. Thuật toán tìm kiếm tham lam điển hình là thuật toán đơn thể hóa Louvain (Louvain modularity) [43]. Thuật toán Louvain là một phương pháp phân chia cộng đồng và thực hiện lặp đi lặp lại việc phân chia cộng đồng nhiều lần để có được mô đun tối đa của toàn bộ mạng. Nó gán các cộng đồng khác nhau cho mỗi đỉnh và lặp lại việc hợp nhất các đỉnh dựa trên mức độ tăng của độ đo đơn thể. Thuật toán được lặp đi lặp lại cho đến khi không thể cải tiến thêm nữa. Độ phức tạp tính toán của thuật toán là O*(nlog(n))*. Các bước chính của thuật toán như sau:

Bước 1. Đầu tiên, gán một cộng đồng khác nhau cho mỗi đỉnh của mạng, vì vậy trong phân vùng ban đầu, số lượng cộng đồng nhiều như số đỉnh.

Bước 2. Đối với mỗi đỉnh, xem xét các đỉnh lân cận của nó và đánh giá giá trị tính mô đun sau khi loại bỏ một đỉnh khỏi cộng đồng của đỉnh đó và đặt đỉnh đó vào một trong những cộng đồng lân cận của nó. Nếu ΔQ dương, đỉnh vẫn ở trong cộng đồng ban đầu của nó, nếu không đỉnh được đặt trong cộng đồng cập nhật.

Bước 3. Lặp lại Bước 2 đến khi cộng đồng của tất cả các đỉnh không còn thay đổi.

Bước 4. Xây dựng một đồ thị mới và mỗi đỉnh đại diện cho một cộng đồng được phân vùng bởi Bước 3. Thực hiện Bước 2 và Bước 3 liên tục cho đến khi đạt được giá trị mô đun lớn nhất.

Thuật toán Louvain cần lặp lại liên tục để tìm ra cộng đồng tốt nhất cho mỗi đỉnh trong Bước 2 và quá trình này mất quá nhiều thời gian trong các mạng lớn. Vì vậy, các thuật toán cải tiến thuật toán Louvain là thuật toán Louvain-LPA [43], thuật toán IG [62] đã được đề xuất để khắc phục các hạn chế trên.

- **Thuật toán mô phỏng luyện kim (Simulated Annealing)**

Thuật toán mô phỏng luyện kim được giới thiệu bởi Kirkpatrich, Gellatt và Vecchi [54]. Thuật toán được xây dựng dựa trên mô phỏng luyện kim với các tham số điều khiển biến thiên theo chu trình tiến hóa của giải thuật. Đây là một công cụ hiệu quả để xử lí các bài toán tìm kiếm và tối ưu, đặc biệt là những bài toán có không gian tìm kiếm lớn. Hạn chế của thuật toán mô phỏng luyện kim là xác suất chấp nhận

lời giải kém hơn tùy thuộc vào tham số nhiệt độ, thuật toán dễ rơi vào cực trị địa phương, tốn nhiều thời gian tính toán vì phụ thuộc vào quá trình ngẫu nhiên hóa và thuật toán chỉ dừng lại ở mức chấp nhận được nên kết quả cuối cùng chưa cao.

- **Thuật toán tối ưu hóa mở rộng (Extremal Optimisation)**

Boettcher và các cộng sự [16] đã đề xuất thuật toán tối ưu hóa mở rộng như một kỹ thuật tìm kiếm heuristic cho mục đích chung cho các vấn đề tối ưu hóa và tổ hợp. Thuật toán tập trung vào việc tối ưu hóa các biến cục bộ. Duch và các cộng sự [27] đã sử dụng để tối ưu hóa độ đo đơn thể. Nó bắt đầu bằng cách chia ngẫu nhiên mạng thành hai phân vùng có cùng thứ tự và lặp lại việc đưa các đỉnh với liên kết thấp nhất đến các phân vùng khác. Sau mỗi lần thay đổi, các phân vùng đều có sự thay đổi, do đó nó tính toán lại liên kết cục bộ của nhiều đỉnh. Quá trình lặp lại cho đến khi đạt được giá trị tối ưu của độ đo đơn thể toàn mạng. Thuật toán tối ưu hóa mở rộng có độ phức tạp là O($n^2 logn$) với n là số đỉnh của đồ thị.

- **Các thuật toán tiến hóa (Evolutionary algorithms)**

Thuật toán tiến hóa là một thuật toán tối ưu hóa. Thuật toán này được chia thành hai nhóm dựa trên tối ưu hóa đơn mục tiêu và tối ưu hóa đa mục tiêu. Tối ưu hóa đơn mục tiêu trong đó mỗi điểm trong không gian tìm kiếm của bài toán được ánh xạ thành một giá trị mục tiêu vô hướng. Tối ưu hóa đa mục tiêu trong đó mỗi điểm trong không gian tìm kiếm của bài toán được ánh xạ thành một véc tơ các giá trị mục tiêu [20], [46].

Hạn chế của thuật toán tiến hóa là tập trung vào việc giải bài toán tối ưu tại mỗi thời điểm dựa trên một quần thể mà chưa có sự quan tâm đến việc giải quyết nhiều bài toán tối ưu khác nhau cùng lúc trên cùng một quần thể. Đồng thời đối với mỗi bài toán, mỗi giai đoạn khác nhau thì việc lựa chọn thuật toán heuristic hoặc thay đổi các tham số phù hợp để có được kết quả tối ưu là không dễ dàng thực hiện.

1.3.2.3. Nhóm thuật toán phát hiện cộng đồng dựa vào độ đo trung tâm trung gian

Bài toán phát hiện cộng đồng tập trung vào việc từ một mạng xã hội, tìm ra những cụm, nhóm cộng đồng có mối liên hệ chặt chẽ với nhau. Qua trực quan có thể

dễ dàng tìm ra những nhóm cộng đồng tập trung, nhưng không phải cộng đồng nào cũng được hình thành bằng các mối liên hệ chặt chẽ và dễ thấy, một số cộng đồng có thể được hình thành ẩn. Điều quan trọng là phải tìm được các cộng đồng tồn tại trong mạng xã hội.

Thay vì việc tìm kiếm những đỉnh trong đồ thị có độ gắn kết cao với nhau, phương pháp phát hiện cộng đồng bằng thuật toán phân chia được đưa ra như một cách giải quyết hữu hiệu. Để tránh các hạn chế của phương pháp phân cụm phân cấp, thay vì cố gắng để xây dựng một giải pháp tìm cạnh trung tâm của cộng đồng, chúng ta đi tìm những cạnh có độ đo trung tâm trung gian cao nhất, cạnh đó được gọi tên là cạnh nối giữa các cộng đồng. Girvan-Newman [76] cho rằng khi các cộng đồng được gắn kết với nhau thì đường đi giữa cộng đồng này đến cộng đồng khác sẽ đi qua các cạnh nối giữa các cộng đồng với tần suất cao. Mục đích chính của thuật toán là tìm những cạnh nối đó. Thay vì việc xây dựng cộng đồng bằng cách thêm vào các cạnh có độ đo trung tâm cao nhất, chúng ta sẽ xây dựng bằng cách loại bỏ dần các cạnh nối từ đồ thị ban đầu. Khi đó, các cộng đồng trong mạng sẽ bị ngắt kết nối với nhau, qua đó ta có thể xác định được cách phân vùng đồ thị thành các phần nhỏ riêng biệt. Để làm được việc này, điều quan trọng nhất của thuật toán là việc tính toán như thế nào, sử dụng tính chất nào để phát hiện ra những cạnh nối này, từ đó loại bỏ chúng ra khỏi đồ thị. Thuật toán đầu tiên được đề xuất bởi Freeman [32], [34]. Theo Freeman, các cạnh nối là cạnh có số lượng con đường ngắn nhất giữa các cặp đỉnh khác nhau chạy qua nó. Cạnh nối có ảnh hưởng rất lớn đến dòng chảy của thông tin giữa các đỉnh khác, đặc biệt là trong trường hợp thông tin lưu truyền trong mạng chủ yếu theo con đường ngắn nhất. Thuật toán điển hình, phổ biến nhất là thuật toán Girvan-Newman [37], [76].

Thuật toán Girvan-Newman (GN) [76] là phương pháp phân cụm phân cấp, và là một trong những thuật toán phát hiện cộng đồng mạng xã hội điển hình, phổ biến, và được sử dụng rộng rãi, thường xuyên. GN phát hiện các cộng đồng trên đồ thị mạng xã hội bằng cách sử dụng độ đo trung tâm trung gian của cạnh để loại bỏ các cạnh trung gian cao nhất trong mỗi lần lặp. Quá trình này sẽ được

tiếp tục, cho đến khi nó đạt đến các cộng đồng có độ đơn thể cao. Nói cách khác, các chức năng mô đun được sử dụng để đánh giá chất lượng của các cộng đồng được phát hiện.

Input: Đồ thị mạng xã hội G = (V, E) gồm V là tập các đỉnh và E là tập các cạnh.

Output: Tập các cộng đồng C_i và tập hợp các đỉnh thuộc các cộng đồng đó.

Thuật toán Girvan-Newman duyệt qua mỗi đỉnh v một lần và tính số đường đi ngắn nhất từ v tới những đỉnh khác có đi qua từng cạnh đó. Tư tưởng chính của thuật toán GN được thực hiện theo kỹ thuật phân cụm phân cấp như sau:

Bước 1. Tính độ đo trung tâm trung gian của tất cả các cạnh trong mạng,

Bước 2. Tìm những cạnh có độ đo trung tâm trung gian lớn nhất và loại bỏ chúng,

Bước 3. Tính lại độ đo trung tâm trung gian của tất cả các cạnh trong các thành phần còn lại,

Bước 4. Lặp lại từ bước 2 cho đến khi đến khi không có cạnh nào vượt qua ngưỡng của độ đo trung tâm trung gian cho trước hoặc không còn cạnh trung gian.

Thuật toán Girvan-Newman cho kết quả tương đối tốt trong nhiều trường hợp, mặc dù vậy nó vẫn gặp phải một số nhược điểm:

- Thuật toán Girvan-Newman sử dụng phương pháp loại trừ dần đến khi không có cạnh nào vượt qua ngưỡng của độ đo trung tâm trung gian cao, vì vậy nên số lượng cộng đồng không kiểm soát trước được. Bên cạnh đó, thuật toán sử dụng nhiều phép phân vùng, khó có thể xác định được phép phân vùng nào mang lại hiệu quả tốt nhất.
- Do tại mỗi lượt thực hiện, thuật toán tính lại độ đo trung tâm trung gian của mỗi cạnh liên quan sau khi xóa đi cạnh có độ đo trung tâm trung gian lớn nhất nên độ phức tạp thời gian tính toán cao. Giả sử với đồ thị n đỉnh, số cạnh phải xóa đi khỏi đồ thị là m cạnh thì ta cần lượng thời gian tính toán $O(mn)$ cho mỗi lần lặp. Tổng thời gian chạy của thuật toán là $O(m^2n)$.

Trong thời gian qua có khá nhiều thuật toán cải tiến, phát triển thuật toán GN được đề xuất [6], [20], [22], [38], [40], [41], ... Dựa trên những ưu điểm và

nhược điểm trên của Girvan-Newman, các nhà khoa học đã đưa ra nhiều cách để cải tiến thuật toán nhằm khắc phục những nhược điểm đó.

Gần đây, thuật toán mới nhất phát hiện cộng đồng thực hiện cải tiến từ thuật toán Girvan-Newman dựa trên độ đo trung tâm trung gian của cạnh là của tác giả Majid Arasteh và các cộng sự (năm 2018) [6] gọi tắt là thuật toán MAA. Thuật toán MAA có một số tính năng quan trọng như sau:

- Thuật toán dựa trên phương pháp tham lam, phân cụm và phân cấp.
- Thuật toán không phụ thuộc vào việc yêu cầu các thông tin xác định trước số lượng các cộng đồng.
- Các thành phần của các cộng đồng được phát hiện không thể thuộc về nhiều hơn một cộng đồng, nghĩa là không có cộng đồng chồng chéo nhau.
- Đề xuất một độ đo mới thực hiện tính toán tỷ lệ độ đo trung tâm của cạnh.
- Nhiều cạnh được loại bỏ sau mỗi lần lặp. Một số cạnh có độ đo trung tâm cao nhất sẽ bị xóa bỏ ở cùng một lần lặp.

Thuật toán MAA có độ phức tạp tính toán là O(m^2) đối với cả đồ thị có trọng số và không có trọng số. Trong đó m là tổng số cạnh của đồ thị mạng xã hội. Thuật toán MAA được sử dụng để so sánh trong chương 2 của luận án.

Dựa trên ý tưởng của phương pháp phát hiện cộng đồng dựa vào độ đo trung tâm trung gian, nghiên cứu sinh nhận thấy trên đồ thị mạng xã hội có khá nhiều đỉnh tương đương với nhau theo cấu trúc có cùng độ đo trung tâm trung gian, chúng tạo thành các lớp tương đương và có thể kết hợp chúng lại với nhau thành một đỉnh đại diện duy nhất cho cả lớp đỉnh. Do vậy giảm thiểu được đáng kể số đỉnh và cạnh của đồ thị mạng xã hội ban đầu, giảm thiểu được chi phí tính toán mà lại không ảnh hưởng đến cấu trúc của đồ thị mạng xã hội ban đầu. Vì vậy trong chương 2 của luận án nghiên cứu sinh đề xuất thuật toán rút gọn đồ thị mạng xã hội dựa vào độ đo trung tâm trung gian nhằm cải tiến thời gian tính toán độ đo trung tâm trung gian và áp dụng để phát hiện nhanh và hiệu quả các cộng đồng trên mạng xã hội.

1.3.2.4. Nhóm thuật toán phát hiện cộng đồng dựa trên lan truyền nhãn

Thuật toán điển hình và phổ biến tiếp theo trong lĩnh vực phát hiện cộng đồng mạng xã hội là thuật toán lan truyền nhãn LPA (Label Propagation Algorithm) được giới thiệu bởi Raghavan và các cộng sự [85]. Lan truyền nhãn là thuật toán phát hiện các cộng đồng trên mạng xã hội trong thời gian gần tuyến tính. Thuật toán được nhiều nhà khoa học quan tâm nghiên cứu, phát triển và áp dụng cho nhiều các trường hợp khác nhau.

Thuật toán LPA thực hiện theo các bước: Bước đầu tiên là gán cho mỗi đỉnh một nhãn duy nhất. Nhãn sẽ đại diện cho cộng đồng của mỗi đỉnh. Tiếp theo, các đỉnh được sắp xếp theo thứ tự ngẫu nhiên và cập nhật cộng đồng theo trọng số của các đỉnh lân cận của chúng. Quá trình này được lặp lại cho đến khi không có thêm đỉnh nào được cập nhật theo cộng đồng.

Sau khi Newman giới thiệu tính mô đun để đo lường chất lượng phân chia mạng, Barber và Clark đã đề xuất thuật toán LPAm [13] để kiểm soát quá trình lan truyền nhãn, Liu và các cộng sự giới thiệu thuật toán cải tiến lan truyền nhãn LPAm+ [65]. Tiếp theo, Zhang và các cộng sự tiếp tục đề xuất thuật toán LPAp [116].

Tuy nhiên, nguyên lý lan truyền nhãn vẫn tồn tại hai vấn đề. Đầu tiên, nhãn được lan truyền sẽ không hội tụ trong một số mạng (ví dụ đồ thị mạng bi-partite) và thứ hai, thứ tự ngẫu nhiên các đỉnh được xem xét để lan truyền nhãn có thể ảnh hưởng lớn đến việc phát hiện các cộng đồng.

Gần đây, thuật toán cải tiến LPA của Matin Pirouz và cộng sự [82] đề xuất năm 2018 được gọi là OLP. Thuật toán OLP gồm 2 bước: Trong bước đầu tiên, một nhãn ngẫu nhiên sẽ được gán cho những đỉnh trong đồ thị. Bằng cách này, mỗi đỉnh được đặt trong cộng đồng riêng của nó. Tiếp theo, mỗi đỉnh trong mạng có cơ hội tham gia một trong những cộng đồng lân cận của nó. Thuật toán lựa chọn nhãn dựa trên lân cận nhãn của đỉnh mà nó kế thừa. Với hệ thống lan truyền này, các đỉnh có bậc cao hơn sẽ được ưu tiên cao hơn. Do đó, các nhãn thuộc về các đỉnh phổ biến hơn sẽ trở thành phổ biến hơn và cộng đồng của mạng sẽ được tạo ra. Trong bước thứ hai, kiểm tra sự hội tụ của thuật toán. Trong bước này, mỗi đỉnh sẽ kiểm tra xung quanh các đỉnh lân cận của chúng để kiểm tra xem nhãn của chúng đã khớp với nhãn của phần

lớn đỉnh lân cận chưa. Nếu nhãn chưa khớp với phần lớn các đỉnh lân cận, thì nhãn của đỉnh này sẽ tiếp tục được cập nhật cho đến khi thỏa mãn yêu cầu và hội tụ. Đến cuối mỗi lần lặp, tất cả các đỉnh giữ cùng một nhãn được tính là một cộng đồng. Độ phức tạp của thuật toán OLP là O(n) với n là số đỉnh của đồ thị. Thuật toán OLP được sử dụng để so sánh trong Chương 3 của luận án.

Tuy nhiên, các nghiên cứu nêu trên hầu hết chỉ giải quyết bài toán tìm cộng đồng trực tiếp trên đồ thị mà rất ít có công trình nghiên cứu tính đến việc giảm thiểu không gian đỉnh và cạnh của đồ thị để giảm thiểu thời gian phân tích, phát hiện các cộng đồng mạng xã hội. Trên đồ thị mạng xã hội có khá nhiều đỉnh có nhãn giống với nhãn (trong cùng một cấu trúc cộng đồng) của một trong số các đỉnh lân cận, và nhãn của chúng luôn được cập nhật lại theo những đỉnh đó suốt trong quá trình lan truyền nhãn. Những đỉnh này tương đương với nhau theo cấu trúc, luôn có cùng nhãn trong các bước lan truyền nhãn, sẽ tạo thành các lớp tương đương và do vậy, có thể kết hợp chúng với nhau thành một đỉnh đại diện duy nhất cho cả lớp đỉnh nhằm giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội ban đầu mà không ảnh hưởng đến cấu trúc của đồ thị mạng xã hội ban đầu. Vì vậy, Chương 2 luận án đề xuất phát triển thuật toán rút gọn đồ thị mạng xã hội dựa vào nguyên lý lan truyền nhãn và áp dụng để phát triển thuật toán phát hiện nhanh, hiệu quả các cộng đồng mạng xã hội.

Tổng hợp một số thuật toán phổ biến phát hiện cộng đồng mạng xã hội được trình bày trong Bảng 1.1 sau đây.

Bảng 1.1. Một số thuật toán phổ biến phát hiện cộng đồng mạng xã hội

Stt	**Nhóm thuật toán**	**Tên thuật toán**	**Thuật toán minh họa điển hình**
1	Nhóm thuật toán phát hiện cộng đồng truyền thống	Thuật toán phân cụm đồ thị	Kernighan - Lin
		Thuật toán phân cụm phân cấp	• BIRCH • EAGLE
		Thuật toán phân cụm phân hoạch	• K-means • K-means mờ
		Thuật toán phân cụm theo phổ	• Phân cụm theo phổ không chuẩn • Phân cụm theo phổ chuẩn hoá

2	Nhóm thuật toán phát hiện cộng đồng dựa trên tối ưu hoá độ đo đơn thể	Thuật toán tìm kiếm tham lam	• Louvain • Louvain-LPA
		Thuật toán mô phỏng luyện kim	Mô phỏng luyện kim
		Thuật toán tối ưu hóa mở rộng	Tối ưu hóa mở rộng
		Các thuật toán tiến hóa	• Tối ưu hóa đơn mục tiêu • Tối ưu hóa đa mục tiêu
3	Nhóm thuật toán phát hiện cộng đồng dựa vào độ đo trung tâm trung gian	Họ thuật toán Girvan-Newman	• Girvan-Newman • MAA
4	Nhóm thuật toán phát hiện cộng đồng dựa trên lan truyền nhãn	Họ thuật toán lan truyền nhãn	• LPA • LPAm • LPAm+ • LPAp • OLP

1.4. Bài toán rút gọn đồ thị

Bài toán rút gọn đồ thị nhằm giảm thiểu không gian, thời gian tính toán của những đồ thị lớn, phức tạp là một hướng nghiên cứu quan trọng được nhiều người quan tâm nghiên cứu và được ứng dụng trong nhiều lĩnh vực khác nhau như trong hệ thống quản lý luồng công việc, xử lý ảnh, mạng ngữ nghĩa, xử lý ngôn ngữ tự nhiên, phát hiện mẫu, phân tích mạng xã hội [7], [58], [61], [90], [100], [103].

1.4.1. Sự cần thiết phải rút gọn đồ thị mạng xã hội

Rút gọn đồ thị mạng xã hội là bài toán quan trọng trong lĩnh vực phân tích dữ liệu. Mục tiêu của bài toán rút gọn đồ thị mạng xã hội là giảm thiểu chi phí, thời gian tính toán mà không làm giảm chất lượng giải pháp hoặc sửa đổi cấu trúc của đồ thị mạng xã hội ban đầu. Rút gọn đồ thị cũng là một giải pháp hữu hiệu để tăng tốc các thuật toán thực thi trên đồ thị đồng thời giảm kích thước của dữ liệu.

Do tính chất của mạng xã hội có cấu trúc khá tự do và kích thước rất lớn không ngừng phát triển theo thời gian, vì vậy các thuật toán phát hiện cộng đồng mất rất nhiều thời gian tính toán và chưa hiệu quả. Một trong những cách tiếp cận để khắc

phục hạn chế trên là phương pháp rút gọn đồ thị mạng xã hội để giảm thiểu thời gian tính toán. Tuy nhiên, việc rút gọn đồ thị mạng xã hội và vẫn bảo toàn được các tính chất của cộng đồng là một thách thức lớn và còn tùy thuộc vào cách tiếp cận của phương pháp phát hiện cộng đồng trên mạng xã hội.

1.4.2. Các thuật toán rút gọn đồ thị

Rút gọn đồ thị thường được được nghiên cứu và ứng dụng trong các hệ thống quản lý luồng công việc (workflow management system), thị giác máy tính (computer vision), đồ thị ngữ nghĩa (semantic graphs), phát hiện mẫu trong các đồ thị lớn (sampling from large graphs), và các ứng dụng khác.

1.4.2.1. *Thuật toán rút gọn đồ thị trong hệ thống quản lý luồng công việc*

Sadiq và Orlowska trong [89] đã trình bày một thuật toán rút gọn đồ thị và áp dụng để để phát hiện sự tồn tại của các xung đột cấu trúc (structural conflicts) trong đồ thị quy trình làm việc. Ý tưởng cơ bản của phương pháp này là loại bỏ một số đỉnh và hay hoặc cạnh khỏi đồ thị luồng công việc và lặp lại để rút gọn đồ thị luồng công việc. Thuật toán rút gọn một đồ thị luồng công việc không có xung đột cấu trúc sẽ rút gọn thành một đồ thị trống. Ngược lại, nếu luồng công việc có chứa tắc nghẽn (deadlock) hoặc không đồng bộ hóa, xung đột cấu trúc có thể xác định tường minh trên đồ thị rút gọn. Lin và các cộng sự [63] đưa ra một tập các quy tắc rút gọn đồ thị luồng công việc và xác định các xung đột, các tắc nghẽn và thiếu đồng bộ trong một quá trình nghiệp vụ (business process). Thuật toán rút gọn đồ thị có thể loại bỏ tất cả các đỉnh tắc nghẽn và xung đột cấu trúc khỏi đồ thị quy trình nghiệp vụ. Thuật toán rút gọn đồ thị quy trình nghiệp vụ theo chu kỳ (cyclic workflow graph) thành đồ thị phi chu trình, sau đó kiểm chứng đồ thị quy trình nghiệp vụ phi chu trình kết quả để khẳng định tính hợp lý, hiệu quả của quy trình nghiệp vụ. Các thuật toán này được ứng dụng hiệu quả trong việc xác minh quy trình quy trình nghiệp vụ dựa trên đồ thị luồng công việc thực tế.

1.4.2.2. *Thuật toán rút gọn đồ thị trong thị giác máy tính*

Thị giác máy tính bao gồm các công việc chính như phân đoạn (segmentation), phục hồi hình ảnh (image restoration), ước lượng âm thanh nổi (stereo) và đối sánh

hình dạng, ... Thuật toán GraphCuts phân đoạn ảnh bằng đồ thị Slim có thể được sử dụng để giải quyết hiệu quả các vấn đề ghi nhãn trên hình ảnh có độ phân giải cao hoặc các hệ thống hạn chế tài nguyên [93]. Đồ thị Slim được xây dựng bằng cách hợp nhất các đỉnh được kết nối với nhau bằng các cạnh đơn. Các tác giả đã chứng được rằng giá trị của luồng cực đại trên đồ thị Slim bằng với luồng cực đại của đồ thị gốc. Các đỉnh được kết nối bởi một cạnh đơn sẽ có cùng nhãn trong phân đoạn cuối cùng và có thể được hợp nhất thành một đỉnh duy nhất. Phương pháp đề xuất yêu cầu ít bộ nhớ hơn nhiều cho bài toán phân đoạn ảnh truyền thống và có kích thước hợp lý ngay cả trên thiết bị di động. Việc giảm thời gian tính toán cũng có thể đạt được bằng cách sử dụng kiến trúc phần cứng xử lý song song.

1.4.2.3. Thuật toán rút gọn đồ thị trong mạng ngữ nghĩa

Sự sẵn có ngày càng tăng của thông tin trực tuyến đã đòi hỏi các nhà nghiên cứu tập trung chuyên sâu bài toán tóm tắt văn bản (text summarization) tự động trong lĩnh vực xử lý ngôn ngữ tự nhiên. Tóm tắt văn bản tự động có thể được thực hiện hiệu quả thông qua việc sử dụng kỹ thuật rút gọn mạng ngữ nghĩa được biểu diễn dưới dạng đồ thị ngữ nghĩa (semantic graph). Trong việc xây dựng một đồ thị ngữ nghĩa rút gọn, một bước quan trọng là nhóm các khái niệm tương tự có cùng một nhãn và kết nối chúng với các kho lưu trữ bên ngoài. Một phương pháp mới [111] được đề xuất để giảm đồ thị có hướng lớn thành đồ thị đơn giản hơn với ít đỉnh hơn. Việc rút gọn được thực hiện thông qua tập hợp đỉnh và cạnh dựa trên nguyên tắc entropy cực đại. Phương pháp này áp dụng cho bài toán rút gọn mô hình chuỗi Markov (Markov chain model-reduction problem) [111], cung cấp một phương pháp phân cụm mềm cho phép tổng hợp các ma trận chuyển đổi trạng thái tốt hơn các phương pháp hiện có. Rút gọn đồ thị được sử dụng để hiệu quả trong lập chỉ mục và các hệ thống tìm kiếm (indexing and retrieval).

1.4.2.4. Thuật toán rút gọn đồ thị trong phát hiện mẫu

Bài toán phát hiện (lấy) mẫu bao gồm phát hiện mẫu đỉnh (vertex sampling), phát hiện mẫu cạnh (edge sampling) và phát hiện mẫu dựa trên truyền tải (traversal - based sampling).

Kỹ thuật phát hiện mẫu đỉnh lựa chọn p đỉnh đầu tiên của đồ thị và giữ lại các cạnh (liên kết) giữa chúng. Kỹ thuật phát hiện mẫu đỉnh điển hình là lấy mẫu đỉnh ngẫu nhiên RN (Random node sampling) và lấy mẫu theo bậc của đỉnh ngẫu nhiên RDN (Random degree node) [100]. RDN lựa chọn một tập các đỉnh đồng nhất ngẫu nhiên. Sử dụng tập hợp đỉnh tạo ra một đồ thị con bao gồm các cạnh là các kết nối giữa các đỉnh đồng nhất. Stumpf và các cộng sự [100] đã chứng minh rằng nhược điểm của RN là không giữ nguyên được quy luật phân bố của mạng cũng như không bảo toàn tính chất của đồ thị ban đầu. Thay vì lấy mẫu đỉnh đồng nhất, các phương pháp cải tiến lấy mẫu đỉnh lựa chọn những đỉnh có xác suất khác nhau dựa trên các tính chất của đồ thị ban đầu. Trong RDN, các đỉnh được lựa chọn để rút gọn tỷ lệ thuận với bậc của đỉnh. Tuy nhiên, đồ thị rút gọn có số lượng đỉnh có bậc của đỉnh cao hơn và phân bố đồ thị thu được trở nên dày đặc hơn.

Phát hiện mẫu cạnh ngẫu nhiên RE (Random edge sampling) [61] gặp hạn chế là các đồ thị được lấy mẫu có rất ít kết nối với nhau do đó sẽ ưu tiên những đồ thị con có kích thước lớn hay những cộng đồng lớn mà bỏ qua những đồ thị con có kích thước nhỏ hay những cộng đồng nhỏ. Một biến thể của RE là RNE (Random node - edge sampling), đầu tiên chọn ngẫu nhiên một đỉnh đồng nhất và ngẫu nhiên một cạnh đồng nhất nối với đỉnh đó. Theo các công trình nghiên cứu trước đó, Leskovec và Faloutsos [61] đã chứng minh thì cả RE và RNE đều không bảo toàn được các cộng đồng vì các đồ thị con được lấy mẫu được kết nối với nhau rất ít trong khi đó cả RE và RNE đều ưu tiên lựa chọn các đỉnh có bậc của đỉnh cao do xác suất chọn một đỉnh tăng theo bậc của nó.

Kỹ thuật phát hiện mẫu bắt đầu từ một tập hợp đỉnh ban đầu sau đó mở rộng mẫu theo các quan sát hiện tại. Điển hình là chọn mẫu lan truyền hòn tuyết lăn SES (Snowball Expansion Sampling), thuật toán FFS (Forest Fire Sampling). Trong kỹ thuật lấy mẫu SES, ta bắt đầu với một tập đỉnh ban đầu được lựa chọn ngẫu nhiên. SES giữ lại các mẫu có đặc tính mở rộng tốt, đây là mẫu đại diện cho cấu trúc của đồ thị ban đầu. SES chọn một tỉ lệ cố định các láng giềng được truy cập ở mỗi lần lặp [58]. FFS là một phiên bản xác suất của SES. Trong SES ta lựa chọn cố định các láng

giềng ở mỗi bước thì trong FFS các láng giềng được chọn theo xác suất. Thuật toán lựa chọn một đỉnh ngẫu nhiên làm trung tâm, sau đó các cạnh và các đỉnh tương ứng được ghi đồng thời. Nếu một cạnh được ghi, đỉnh ở đầu kia sẽ ghi nhận các cạnh nối với nó và cứ tiếp tục như vậy. Mô hình này có hai tham số tương ứng với xác suất ghi. Đồ thị lấy mẫu thu được khi đạt đến số đỉnh mong muốn hoặc không thể ghi thêm đỉnh nào nữa. Chất lượng lấy mẫu của phương pháp này phù hợp với các cấu trúc khác nhau của đồ thị gốc như: độ phân bố, hệ số phân cụm và kích thước các thành phần. Phương pháp này được đánh giá tốt hơn RE và RN. Ngoài ra, FFS đáp ứng được 15% các thuộc tính của đồ thị ban đầu. Độ phức tạp của thuật toán không còn là tuyến tính.

Các cách tiếp cận rút gọn đồ thị phần lớn phụ thuộc vào các đặc tính cơ bản của lĩnh vực ứng dụng. Hầu như không có phương pháp rút gọn đồ thị nào nêu trên bảo toàn được cấu trúc thông tin về cộng đồng trên mạng xã hội. Luận án đã đề xuất hai phương pháp rút gọn đồ thị mạng xã hội (chương 2) và áp dụng phát triển hai thuật toán nhanh, hiệu quả phát hiện các cộng đồng trên đồ thị rút gọn mà vẫn bảo toàn được tính chất của các cộng đồng mạng xã hội ban đầu (chương 3).

1.5. Độ đo đánh giá thuật toán phát hiện cộng đồng mạng xã hội

Mục tiêu của rút gọn đồ thị mạng xã hội là áp dụng để cải tiến thuật toán phát hiện cộng đồng trên mạng xã hội. Vì vậy, cần đánh giá tính hiệu quả của thuật toán phát hiện cộng đồng thông qua các độ đo. Độ đo (measures) [71] được dùng để đánh giá hiệu quả, chất lượng của thuật toán phát hiện cộng đồng so với các cộng đồng có trong thực tế. Độ đo cũng được dùng để đo sự tương đồng hay giống nhau giữa kết quả phát hiện cộng đồng của các thuật toán so với các cộng đồng có trong thực tế. Mục này trình bày các độ đo này nhằm đánh giá tính hiệu quả của các thuật toán rút gọn đồ thị được đề xuất trong chương 2 và các thuật toán phát hiện cộng đồng trên mạng xã hội được đề xuất trong chương 3 của luận án.

1.5.1. Độ đo đơn thể mô đun Q

Độ đo đơn thể mô đun Q được đề xuất bởi Girvan - Newman [22], [78] được sử dụng để đo lường mức độ phân chia cộng đồng của toàn mạng. Mạng có độ đo đơn

thể mô đun Q cao cho thấy có rất nhiều các liên kết giữa các đỉnh trong cùng một cộng đồng và có ít liên kết giữa các đỉnh trong các cộng đồng khác nhau. Vì vậy, độ đo đơn thể mô đun Q có giá trị càng cao thì đồng nghĩa kết quả phân vùng trong mạng càng tốt.

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{d_i d_j}{2m})\delta(C_i, C_j) \quad (1.12)$$

Trong đó:

A_{ij} là ma trận liền kề, m là tổng số cạnh của đồ thị, d_i là bậc của đỉnh i, d_j là bậc của đỉnh j, C_i, C_j lần lượt là các cộng đồng mà đỉnh i và đỉnh j ở trong cộng đồng đó.

$\delta(C_i, C_j)$ = 1 nếu đỉnh i và đỉnh j cùng thuộc một cộng đồng ngược lại $\delta(C_i, C_j)$ = 0.

Độ đo đơn thể mô đun Q được sử dụng rộng rãi vì đơn giản để tính toán và có thể đánh giá chính xác chất lượng của các phân vùng trong mạng [64], [112], [114].

1.5.2. Độ đo F-measure

Độ đo F-measure là độ đo dựa trên độ tương tự cặp [41], [112], [114]. Độ đo này được sử dụng từ lâu trong công việc phân cụm dữ liệu, xử lý ngôn ngữ tự nhiên, truy xuất thông tin và học máy. Độ đo F-measure giữa tập các cộng đồng thực C và tập các cộng đồng được phát hiện C' được tính bằng công thức như sau:

$$\text{F-measure (C,C')} = \frac{2 * Precision\,(C,C') * Recall\,(C,C')}{Precision\,(C,C') + Recall(C,C')} \quad (1.13)$$

Trong đó: $Precision\ (C,C') = \frac{|C \cap C'|}{|C'|}$ (1.14)

$$Recall\ (C,C') = \frac{|C \cap C'|}{|C|} \quad (1.15)$$

Trong đó C là tập các cộng đồng thực (ground-truth, known communities) và C' là tập các cộng đồng được phát hiện. Độ đo F-measure để đánh giá chất lượng của các cộng đồng được phát hiện vì có hai yếu tố: độ phức tạp tính toán là tuyến tính và dễ thực hiện cũng như giải thích giữa các thước đo bên ngoài.

Để thực hiện tính giá trị độ đo F-measure trên đồ thị mạng xã hội. Bước đầu tiên thực hiện tính toán độ đo F-measure từng cặp giữa tất cả các cộng đồng được phát hiện bởi thuật toán phát hiện cộng đồng và tất cả các cộng đồng thực trong mạng. Trong trường hợp thực tế mạng xã hội có n cộng đồng thực (ground-truth), mà thuật

toán phát hiện cộng đồng lại tính ra m cộng đồng thì ta tiến hành tính độ đo F-measure lần lượt từng cặp giữa tất cả các cộng đồng thực (ground-truth) với tất cả các cộng đồng được thuật toán phát hiện. Số lần tính độ đo F-measure được thực hiện là n*m lần. Sau đó ta chọn n kết quả F-measure cao nhất ở trên và tính giá trị trung bình. Cuối cùng, lấy giá trị trung bình của độ đo F-measure [64], [88], [113].

1.5.3. Độ đo NMI dựa trên lý thuyết thông tin

Các độ đo dựa trên lý thuyết thông tin đưa ra một cách tiếp cận khác để kiểm chứng chất lượng cộng đồng với phân vùng tham chiếu nhất định. Độ đo dựa trên lý thuyết thông tin thường được sử dụng là độ đo thông tin tương hỗ chuẩn NMI (Normal Mutual Information) [96]. Độ đo NMI định nghĩa N là ma trận sai số với các hàng tương ứng với cộng đồng trong thực tế, các cột tương ứng với cộng đồng được phát hiện, các hàng biểu diễn tập A các cộng đồng thực và các cột tương ứng tập B các cộng đồng được phát hiện. Các phần tử bên trong ma trận N mô tả độ tương tự giữa các nhóm. N_{ij} (hàng thứ i, cột thứ j) là số các đỉnh của cộng đồng thực i xuất hiện trong cộng đồng được phát hiện j. Độ đo NMI giữa hai mạng A và mạng B được tính theo công thức:

$$\text{NMI(A,B)} = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i_s}N_{j_s}}\right)}{\sum_{i=1}^{C_A} N_{i_s}\log\left(\frac{N_{i_s}}{N}\right)+\sum_{j=1}^{C_B} N_{j_s}\log\left(\frac{N_{j_s}}{N}\right)} \quad (1.16)$$

C_A, C_B tương ứng là số cộng đồng thực (ground-truth) và số cộng đồng được phát hiện bởi thuật toán phát hiện cộng đồng trong mạng, N_{i_s} và N_{j_s} đại diện cho tổng số phần tử trên hàng i và cột j tương ứng của ma trận N. Giá trị của NMI nằm trong khoảng từ 0 đến 1. Chỉ số NMI bằng 1 nếu cộng đồng tìm kiếm trùng với cộng đồng thực. Ngược lại NMI bằng 0.

Luận án sử dụng các độ đo: Độ đo đơn thể mô đun Q, đo đo F-measure và độ đo NMI để đánh giá tính hiệu quả của thuật toán phát hiện cộng đồng mạng xã hội vì đây không chỉ là các độ đo được đánh giá là phổ biến, thông dụng, hữu hiệu được sử dụng thường xuyên, liên tục để đánh giá hiệu quả, chất lượng phát hiện cộng đồng mạng xã hội [64], [88], [112], [113], [114]. Ngoài ra một yếu tố quan trọng khác nữa

để đảm bảo yếu tố tin cậy, khách quan của các thuật toán được luận án so sánh, tham chiếu đã công bố các kết quả nghiên cứu liên quan đến các loại độ đo này.

1.6. Độ đo đánh giá thuật toán rút gọn đồ thị

Tỷ lệ rút gọn đồ thị của thuật toán cho thấy hiệu quả đạt được của thuật toán rút gọn đồ thị mạng xã hội đề xuất. Nếu |E| và |Ec| là số cạnh của đồ thị ban đầu và đồ thị sau khi rút gọn, khi đó tỷ lệ rút gọn (Compression) [51] được tính theo công thức:

$$\text{Compression (VN)} = \frac{|E|-|Ec|}{|E|} \quad (1.17)$$

Giá trị của chỉ số này nằm trong khoảng từ 0 đến 1. Giá trị của chỉ số càng cao thì đồng nghĩa hiệu suất rút gọn đồ thị đạt được càng tốt.

1.7. Kết luận chương 1

Chương 1 trình bày một số khái niệm cơ sở về phân tích mạng xã hội và các phương pháp phát hiện cộng đồng mạng xã hội. Phân tích mạng xã hội là một tập hợp các phương pháp phân tích các khái niệm, sử dụng lý thuyết đồ thị để mô tả và phân tích các mối quan hệ giữa các tác nhân (thực thể) trong mạng, xác nhận các quy luật hình thành và biến chuyển của những mối quan hệ đó, và làm sáng tỏ những ảnh hưởng của các mối quan hệ xã hội (hay cấu trúc của mạng) đối với hành vi của các tác nhân. Để xác định được vai trò và mối quan hệ của các tác nhân người ta sử dụng các độ đo trung tâm trung gian, nhất là độ đo trung tâm trung gian của các đỉnh, cạnh trên đồ thị mạng xã hội.

Bài toán phát hiện cộng đồng trên mạng xã hội là một nội dung chính của phân tích mạng xã hội được rất nhiều sự quan tâm, nghiên cứu của các nhà khoa học trong nước và trên thế giới. Chương này đã giới thiệu 4 nhóm thuật toán chính phát hiện các cộng đồng trên mạng xã hội: các thuật toán phân cụm truyền thống, các thuật toán dựa vào độ đo đơn thể hóa, các thuật toán dựa vào độ đo trung tâm trung gian và các thuật toán dựa vào nguyên lý lan truyền nhãn.

Do tính chất của mạng xã hội có cấu trúc khá tự do và kích thước rất lớn không ngừng phát triển theo thời gian, vì vậy bài toán phân tích mạng xã hội, phát hiện cộng đồng mất rất nhiều thời gian và chưa hiệu quả. Một trong những cách tiếp cận để

khắc phục nhược điểm trên là đề xuất phương pháp rút gọn đồ thị để giảm thiểu thời gian tính toán là hết sức cần thiết. Chương này cũng phân tích các phương pháp rút gọn đồ thị và ứng dụng trong nhiều lĩnh vực khác nhau. Tuy nhiên, các phương pháp rút gọn đồ thị truyền thống không bảo toàn được các thông tin về cộng đồng của đồ thị mạng xã hội ban đầu. Các chương sau đề xuất phương pháp rút gọn đồ thị mạng xã hội dựa vào độ đo trung tâm trung gian và nguyên lý lan truyền nhãn, từ đó áp dụng để phát triển các thuật toán phát hiện nhanh, hiệu quả các cộng đồng trên mạng xã hội.

CHƯƠNG 2. THUẬT TOÁN RÚT GỌN ĐỒ THỊ MẠNG XÃ HỘI DỰA VÀO ĐỘ ĐO TRUNG TÂM TRUNG GIAN VÀ NGUYÊN LÝ LAN TRUYỀN NHÃN

2.1. Giới thiệu

Hầu hết các phương pháp phát hiện cộng đồng trên mạng xã hội đều tập trung vào việc nghiên cứu các mối liên kết giữa các thực thể để xác định các cộng đồng. Mạng xã hội rất phong phú, đa dạng, có thành phần tham gia rất lớn và có thể phát triển, mở rộng theo thời gian. Vì vậy các thuật toán phát hiện cộng đồng trên đồ thị mạng xã hội đều mất khá nhiều thời gian tính toán và kém hiệu quả. Một trong các hướng nghiên cứu để giảm độ phức tạp tính toán là hướng rút gọn đồ thị. Nhược điểm chung của hầu hết các phương pháp rút gọn trong lý thuyết đồ thị truyền thống là không bảo toàn được các thuộc tính cấu trúc của đồ thị ban đầu, không bảo toàn được chất lượng cộng đồng và thường có những yêu cầu về các thông tin dự đoán ban đầu. Trong chương này, luận án tập trung nghiên cứu các tính chất của các đỉnh tương đương dựa vào độ đo trung tâm trung gian và nguyên lý lan truyền nhãn từ đó đề xuất thuật toán kết hợp các lớp đỉnh tương đương theo độ đo trung tâm trung gian và nguyên lý lan truyền nhãn để rút gọn đồ thị nhưng vẫn bảo toàn chất lượng cộng đồng và áp dụng rút gọn đồ thị để phát triển thuật toán phát hiện cộng đồng trên đồ thị mạng xã hội dựa vào độ đo trung tâm trung gian và nguyên lý lan truyền nhãn. Các kết quả trong chương này được công bố trong các công trình [CT1], [CT3], [CT4].

Dựa trên ý tưởng của phương pháp phát hiện cộng đồng dựa vào độ đo trung tâm trung gian, nghiên cứu sinh nhận thấy trên đồ thị mạng xã hội có nhiều đỉnh tương đương với nhau theo cấu trúc có cùng độ đo trung tâm trung gian, chúng tạo thành các lớp tương đương và có thể kết hợp chúng lại với nhau thành một đỉnh đại diện duy nhất cho cả lớp đỉnh. Do vậy giảm thiểu được đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội ban đầu, giảm thiểu được chi phí tính toán mà lại không ảnh hưởng đến cấu trúc của đồ thị mạng xã hội ban đầu.

Phát hiện cộng đồng trên mạng xã hội là một nhiệm vụ quan trọng hàng đầu trong phân tích mạng xã hội. Để giải quyết bài toán này, nhiều thuật toán phát hiện cộng đồng trên mạng xã hội đã được đề xuất. Tuy nhiên, các thuật toán này hầu hết chưa đạt hiệu quả trong việc phát hiện cộng đồng trên các mạng xã hội có kích thước rất lớn. Với sự phát triển mạnh mẽ của công nghệ thông tin, việc sử dụng mạng xã hội của chúng ta đang phát triển theo cấp số nhân. Hệ quả là quy mô của mạng xã hội càng phát triển và trở nên khổng lồ. Điều này dẫn đến việc phát hiện cộng đồng trên các mạng xã hội quy mô rất lớn không thể được giải quyết bằng các thuật toán truyền thống do độ phức tạp về thời gian và không gian tính toán. Có nghĩa là, hầu hết các thuật toán phát hiện cộng đồng mạng xã hội hiện không thể được mở rộng đến kích thước khổng lồ của các mạng xã hội. Để giải quyết thách thức lớn này, nghiên cứu sinh đề xuất các phương pháp rút gọn đồ thị mạng xã hội nhằm giảm thiểu kích thước của mạng xã hội để phát hiện các cộng đồng trên mạng xã hội nhanh, hiệu quả tuy nhiên vẫn bảo toàn được các tính chất của cộng đồng mạng xã hội ban đầu.

Bài toán rút gọn đồ thị đã được trình bày tại mục 1.4 của chương 1 trong luận án, chương 2 bao gồm các mục trình bày về các định nghĩa, đề xuất các tính chất, hệ quả của độ đo trung tâm trung gian từ đó đề xuất thuật toán rút gọn đồ thị mạng xã hội dựa vào các lớp đỉnh tương đương theo độ đo trung tâm trung gian. Tiếp theo, chương 2 trình bày nguyên lý lan truyền nhãn để từ đó đề xuất thuật toán rút gọn đồ thị mạng xã hội dựa vào các lớp đỉnh tương đương theo nguyên lý lan truyền nhãn. Các thuật toán đề xuất để rút gọn đồ thị mạng xã hội trong chương 2 nhằm mục đích áp dụng hiệu quả cho bài toán phát hiện cộng đồng trên mạng xã hội rất lớn tuy nhiên vẫn bảo toàn được các tính chất của cộng đồng mạng xã hội ban đầu.

2.2. Các tính chất của độ đo trung tâm trung gian trên đồ thị mạng xã hội

Độ đo trung tâm trung gian đã được giới thiệu ở Chương 1, phần này nghiên cứu một số các tính chất tương đương theo độ đo trung tâm trung gian của các đỉnh trên đồ thị. Từ đó, thuật toán kết hợp các lớp đỉnh tương đương theo độ đo trung tâm trung gian trên đồ thị để thực hiện rút gọn đồ thị mạng xã hội được đề xuất.

Giả thiết mạng xã hội được biểu diễn bởi một đồ thị đơn liên thông G = (V, E), trong đó V là tập các đỉnh, E là tập các cạnh. Ký hiệu σ_{st} là số đường đi ngắn nhất đi từ s tới t, và $\sigma_{st}(v)$ là số đường đi ngắn nhất đi từ s tới t và có đi qua v. Khi đó độ đo trung tâm trung gian của đỉnh v, ký hiệu là $C_B(v)$ [84] được tính như sau:

$$C_B(v) = \sum_{s \neq t \neq v} \sigma_{st}(v) / \sigma_{st} \quad (2.1)$$

Độ đo trung tâm trung gian của cạnh e, ký hiệu là $C_B(e)$ [84], được định nghĩa như sau:

$$C_B(e) = \sum_{s \neq t} \sigma_{st}(e) / \sigma_{st} \quad (2.2)$$

Với hai đỉnh s, t ∈ V, cạnh e ∈ E và $\delta_{st}(e)$ là số đường đi ngắn nhất đi từ đỉnh s tới đỉnh t và đi qua cạnh e.

Độ đo trung tâm trung gian của đỉnh v cũng có thể tính thông qua công thức tính độ đo trung tâm trung gian của cạnh e.

$$C_B(v) = \frac{1}{2} \sum_{e \in \Gamma(v)} C_B(e) - (n - 1) \quad (2.3)$$

Trong đó, Γ(v) là tập các cạnh kề với v và n là số đỉnh của thành phần chứa v.

Trên đồ thị mạng xã hội có nhiều đỉnh tương đương với nhau theo cấu trúc dựa vào độ đo trung tâm trung gian, chúng tạo thành các lớp tương đương và có thể kết hợp chúng với nhau thành một đỉnh đại diện cho cả lớp có cùng độ đo trung tâm trung gian, nhằm giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội.

2.2.1. Các lớp đỉnh treo tương đương

Mục này giới thiệu một số các tính chất, hệ quả về các đỉnh treo tương đương làm cơ sở để thực hiện thuật toán kết hợp lớp đỉnh treo tương đương, có cùng độ đo trung tâm trung gian thành một đỉnh đại diện nhằm giảm thiểu không gian tính toán của đồ thị mạng xã hội. Các tính chất sau đây khẳng định độ đo trung tâm trung gian của các đỉnh trong đồ thị rút gọn cũng chính là độ đo trung tâm trung gian của các đỉnh trên đồ thị ban đầu.

Định nghĩa 2.1. Đỉnh v ∈ V của đồ thị G = (V, E) là đỉnh treo (leaf vertex) [84] nếu bậc của v là 1, kí hiệu deg(v) = 1.

Tính chất 2.1. Nếu v là đỉnh treo của đồ thị G và e = (v, w) ∈ E thì:

(i) $C_B(v) = 0$ (2.4)

(ii) $C_B(e) = (|V| - 1)$ (2.5)

Chứng minh.

(i) Vì v là đỉnh treo nên $\sigma_{st}(v) = 0$, theo công thức (2.1) thì $C_B(v) = 0$

(ii) G là liên thông nên với mọi v' ∈ V - {v} đều có đường đi tới v, nghĩa là có đường đi ngắn nhất giữa v và v'. Vì v là đỉnh treo của đồ thị nên mọi đường đi ngắn nhất giữa chúng đều đi qua e = (v, w). Theo định nghĩa độ đo trung tâm trung gian của cạnh thì $C_B(e) = (|V| - 1)$. ■

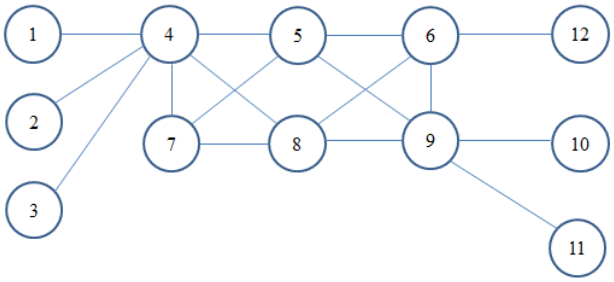
Định nghĩa 2.2. Cho trước đồ thị vô hướng liên thông G = (V, E) với u, w ∈ V là hai đỉnh treo, u tương đương bậc 1 với w, ký hiệu u $\approx_1$ w khi và chỉ khi chúng cùng liền kề với v (N(u) = N(w) = {v}), N(u) là tập các đỉnh lân cận của u. [84]

Dễ nhận ra quan hệ $\approx_1$ là quan hệ tương đương. Ta ký hiệu V1 là tập tất cả các đỉnh treo của đồ thị G thì: V1 = { u ∈ V | deg(u) = 1}.

V1/$\approx_1$ sẽ tạo ra các lớp các đỉnh treo tương đương với nhau, V1/$\approx_1$ = {C_1, C_2, …, C_k}. Những đỉnh treo cùng liền kề với một đỉnh sẽ cùng lớp tương đương và chúng có cùng bậc 1 và cùng độ đo trung tâm trung gian là 0.

Như vậy, u $\approx_1$ w thì u, w ∈ C_i, i = 1..k và deg(u) = deg(w) = 1, $C_B(u) = C_B(w) = 0$.

Ví dụ 2.1. Xét đồ thị vô hướng liên thông G sau:



Hình 2.1. Đồ thị vô hướng liên thông G

Như trên đã phân tích, tất cả các đỉnh treo đều có độ đo trung tâm trung gian bằng 0, nghĩa là:

$C_B(1) = C_B(2) = C_B(3) = C_B(10) = C_B(11) = C_B(12) = 0$

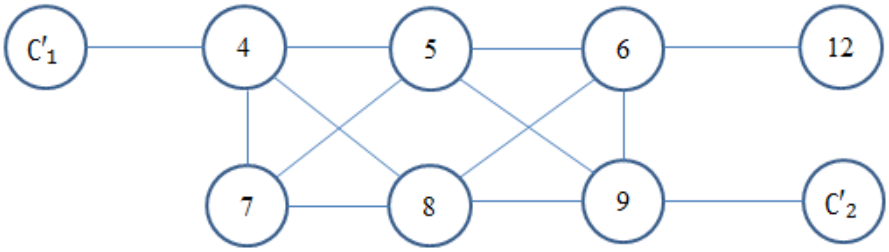
Đỉnh số 4 có 3 đỉnh treo liền kề, N(4) = {1, 2, 3}, $1 \approx_1 2 \approx_1 3$, tương tự N(9) = {10, 11} và $10 \approx_1 11$.

Nhiệm vụ chính là tính độ đo trung tâm trung gian của các đỉnh trên đồ thị, nên việc kết hợp những đỉnh tương đương với nhau (về độ đo trung tâm trung gian) thành một đỉnh đại diện cho những lớp có số phần tử lớn hơn hoặc bằng 2, sẽ làm giảm đáng kể các đỉnh cần tính độ đo trung tâm trung gian. Sau khi kết hợp tất cả những đỉnh tương đương của lớp C_i, $|C_i| \geq 2$, i = 1..k, thành đỉnh đại diện C'_i (cũng là đỉnh treo), ta nhận được đồ thị $G_1 = (V_1, E_1)$, trong đó:

- $V_1 = V - V1 \cup \{C'_1, C'_2, \ldots, C'_k\}$ (*)
- $E_1 = E - \{(u, v) \mid u \in V1, v = N(u)\} \cup \{(v, C'_i) \mid i = 1..k, v = N(u) \text{ với } u \in C_i\}$

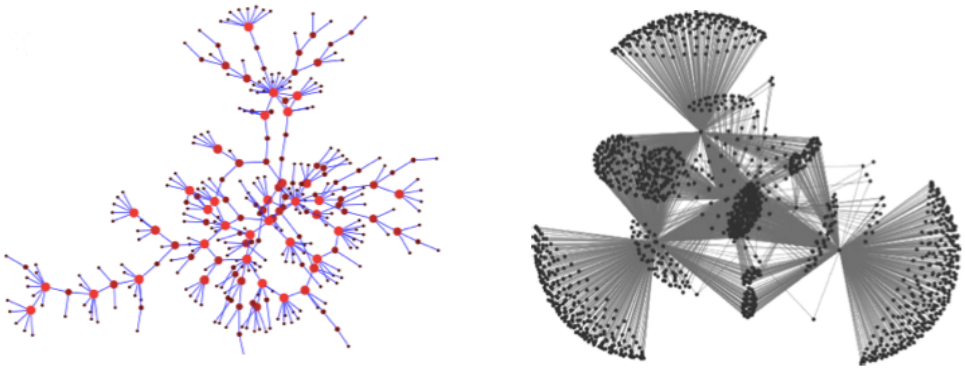
Đồ thị G_1 nhận được từ đồ thị G sau khi loại bỏ đi những đỉnh treo tương đương với nhau và các cạnh liền kề với chúng, thay thế bằng một đỉnh có tên trùng với tên của lớp và một cạnh liền kề với một đỉnh đại diện cho mỗi lớp tương đương.

Ví dụ 2.2. Xét đồ thị G từ ví dụ 2.1 có V = {đỉnh 1, đỉnh 2, …, đỉnh 12}, ta thấy các đỉnh 1, đỉnh 2, đỉnh 3 là các đỉnh treo liền kề với đỉnh 4, do vậy chúng tương đương bậc 1 với nhau. Tương tự, đỉnh 10, đỉnh 11 liền kề với đỉnh 9 và tương đương bậc 1 với nhau. Ở đồ thị này có 2 lớp tương đương bậc 1 có nhiều hơn 1 phần tử là: C_1 = {đỉnh 1, đỉnh 2, đỉnh 3}, C_2 = {đỉnh 10, đỉnh 11}. Kết hợp các đỉnh treo tương đương của, ta có đồ thị G_1.



Hình 2.2. Đồ thị G_1 kết hợp các đỉnh treo tương đương

Ví dụ 2.3. Các mạng xã hội có nhiều đỉnh treo tương đương [22], [83].



Hình 2.3. Minh họa các mạng xã hội xuất hiện nhiều đỉnh treo

Để chứng minh rằng được độ đo trung tâm trung gian của các đỉnh trong đồ thị G_1 cũng chính là độ đo trung tâm trung gian của các đỉnh trên đồ thị G ban đầu, nghĩa là đồ thị rút gọn bảo toàn độ đo trung tâm trung gian của các đỉnh, ta sử dụng các tính chất sau:

Tính chất 2.2. Với mọi đỉnh treo $u \in V$ hay $deg(u) = 1$, $v \in V$ là đỉnh liền kề với đỉnh u. Tập các đỉnh treo liền kề với v ký hiệu $N_1(v) = \{ w \in V \mid (w, v) \in E, deg(w) = 1\}$. Khi đó, ta có các tính chất sau:

(i) $\delta_{ut} = \delta_{vt}$, với mọi $t \in V - \{u, v\}$ (2.6)

(ii) $\delta_{ut}(w) = \delta_{vt}(w)$, với mọi $w \in V - \{s \in V | deg(s) = 1\}$, $t \in V - \{u, v, w\}$ (2.7)

(iii) $\tau_{ut}(v) = 1$, với mọi đỉnh $t \in V - \{u, v\}$ (2.8)

(iv) $C_B(v)=|N_1(v)|*(|N_1(v)|-1)/2+|N_1(v)|*|V-\{v\}-N_1(v)|\sum_{s\neq v\neq t\in V-N_1(v)} \frac{\delta_{st}(v)}{\delta_{st}}$ (2.9)

Chứng minh.

(i) u là đỉnh treo và liền kề với v, nên mọi đường đi từ u đến t đều phải đi qua v, nên $\delta_{ut}(v) = \delta_{uv} * \delta_{vt} = \delta_{vt}$, vì $\delta_{uv} = 1$ ($deg(u) = 1$).

(ii) u là đỉnh treo và liền kề với v, nên mọi đường đi từ u đến t đều phải đi qua v, nên $\delta_{ut}(w) = \delta_{uv} * \delta_{vt}(w) = \delta_{vt}(w)$ vì $\delta_{uv} = 1$.

(iii) Mọi đường đi ngắn nhất đi từ đỉnh treo u ($deg(u) = 1$) đến $t \in V - \{u, v\}$ (những đỉnh còn lại của đồ thị) đều phải đi qua đỉnh liền kề với u là v, vậy suy ra $\delta_{ut}(v) = \delta_{ut}$, nghĩa là $\tau_{ut}(v) = 1$.

(iv) $N_1(v)$ là tập các đỉnh treo liền kề với v. Nếu $|N_1(v)| > 1$ thì giữa 2 đỉnh treo liền kề với v sẽ có 1 đường đi ngắn nhất đi qua v. Khi đó ta có:
$|N_1(v)|*(|N_1(v)|-1)/2|$ đường đi ngắn nhất giữa 2 đỉnh treo liền kề với v và đi qua v. Với mỗi đỉnh treo u liền kề với v, tiếp tục dựa vào tính chất 2.2 (i) ta có: $|V-\{N_1(v), v\}|$ đường đi ngắn nhất đi từ u tới những đỉnh còn lại. Đỉnh v có $|N_1(v)|$ đỉnh treo liền kề, do vậy sẽ có $|N_1(v)|*|V-\{u, v\}|$ đường đi ngắn nhất từ các đỉnh treo liền kề với v, đi qua v để đến những đỉnh còn lại của đồ thị. Còn lại, $\sum_{s\neq v\neq t, s,t\in V-N_1(v)} \frac{\delta_{st}(v)}{\delta_{st}}$ là tổng các độ phụ thuộc các cặp đỉnh còn lại (V-{$N_1(v)$}-{v}) vào v (đi qua v). ■

Tính chất 2.3. Giả sử G_1 là đồ thị rút gọn của đồ thị G sau khi kết hợp đỉnh tương đương bậc một. Ta có tính chất sau:

(i) $\tau_{st}(v) = |C_i| * \tau_{ut}(v)$ nếu $s = C'_i$, $i = 1..k$, $u = N(C'_i)$, $v \neq t \in V_1 - \{u, C'_i\}$ (2.10)

(ii) $\tau_{st}(v) = |C_i| * \tau_{sw}(v)$ nếu $t = C'_i$, $i = 1..k$, $w = N(C'_i)$, $v \neq t \in V_1 - \{w, C'_i\}$ (2.11)

(iii) $\tau_{st}(v) = |C_i| * |C_j| * \tau_{uw}(v)$ nếu $s = C'_i$, $t = C'_j$, $i, j = 1..k$, $u = N(C'_i)$, $w = N(C'_j)$, $v \in V_1 - \{u, w, C'_i, C'_j\}$ (2.12)

Chứng minh.

(i) Đỉnh đầu $s = C'_i$ đại diện cho lớp tương đương có $|C_i| \geq 2$ phần tử. Vì s cũng là đỉnh treo, nên mọi đường đi ngắn nhất từ $s = C'_i$ đến t qua v đều phải đi qua một đỉnh liền kề của s là u. Trên đồ thị G_1, mỗi đỉnh C'_i sẽ đại diện cho $|C_i|$ đỉnh treo tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v)$ sẽ tương ứng với $|C_i| * \tau_{ut}(v)$.

(ii) Đỉnh cuối $t = C'_i$ đại diện cho lớp tương đương có $|C_i| \geq 2$ phần tử. Vì t cũng là đỉnh treo, nên mọi đường đi ngắn nhất từ s đến $t = C'_i$ qua v đều phải đi qua một đỉnh liền kề (tiền tố) của t là w. Trên đồ thị G_1, mỗi đỉnh C'_i sẽ đại diện cho $|C_i|$ đỉnh treo tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |C_i| * \tau_{sw}(v)$.

(iii) Trường hợp đỉnh đầu $s = C'_i$ đại diện cho lớp tương đương có $|C_i| \geq 2$ phần tử và đỉnh cuối $t = C'_j$ đại diện cho lớp tương đương có $|C_j| \geq 2$ phần tử. Vì s, t

cũng đều là đỉnh treo, nên mọi đường đi ngắn nhất từ s = C'_i đến t qua v đều phải đi qua một đỉnh liền kề (hậu tố) của s là u và phải đi qua một đỉnh liền kề (tiền tố) của t là w. Trên đồ thị G_1, mỗi đỉnh C'_i sẽ đại diện cho $|C_i|$ đỉnh treo tương đương và mỗi đỉnh C'_j sẽ đại diện cho $|C_j|$ đỉnh treo tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |C_i| * |C_j| * \tau_{uw}(v)$. ■

2.2.2. Các lớp đỉnh sườn tương đương

Mục này đề xuất một số các tính chất, hệ quả về lớp đỉnh sườn tương đương trên đồ thị làm cơ sở để thực hiện thuật toán kết hợp lớp đỉnh sườn tương đương về độ đo trung tâm trung gian thành một đỉnh đại diện bảo toàn độ đo trung tâm trung gian, nhằm giảm thiểu không gian tính toán của đồ thị mạng xã hội. Các tính chất sau khẳng định độ đo trung tâm trung gian của các đỉnh đại diện trong đồ thị rút gọn cũng chính là độ đo trung tâm trung gian của các đỉnh trong lớp tương đương trên đồ thị ban đầu.

Định nghĩa 2.3. Cho đồ thị vô hướng, liên thông G = (V, E), u ∈ V được gọi là đỉnh sườn (Side vertex) [84] của G nếu đồ thị con sinh bởi tập các đỉnh liền kề N(u) là clique (đồ thị con đầy đủ).

Ở đây, ta chỉ xét những đỉnh sườn có $|N(u)| \geq 2$, vì trường hợp ngược lại, $|N(u)| = 1$ thì u là đỉnh treo đã được nghiên cứu ở trên.

Nhận xét 2.1. Nếu u là đỉnh sườn và G không phải là clique thì chắc chắn có ít nhất một đỉnh v ∈ N(u) có bậc khác với bậc của đỉnh sườn u (deg(v) > deg(u)) trên đồ thị G.

Ký hiệu Γ(u) = {u} ∪ N(u) là tập các đỉnh liền kề với u và kể cả u.

Nhận xét 2.2. Đồ thị con sinh bởi Γ(u) cũng là clique, vì bản thân N(u) đã sinh ra là clique, và u lại liền kề với tất cả các đỉnh của N(u).

Γ1(u) = { v ∈ Γ(u) | deg(v) = deg(u)} - Tập những đỉnh có cùng bậc với đỉnh sườn u trong clique sinh bởi Γ(u).

Nhận xét 2.3. Nếu G không phải là clique thì Γ2(u) = Γ(u) - Γ1(u) ≠ ϕ, nghĩa là chắc chắn có ít nhất một đỉnh v ∈ Γ2(u) trên clique sinh bởi N(u) (hay Γ(u)) có bậc khác với bậc của đỉnh sườn (deg(v) > deg(u)).

Để thực hiện thuật toán tính độ đo trung tâm trung gian của các đỉnh trên đồ thị một cách hiệu quả, người ta thường sử dụng phương pháp duyệt theo chiều rộng BFS (Breadth-First Search) [55]. Thuật toán duyệt theo chiều rộng tìm kiếm các đường đi ngắn nhất từ đỉnh gốc qua các cạnh tới tất cả các đỉnh khác trong đồ thị. Các cạnh giữa các mức của quá trình duyệt BFS bắt đầu từ đỉnh gốc X sẽ tạo thành đồ thị định hướng, phi chu trình, được gọi DAG_X.

Tính chất 2.4. Nếu u là đỉnh sườn của đồ thị G = (V, E), thì u hoặc là gốc hoặc là lá trên cây DAG duyệt theo chiều rộng (BFS).

Chứng minh.

Giả sử u không phải là đỉnh gốc và cũng không phải là lá trên cây DAG duyệt theo BFS. Vì u là đỉnh trong trên DAG, nên có đường đi ngắn nhất đi qua u. Mọi đường đi ngắn nhất từ gốc đi qua đỉnh u thì phải đi qua ít nhất hai đỉnh v, w liền kề với đỉnh u. Khi u là đỉnh sườn, theo định nghĩa thì N(u) là clique, nghĩa là $(v, w) \in E$, với mọi v, w liền kề với u. Khi đó đường đi qua u không phải đường đi ngắn nhất trên DAG, điều này mâu thuẫn với tính chất là mọi đường đi trên DAG là đường đi ngắn nhất. ■

Tính chất 2.5. Giả sử S là tập các đỉnh sườn tương đương, $S = \{v_1, v_2, \ldots, v_h\}$ và nếu chọn một đỉnh sườn v_i, i = 1..h, làm gốc để duyệt BFS, thì h-1 đỉnh còn lại đều là lá có độ dài từ gốc là 2 và độ đo trung tâm trung gian của các cạnh liền kề của đỉnh sườn với các đỉnh liền kề tương ứng trên DAG_{vi} là như nhau, $C_B((v, v_j)) = 1/ |N(S)|$, với mọi $j \neq i$, $v \in N(S)$.

Chứng minh.

Theo giả thiết, các đỉnh v_i, i = 1..h, là các đỉnh sườn tương đương, chúng có cùng tập các đỉnh liền kề $N(S) = N(v_i)$, i = 1..h. Khi lấy một đỉnh v_i để duyệt BFS, thì có |N(S)| đỉnh liền kề với v_i đều ở mức 1 (có khoảng cách tới gốc là 1), đồng thời $v \in N(S)$ lại là cha của tất cả các đỉnh sườn tương đương còn lại v_j (ở mức 2), nghĩa là v_j sẽ có |N(S)| đỉnh cha, và do vậy, tỷ số đường đi ngắn nhất từ v_i (đỉnh gốc) đến các đỉnh sườn tương đương khác trên DAG_{vi} và đi qua mỗi cạnh (v, v_j) là 1 / |N(S)|. ■

Tính chất 2.6. Giả sử S là tập các đỉnh sườn tương đương, $S = \{v_1, v_2, \ldots, v_h\}$ và $N(S) = N(v_i)$, i = 1..h, thì độ đo trung tâm trung gian của các cạnh nối đỉnh sườn với các đỉnh liền kề tương ứng là như nhau: $C_B((v_i, v)) = C_B((v_j, v))$, với mọi $v_i, v_j \in S$, $v \in N(S)$.

Chứng minh.

Theo giả thiết, các đỉnh v_i, i = 1..h là các đỉnh sườn tương đương, nên chúng có cùng tập các đỉnh liền kề $N(S) = N(v_i)$, i = 1..h. Khi đó, với mọi đỉnh $v \in N(S)$, đều có cạnh $e_i = (v_i, v) \in E$, với mọi i = 1..h. Theo tính chất 2.4 mọi đỉnh sườn chỉ có thể là đỉnh gốc hoặc là đỉnh lá trên các DAG, suy ra $C_B(e_i) = C_B(e_j)$, với mọi i, j = 1..h. ■

Tính chất 2.7. Giả sử S là tập các đỉnh sườn tương đương, $S = \{v_1, v_2, \ldots, v_h\}$ và $N(S) = N(v_i)$, i = 1..h. Khi đó các đồ thị DAG_v duyệt theo BFS, với mọi $v \in S$ đều có chung một đồ thị con sinh bởi tập đỉnh $V_S = V - S$.

Đồ thị con sinh chung của DAG_v, $v \in S$, là đồ thị thu được từ DAG_v, sau khi bỏ đi các đỉnh sườn tương đương S cùng các cạnh liên thuộc với chúng. Đồ thị chung này có thể không liên thông.

Tính chất 2.8. Nếu u là đỉnh sườn của đồ thị G, thì

(i) $\delta_{st}(v) = 0$, với mọi $v \in \Gamma 1(u)$, $s \neq u \neq t \in V$ (2.13)

(ii) $C_B(v) = 0$, với mọi $v \in \Gamma 1(u)$ (2.14)

Chứng minh.

(i) Trường hợp $s \neq v \neq t \in V$, mọi đường đi ngắn nhất đi qua một đỉnh u thì chúng phải đi qua ít nhất 2 đỉnh v, w liền kề với đỉnh u. Khi u là đỉnh sườn, theo định nghĩa thì N(u) là clique, nghĩa là $(v, w) \in E$. Theo các tính chất của đường đi ngắn nhất thì $\delta_{vw}(u) = 0$ vì $d_G(v, w) < d_G(v, u) + d_G(u, w)$. Như vậy ta có $\delta_{st}(u) = \delta_{sv} * \delta_{vw}(u) * \delta_{wt} = 0$.

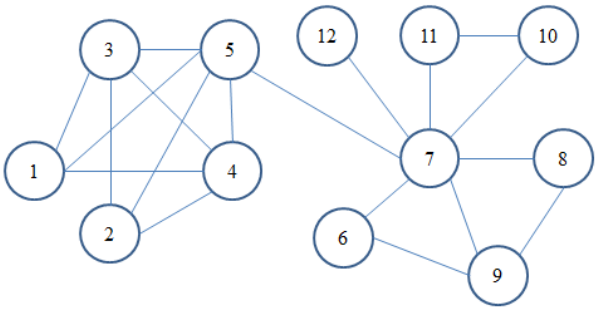
(ii) Như nhận xét ở trên thì đồ thị con sinh bởi Γ(u) là clique và deg(v) = deg(u), nên theo tính chất của đường đi ngắn nhất và tính chất 2.8 (i) suy ra $\delta_{st}(v) = 0$, với mọi $v \in \Gamma 1(u)$. ■

Định nghĩa 2.4. Cho u, v ∈ V, có quan hệ $\approx_2$ với nhau, ký hiệu u $\approx_2$ v khi và chỉ khi u, v là hai đỉnh sườn của G và N(u) = N(v). [84]

Nhận xét: Quan hệ $\approx_2$ là quan hệ tương đương.

Từ định nghĩa và tính chất 2.8 suy ra, u $\approx_2$ v thì deg(u) = deg(v) = |N(u)| - 1| và $C_B(u) = C_B(v) = 0$.

Ví dụ 2.4. Xét đồ thị G được cho như trong Hình 2.4 dưới đây.



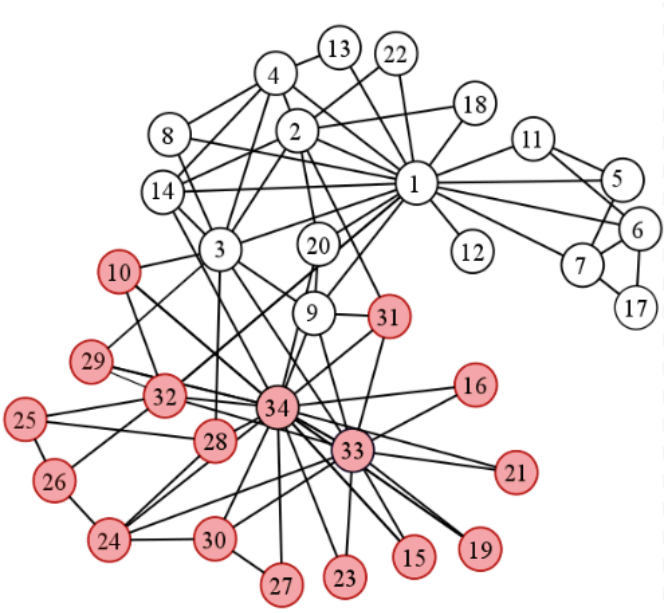
Hình 2.4. Đồ thị G có các đỉnh sườn tương đương

Các đỉnh 1, 2 là hai đỉnh sườn tương đương, 1 $\approx_2$ 2, bởi N(1) = N(2) = {3, 4, 5} và đồ thị con sinh bởi {3, 4, 5} là clique. Ngoài ra, các đỉnh 6, 8, 9, 10, 11 cũng là các đỉnh sườn vì đồ thị con sinh bởi tập liền kề với chúng cũng đều là clique (2 đỉnh có 1 cạnh nối với nhau), trong đó 6 $\approx_2$ 8, bởi N(6) = N(8) = {7, 9}.

Những đỉnh sườn tương đương có thể kết hợp thành một đỉnh đại diện để rút gọn số đỉnh sườn tương đương trên đồ thị. Giả sử G = (V, E) có các lớp đỉnh sườn tương đương S_i, i = 1..h, mỗi lớp có ít nhất 2 đỉnh sườn tương đương với nhau. Kết hợp những đỉnh tương đương trong cùng lớp thành một đỉnh sườn đại diện, ta nhận được đồ thị $G_2 = (V_2, E_2)$, trong đó:

- $V_2 = V - V2 \cup \{S'_1, S'_2, \ldots, S'_h\}$, với $V2 = S_1 \cup S_2 \cup \ldots \cup S_h$. (**)
- $E_2 = E - \{(u, v) \mid u \in V2, v \in N(u)\} \cup \{(v, S'_i) \mid i = 1..h, v \in N(u) \text{ với } u \in S_i\}$

Ví dụ 2.5. Đồ thị mạng xã hội câu lạc bộ Karate của Zachary [31] xuất hiện các đỉnh sườn tương đương.



Hình 2.5. Đồ thị mạng xã hội câu lạc bộ karate của Zachary xuất hiện nhiều đỉnh sườn.

Trên đồ thị mạng xã hội câu lạc bộ karate của Zachary [31] gồm có 34 đỉnh và 68 cạnh thì đỉnh số 18 và đỉnh số 22 là những đỉnh sườn và chúng có thể kết hợp với nhau thành một đỉnh đại diện. Tương tự như vậy các đỉnh số 15, đỉnh số 16, đỉnh số 19, đỉnh số 21 và đỉnh số 23 cũng là những đỉnh sườn tương đương và chúng có thể kết hợp với nhau thành một đỉnh đại diện duy nhất. Như vậy có thể kết hợp các đỉnh sườn tương đương ta thu được đồ thị mạng xã hội rút gọn gồm 29 đỉnh và 58 cạnh đã giảm 5 đỉnh và 10 cạnh so với đồ thị mạng xã hội ban đầu.

Để chứng minh được độ đo trung tâm trung gian của các đỉnh trong đồ thị G_2 rút gọn cũng chính là độ đo trung tâm trung gian của các đỉnh trên đồ thị G ban đầu, nghĩa là đồ thị rút gọn bảo toàn độ đo trung tâm trung gian của đồ thị ban đầu, ta sử dụng các tính chất sau:

Tính chất 2.9. Giả sử G_2 là đồ thị rút gọn của đồ thị G sau khi kết hợp các đỉnh sườn tương đương của các lớp S_i thành một đỉnh đại diện S'_i, i = 1..h. Ký hiệu $\Gamma 2(S'_i) = \Gamma 2(u)$, với $u \in S_i$. Ta có tính chất sau:

(i) $\tau_{S'_i t}(v) = |S_i| * \delta_{ut}$, $u \in \Gamma 2(S'_i)$, i =1..h, u = v, $t \notin \{u, S'_1, S'_2, \ldots, S'_h\}$ (2.15)

(ii) $\tau_{S'_i t}(v) = |S_i| * \tau_{ut}(v)$, $u \in \Gamma 2(S'_i)$, i = 1..h, $u \neq v$, $t \notin \{u, v, S'_1, S'_2, \ldots, S'_h.$ (2.16)

(iii) $\tau_{s S'_i}(v) = |S_i| * \delta_{sw}$, $w \in \Gamma 2(S'_i)$, $i = 1..h$, $w = v$, $s \notin \{v, S'_1, S'_2, \ldots, S'_h\}$ (2.17)

(iv) $\tau_{s S'_i}(v) = |S_i| * \tau_{sw}(v)$, $w \in \Gamma 2(S'_i)$, $i = 1..h$, $v \neq w$, $s \notin \{w, S'_1, S'_2, \ldots, S'_h\}$ (2.18)

(v) $\tau_{S'_i S'_j}(v) = |S_i| * |S_j| * \tau_{u w}(v)$, $u \in \Gamma 2(S'_i)$, $w \in \Gamma 2(S'_j)$, $i, j = 1..h$, $v \notin \{u, w, S'_i, S'_j\}$ (2.19)

Chứng minh.

(i) Đỉnh đầu $s = S'_i$ đại diện cho lớp tương đương có $|S_i| \geq 2$ phần tử. Vì s cũng là đỉnh sườn, nên mọi đường đi ngắn nhất từ $s = S'_i$ đến t qua v đều phải đi qua một đỉnh liền kề của s là u, $u \in \Gamma 2(S'_i)$, u không phải là đỉnh sườn. Trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh sườn tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |S_i| * \delta_{ut}$ khi $v = u$.

(ii) Tương tự (2.15), trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh sườn tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |S_i| * \tau_{ut}(v)$, khi $v \neq u$.

(iii) Đỉnh cuối $t = S'_i$ đại diện cho lớp tương đương có $|S_i| \geq 2$ phần tử. Vì t cũng là đỉnh sườn, nên mọi đường đi ngắn nhất từ s đến $t = S'_i$ qua v đều phải đi qua một đỉnh liền kề (tiền tố) của t là $w \in \Gamma 2(S'_i)$, w không phải là đỉnh sườn. Trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh sườn tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |S_i| * \delta_{sw}$ khi $v = w$.

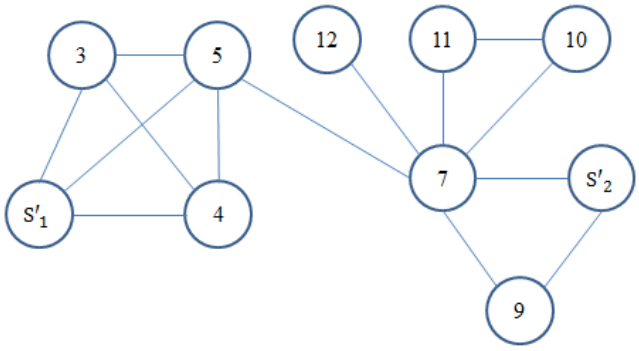
(iv) Tương tự (2.17), trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh sườn tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |S_i| * \tau_{sw}(v)$ khi $v \neq w$.

(v) Trường hợp đỉnh đầu $s = S'_i$ đại diện cho lớp tương đương có $|S_i| \geq 2$ phần tử và đỉnh cuối $t = S'_j$ đại diện cho lớp tương đương có $|S_j| \geq 2$ phần tử. Vì s, t cũng đều là đỉnh sườn, nên mọi đường đi ngắn nhất từ $s = S'_i$ đến t qua v đều phải đi qua một đỉnh liền kề (hậu tố) của s là u và phải đi qua một đỉnh liền kề (tiền tố) của t là w. Trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh sườn tương đương và mỗi đỉnh S'_j sẽ đại diện cho $|S_j|$ đỉnh sườn tương đương ở đồ thị G,

nên số đường đi ngắn nhất từ s đến t qua v: $\tau_{st}(v) = |C_i| * |C_j| * \tau_{uw}(v)$ khi $v \neq u$, $v \neq w$. ■

Ví dụ 2.6. Xét đồ thị G_2 được rút gọn từ G cho trước trong Hình 2.1.

Các đỉnh 1 và đỉnh 2 là hai đỉnh sườn tương đương, đỉnh 1 $\approx_2$ đỉnh 2, được kết hợp với nhau thành đỉnh đại diện S'_1 và các đỉnh 6 kết hợp với đỉnh 8 thành S'_2 như trong hình 2.6



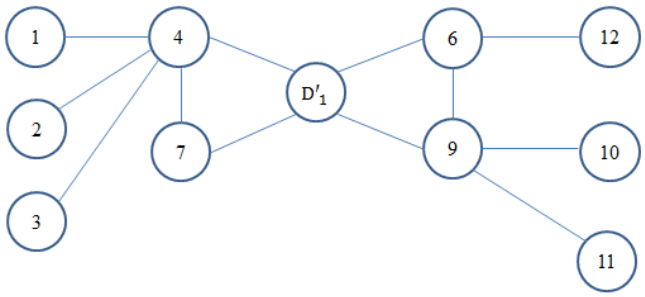
Hình 2.6. Đồ thị G_2 được rút gọn bằng cách kết hợp đỉnh 1 và đỉnh 2 thành đỉnh sườn S'_1, còn đỉnh 6 và đỉnh 8 kết hợp thành S'_2.

2.2.3. Các lớp đỉnh đồng nhất tương đương

Định nghĩa 2.5. Cho đồ thị vô hướng, liên thông G = (V, E). Hai đỉnh u, v ∈ V được gọi là đồng nhất (Identical vertex) [84] trên G, ký hiệu là u $\approx_3$ v khi và chỉ khi N(u) = N(v) ≥ 2 và đồ thị con sinh bởi N(u) không phải clique (đồ thị con đầy đủ).

Ví dụ 2.7. Xét đồ thị G được cho như trong Hình 2.1.

Ta thấy các đỉnh 5, đỉnh 8 là hai đỉnh đồng nhất, đỉnh 5 $\approx_3$ đỉnh 8, bởi N(đỉnh 5) = N(đỉnh 8) = {đỉnh 4, đỉnh 6, đỉnh 7, đỉnh 9}. Đồ thị G có một lớp D'_1 = {đỉnh 5, đỉnh 8} hai đỉnh đồng nhất với nhau. Kết hợp đỉnh 5 và đỉnh 8 thành một đỉnh đại diện D'_1 để được đồ thị rút gọn G_3 như hình 2.7



Hình 2.7. Đồ thị G_3 sau khi kết hợp các đỉnh đồng nhất tương đương

Những đỉnh đồng nhất có thể kết hợp thành một đỉnh đại diện để rút gọn số đỉnh trên đồ thị. Giả sử G = (V, E) có các lớp đỉnh đồng nhất D_i, i = 1..l, mỗi lớp có ít nhất 2 đỉnh đồng nhất với nhau. Kết hợp những đỉnh đồng nhất (trong cùng lớp) thành một đỉnh đồng nhất đại diện, ta nhận được đồ thị $G_3 = (V_3, E_3)$, trong đó:

- $V_3 = V - V3 \cup \{D'_1, D'_2, \ldots, D'_h\}$, với $V3 = D_1 \cup D_2 \cup \ldots \cup D_l$. (***)
- $E_3 = E - \{(u, v) \mid u \in V3, v \in N(u)\} \cup \{(v, D'_i) \mid i = 1..l, v \in N(u), \text{với } u \in S_i\}$

Ký hiệu $N(D'_i) = N(u)$, $u \in D_i$.

Tính chất 2.10. Nếu u, v là hai đỉnh đồng nhất ($u \approx_3 v$) trên đồ thị G, thì:

$\delta_{st}(u) = \delta_{st}(v)$, với mọi $s \neq v, u \neq t \in V$ (2.20)

Chứng minh.

Trường hợp $s \neq v, u \neq t \in V$, mọi đường đi ngắn nhất đi qua một đỉnh u thì, trước khi đi đến u, chúng phải đi qua ít nhất một đỉnh p liền kề với đỉnh u ($p \in N(v)$), sau đó phải đi qua một đỉnh q khác liền kề với u ($q \in N(u)$) sau khi đã qua u. Vì $u \approx v$, nên theo định nghĩa N(u) = N(v), do đó cũng sẽ có đường đi ngắn nhất. qua đỉnh v. Do vậy, $\delta_{st}(u) = \delta_{sp} * \delta_{pq}(u) * \delta_{qt} = \delta_{sp} * \delta_{pq}(v) * \delta_{qt} = \delta_{sp}(v)$. ■

Tính chất 2.11. Nếu u, v là hai đỉnh đồng nhất ($u \approx_3 v$) trên đồ thị G, thì:

$\delta_{st}(e_1) = \delta_{st}(e_2)$, với mọi $s \neq v, u \neq t \in V$, với mọi $w \in N(u) = N(v)$, $e_1 = (u, w)$, $e_2 = (v, w)$ (2.21)

Chứng minh.

Tương tự như tính chất 2.10, nhưng thay vì xét số đường đi ngắn nhất qua đỉnh, ở tính chất 2.11 xét số đường đi ngắn nhất qua cạnh liên thuộc với hai đỉnh tương đương

u, v và vì tập các đỉnh liền của u, v bằng nhau, nên số các đường đi ngắn nhất qua e_1, e_2 luôn bằng nhau. ■

Tính chất 2.12. Giả sử G_3 là đồ thị rút gọn của đồ thị G sau khi kết hợp các đỉnh đồng nhất của các lớp D_i thành một đỉnh đại diện D'_i, i = 1…l. Ta có các tính chất sau:

(i) $\delta_{D'it}(v) = |D_i| * \delta_{ut}$, $u \in N(D'_i)$, i =1..l, u = v, $t \notin \{u, D'_1, D'_2, \ldots, D'_l\}$ (2.22)

(ii) $\delta_{D'it}(v) = |D_i| * \delta_{ut}(v)$, $u \in N(D'_i)$, i =1..l, $u \neq v$, $t \notin \{u, D'_1, D'_2, \ldots, D'_l\}$ (2.23)

(iii) $\delta_{sD'i}(v) = |D_i| *_{\delta sw}$, $w \in N(D'_i)$, i = 1..l, w = v, $s \notin \{v, D'_1, D'_2, \ldots, D'_l\}$ (2.24)

(iv) $\delta_{sD'i}(v) = |D_i| *_{\delta sw}(v)$, $w \in N(D'_i)$, i =1..l, $v \neq w$, $s \notin \{w, D'_1, D'_2, \ldots, D'_l\}$ (2.25)

(v) $\delta_{D'iD'j}(v) = |D_i| * |D_j| * \delta_{uw}(v)$, $u \in N(D'_i)$, $w \in N(D'_j)$, i, j =1..l, $v \notin \{u, w, D'_i, D'_j\}$ (2.26)

Chứng minh.

(i) Đỉnh đầu $s = D'_i$ đại diện cho lớp tương đương có $|D_i| \geq 2$ phần tử, nên mọi đường đi ngắn nhất từ $s = D'_i$ đến t qua v đều phải đi qua một đỉnh liền kề của s là u, $u \in N(D'_i)$. Trên đồ thị G_3, mỗi đỉnh D'_i sẽ đại diện cho $|D_i|$ đỉnh đồng nhất ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v sẽ là: $\delta_{st}(v) = |D_i| * \delta_{ut}$ khi v = u.

(ii) Tương tự (2.22), trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh sườn tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\delta_{st}(v) = |S_i| * \delta_{ut}(v)$, khi $v \neq u$.

(iii) Đỉnh cuối $t = S'_i$ đại diện cho lớp tương đương có $|S_i| \geq 2$ phần tử. Vì t cũng là đỉnh sườn, nên mọi đường đi ngắn nhất từ s đến $t = S'_i$ qua v đều phải đi qua một đỉnh liền kề (tiền tố) của t là w không phải là đỉnh đồng nhất. Trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh đồng nhất tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\delta_{st}(v) = |S_i| * \delta_{sw}$ khi v = w.

(iv) Tương tự (2.24), trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\delta_{st}(v) = |S_i| * \tau_{sw}(v)$ khi $v \neq w$.

(v) Trường hợp đỉnh đầu $s = S'_i$ đại diện cho lớp tương đương có $|S_i| \geq 2$ phần tử và đỉnh cuối $t = S'_j$ đại diện cho lớp tương đương có $|S_j| \geq 2$ phần tử. Vì s, t cũng đều là

đỉnh đồng nhất, nên mọi đường đi ngắn nhất từ s = S'_i đến t qua v đều phải đi qua một đỉnh liền kề (hậu tố) của s là u và phải đi qua một đỉnh liền kề (tiền tố) của t là w. Trên đồ thị G_2, mỗi đỉnh S'_i sẽ đại diện cho $|S_i|$ đỉnh tương đương và mỗi đỉnh S'_j sẽ đại diện cho $|S_j|$ đỉnh tương đương ở đồ thị G, nên số đường đi ngắn nhất từ s đến t qua v: $\delta_{st}(v) = |C_i| * |C_j| * \tau_{uw}(v)$ khi $v \neq u, v \neq w$. Trên đồ thị G_3, đỉnh D'_i đại diện cho $|D_i| \geq 2$ đỉnh đồng nhất và trên đồ thị G mỗi đỉnh $v \in D_i$ liền kề với $|N(D'_i)| = |N(v)|$ đỉnh. Do vậy, luôn có $|N(D'_i)|$ đường đi ngắn nhất từ đỉnh đại diện D'_i tới chính nó, trong đó có một đường đi ngắn nhất đi qua đỉnh liền kề v của đỉnh đại diện D'_i trên đồ thị G_3. Từ đó suy ra $\delta_{D'iD'j}(v) = 1/|N(D'_i)|$ với $v \in N(D'_i)$. ■

2.3. Thuật toán rút gọn đồ thị dựa vào độ đo trung tâm trung gian

Độ đo trung tâm trung gian của cạnh được tính dựa vào số đường đi ngắn nhất giữa các cặp đỉnh khác nhau trên đồ thị. Đồ thị mạng xã hội thường rất phức tạp, có số đỉnh và cạnh là rất lớn, nên việc tính các độ đo trung tâm trung gian rất tốn thời gian. Bài toán tìm đường đi ngắn nhất giữa các đỉnh trên đồ thị là bài toán thuộc lớp NP-khó.

Dựa trên các tính chất của các đỉnh tương đương theo độ đo trung tâm trung gian được trình bày ở Mục 2.2, Mục này trình bày đề xuất thuật toán REG (Reduce Equivalance Graph) thực hiện kết hợp các đỉnh tương đương theo độ đo trung tâm trung gian trong đồ thị thành một đỉnh đại diện. Công việc rút gọn đồ thị này khác với rút gọn đồ thị thông thường ở chỗ rút gọn các lớp đỉnh tương đương theo độ đo trung tâm trung gian không làm thay đổi tính chất của đồ thị ban đầu và bảo toàn được giá trị của độ đo trung tâm trung gian.

Như vậy thuật toán REG thực hiện kết hợp các lớp đỉnh tương đương theo độ đo trung tâm trung gian trên đồ thị, giảm thiểu được số đỉnh và số cạnh trên đồ thị mạng xã hội. Qua đó làm tăng hiệu quả, rút gọn thời gian tính toán của các thuật toán tính độ đo trung tâm trung gian trên đồ thị. Đồng thời giúp tăng hiệu quả của nhóm các thuật toán phân tích, phát hiện các cấu trúc cộng đồng trên đồ thị mạng xã hội sử dụng độ đo trung tâm trung gian.

Thuật toán REG (Reduce Equivalance Graph)

Input: Đồ thị mạng xã hội G = (V, E)

Output: Đồ thị mạng xã hội $G_2 = (V_2, E_2)$ là đồ thị thu được sau khi thực hiện thuật toán rút gọn các lớp các đỉnh treo và đỉnh sườn tương đương về độ đo trung tâm trung gian trên đồ thị mạng xã hội.

Bước 1. Tìm tất cả các đỉnh treo và đỉnh sườn trên đồ thị

Bước 2. Tìm các lớp tương đương các đỉnh treo và đỉnh sườn trên đồ thị.

Bước 3. Kết hợp các lớp tương đương các đỉnh treo thành đỉnh treo đại diện và kết hợp các lớp đỉnh sườn thành đỉnh sườn đại diện. (Dựa vào (*) và (**)).

Giả mã thuật toán Reduce Equivalence Graph (REG)

```
Input: G = (V, E)
Output: G1 = (V1, E1) - đồ thị thu được sau khi kết hợp các đỉnh tương đương
        VC, VS - Tập các đỉnh treo, đỉnh sườn đại diện cho lớp tương đương
  V1 = V;
  E1 = E;
  P1 = ∅;            //Stack lưu các cặp (đỉnh treo, đỉnh liền kề)
  P2 = ∅;            //Stack lưu các cặp (đỉnh sườn, tập đỉnh liền kề)
      for u ∈ V1 do{
       N[u] = Neighbor(G, u);        // Tìm các đỉnh liền kề với u
        if(deg(u) == 1) then  { // Tìm tất cả các đỉnh treo
          v = N[u]; // N[u] là một đỉnh liền kề với u
             P1.push(u, v); // Lưu cặp (đỉnh treo u, đỉnh liền kề) vào P1
          V1 = V1 - {u}; E1 = E1 - {(u, v)}; // Loại bỏ đỉnh treo và cạnh liền kề
        } // if(deg(u) == 1)
       // Tìm tất cả các đỉnh sườn
        if (Clique(N[u])) then {      // Kiểm tra đồ thị cảm sinh N[u] là clique
          V1 = V1 - {u};       // Loại bỏ đỉnh sườn u khỏi đồ thị G
          for v ∈ N[u] do{    // Loại bỏ các cạnh liền kề với u
                E1 = E1 - {(u, v)};          // Loại bỏ các cạnh liền kề với u
                P2.push(u, N[u]); // Lưu cặp đỉnh sườn và tập liền kề (u, N[u]) vào P2
           }
        }
      } // for u ∈ V1
     // Tìm các lớp tương đương của các đỉnh treo
      k = 1;
      (u, v) = P1.pop(); // Lấy ra từng cặp đỉnh tương ứng của đỉnh treo
      C[k] = {u}          // Mảng lớp các đỉnh treo tương đương
      N[k] = v;           // Đỉnh liền kề với đỉnh thuộc lớp C[k]
```

```
while( P1 != ∅) do {
    (u, v) = P1.pop();
    j = 1;
    loop = true;
    while (j <= k && loop) do {
        if (N[j] == v) then {    // Kiểm tra những đỉnh treo có đỉnh liền kề là v
            C[j] = C[j] ∪ {u}; // Để đưa vào cùng một lớp
            loop = false;
        } else j = j + 1;
    }
    if (loop) then {        // Khi đỉnh các lớp trước không tương đương với u
        k = k + 1;// Tìm lớp tiếp theo
        C[k] = {u};
        N[k] = v;
    }
} // while( P1 != ∅)
// Kết hợp các đỉnh tương tương đương ở lớp C[j] thành đỉnh treo đại diện C'j
V_C = ∅;
for j = 1 to k do {  // k lớp tương đương các đỉnh treo
    V_C = V_C ∪ {C'j};
    E_1 = E_1 ∪ {(C'j, N[j])} // Bổ sung thêm cạnh nối đỉnh đại diện lớp tương
    V_1 = V_1 ∪ V_C;
}// for j = 1 to k
// Tìm các lớp tương đương của các đỉnh sườn
h = 1;
(u, M) = P2.pop();          // Lấy ra từng cặp (đỉnh sườn, tập đỉnh liền kề)
S[h] = {u}                  // Mảng các lớp đỉnh sườn
N[h] = M;                   // Tập các đỉnh liền kề với đỉnh sườn thuộc lớp S[h]
while( P2 != ∅) do {
    (u, M) = P2.pop();
    j = 1;
    loop = true;
    while (j <= h && loop) do {
        if(N[j] == M) then {   // Kiểm tra những đỉnh sườn tương đương
            S[j] = S[j] ∪ {u};  // Đỉnh u vào lớp tương đương S[j]
            loop = false;
        }else j = j +1;
    }
    if (loop) then {
        h = h + 1;
        S[h] = {u};
        N[h] = M;
    }
}
// Kết hợp các đỉnh tương tương đương ở lớp S[j] thành đỉnh sườn đại diện Sj
V_S = ∅;
```

```
for j = 1 to h do {
    Vs = Vs ∪ {S'j}; // Tập các đỉnh sườn đại diện tương đương
    for v ∈ N[j] do {
        E1 = E1 ∪ {(S'j, v)}
    }
}
for u ∈ V1 do {
    W1(u) = W(u);   // Những đỉnh không là đỉnh sườn vẫn giữ nguyên trọng số
}
V1 = V1 ∪ Vs;
```

Thủ tục tính toán Neighbor(G, u): Tìm các đỉnh liền kề của u trong đồ thị G
Input: Đồ thị G = (V, E, W) và đỉnh u ∈ V
Output: N - tập các đỉnh liền của u trong đồ thị G

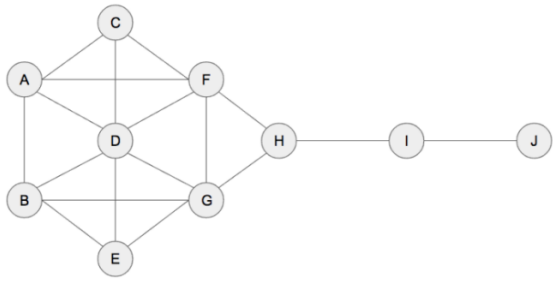
```
N = ∅;
for v ∈ V do {
    if ((u, v) ∈ E) then {
        N = N ∪ {v};
    }
}
return N;
```

Thủ tục tính toán Clique(G, N): Kiểm tra xem đồ thị con sinh bởi tập N trong đồ thị G có là đồ thị con đầy đủ hay không.
Input: Đồ thị G = (V, E, W) và tập đỉnh N ⊆ V
Output: **True** nếu đồ thị con sinh bởi tập đỉnh N trong đồ thị G_1 là clique, ngược lại **False**

```
for u ∈ N do {
    for v ∈ N - {u} do {
        if ((u, v) ∉ E) then {
            return false;
        }
    }
}
return true;
```

Ví dụ 2.8. Minh họa thuật toán REG

Xét đồ thị mạng xã hội Kite [56] như sau:



Hình 2.8. Đồ thị mạng xã hội Kite

Thực hiện tính độ đo trung tâm trung gian của các đỉnh trên đồ thị mạng xã hội Kite ta có được Bảng 2.1 như sau.

Bảng 2.1. Độ đo trung tâm trung gian các đỉnh trên đồ thị mạng xã hội Kite

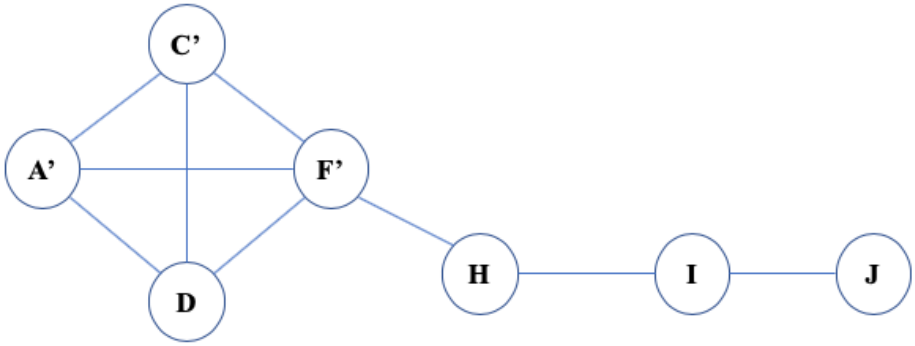
Stt	Đỉnh của đồ thị mạng xã hội	Độ đo trung tâm trung gian của đỉnh
1	Đỉnh A	0.023
2	Đỉnh B	0.023
3	Đỉnh C	0.000
4	Đỉnh D	0.102
5	Đỉnh E	0.000
6	Đỉnh F	0.231
7	Đỉnh G	0.231
8	Đỉnh H	0.389
9	Đỉnh I	0.222
10	Đỉnh J	0.000

Qua số liệu tại Bảng 2.1 ta thấy độ đo trung tâm trung gian của các đỉnh trong đồ thị mạng xã hội như sau:

$C_B(A) = C_B(B) = 0.023$, $C_B(C) = C_B(E) = 0.000$, $C_B(F) = C_B(G) = 0.231$

Như vậy các đỉnh tương đương với nhau về độ đo trung tâm trung gian: gồm đỉnh A với đỉnh B, đỉnh C với đỉnh E, đỉnh F với đỉnh G.

Thuật toán REG thực hiện kết hợp các đỉnh A và đỉnh B thành một đỉnh đại diện A’, tương tự kết hợp đỉnh C và đỉnh E thành một đỉnh đại diện C’ và cuối cùng kết hợp đỉnh F và đỉnh G thành một đỉnh F’. Kết quả thu được đồ thị mạng xã hội Kite rút gọn như sau:



Hình 2.9. Đồ thị mạng xã hội Kite rút gọn

Độ phức tạp của thuật toán REG.

Thuật toán REG (G) thực hiện qua ba bước.

Bước 1. Có độ phức tạp tính toán là O(n * ($d_1 + d_2$)), với n = |V| và $d_1$là độ phức tạp tính toán của thủ tục Neighbor (G, u) và d_2 là độ phức tạp tính toán của thủ tục Clique (G, N).

Bước 2. Duyệt lần lượt các cặp (đỉnh, tập các đỉnh lân cận) được lấy ra từ S để tìm các lớp tương đương có độ phức tạp tính toán là O(n * k), với k là bậc của các đỉnh trên đồ thị.

Bước 3. Rút gọn h lớp tương đương nên có độ phức tạp tính toán sẽ là O(h * k), thông thường h << n. Đối với những đồ thị mạng xã hội thường là dạng đồ thị có số các đỉnh lân cận (bậc của mỗi đỉnh) d = k << m, với d và m là các hằng số, nên thuật toán REG có độ phức tạp thời gian tuyến tính (O(n)).

2.4. Thuật toán rút gọn đồ thị mạng xã hội dựa vào nguyên lý lan truyền nhãn

Trên đồ thị mạng xã hội có khá nhiều đỉnh có nhãn giống với nhãn (trong cùng một cấu trúc cộng đồng) của một trong số các đỉnh lân cận, và nhãn của chúng luôn được cập nhật lại theo những đỉnh đó suốt trong quá trình lan truyền nhãn. Những đỉnh này tương đương với nhau theo cấu trúc, luôn có cùng nhãn trong các bước lan truyền nhãn, sẽ tạo thành các lớp tương đương và do vậy, có thể kết hợp chúng với nhau thành một đỉnh đại diện duy nhất cho cả lớp đỉnh nhằm giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội ban đầu mà không ảnh hưởng đến cấu trúc của đồ thị mạng xã hội ban đầu. Chương 1 luận án đã giới thiệu một số các thuật toán phổ biến để phát hiện cấu trúc cộng đồng trên đồ thị mạng xã hội. Một trong những thuật toán hiệu quả trong lĩnh vực này là thuật toán lan truyền nhãn LPA (Label Propagation Algorithm) [85] dựa vào phương pháp học bán giám sát trên đồ thị.

2.4.1. Thuật toán lan truyền nhãn

Nguyên lý cơ bản phương pháp lan truyền nhãn là sử dụng thông tin nhãn của các đỉnh được gắn nhãn để dự đoán thông tin về nhãn của những đỉnh chưa được gắn nhãn. LPA nói chung hoạt động như sau: ban đầu mỗi đỉnh trong mạng được gán một nhãn duy nhất. Trong mỗi lần lặp, các đỉnh sẽ cập nhật nhãn của nó thành nhãn xuất

hiện thường xuyên nhất trong vùng lân cận, nếu có nhiều đỉnh có nhãn thường xuyên nhất thì chọn một cách ngẫu nhiên. Phương pháp lan truyền nhãn đề cập đến ba tính năng chính như sau:

(i) Tính năng đầu tiên là độ phức tạp thời gian gần tuyến tính. Đối với một mạng gồm n đỉnh và m cạnh, độ phức tạp thời gian của thuật toán lan truyền nhãn là $O(m+n)$.

(ii) Tính năng thứ hai là khả năng phát hiện cấu trúc cộng đồng của nó không phụ thuộc vào quy mô, độ lớn của mạng. Nó không bị ảnh hưởng bởi giới hạn độ phân giải như các phương pháp dựa trên mô đun.

(iii) Tính năng thứ ba là tính ngẫu nhiên của nó, bao gồm nhãn ban đầu ngẫu nhiên, thứ tự cập nhật nhãn ngẫu nhiên và chọn ngẫu nhiên một trong các nhãn tối đa làm nhãn của đỉnh khi nhãn tối đa không phải là duy nhất.

Ưu điểm lớn nhất của thuật toán lan truyền nhãn là thuật toán không yêu cầu bất kỳ tham số nào và thuật toán có độ phức tạp thời gian tính toán gần tuyến tính, do đó hiệu quả thực hiện của thuật toán lan truyền nhãn trong các mạng xã hội lớn là tương đối cao.

Mạng xã hội được biểu diễn bởi đồ thị vô hướng liên thông G = (V, E), V là tập các đỉnh và E là tập các cạnh. Đỉnh v liền kề (lân cận) với w nếu (v, w) ∈ E (hoặc (w,v) ∈ E). Giả sử đỉnh v có k đỉnh liền kề, ký hiệu N(v) = {v_1, v_2, ..., v_k}. Mỗi đỉnh v_j liền kề với v mang nhãn L(v_j) biểu thị cho cộng đồng mà v_j thuộc về.

Mỗi đỉnh được khởi tạo bằng một nhãn duy nhất và để các nhãn truyền qua mạng. Khi các nhãn lan truyền, các nhóm đỉnh được kết nối dày đặc nhanh chóng đạt được sự đồng thuận về một nhóm cùng nhãn duy nhất được tạo trên mạng, chúng tiếp tục mở rộng ra bên ngoài cho đến khi có thể lan truyền tiếp. Vào cuối quá trình lan truyền, các đỉnh có cùng nhãn được nhóm lại thành một cộng đồng.

Trong quá trình cập nhật lan truyền nhãn, đỉnh x tại lần lặp thứ t cập nhật nhãn của nó dựa trên nhãn của các lân cận của nó tại lần lặp thứ t - 1. Lý tưởng nhất là quá trình lặp lại sẽ tiếp tục đến khi không có đỉnh nào trong đồ thị thay đổi nhãn.

Thuật toán lan truyền nhãn LPA [85]

Input: Đồ thị mạng xã hội G = (V, E)

Output: Các cộng đồng mạng xã hội

Bước 1. Khởi tạo nhãn duy nhất cho tất cả các đỉnh trong mạng, L(i) = i, i ∈ V.

Bước 2. Đặt X là danh sách (dãy) các đỉnh được sắp xếp theo thứ tự ngẫu nhiên.

Bước 3. Với mỗi v ∈ X được chọn theo thứ tự ngẫu nhiên, cập nhật lại L(v) là nhãn của đỉnh lân cận xuất hiện thường xuyên nhất.

Bước 4. Nếu mỗi đỉnh có nhãn là số lượng tối đa mà các đỉnh lân cận của nó có, thì dừng thuật toán, chuyển sang Bước 5; Ngược lại tiếp tục thực hiện Bước 2.

Bước 5. Những đỉnh có cùng nhãn sẽ tạo thành một cộng đồng trên mạng xã hội.

Độ phức tạp của thuật toán LPA là $O(m+n)$ và đối với những đồ thị thưa là $O(n)$, với n = |V|, m = |E|; Nghĩa là độ phức tạp của thuật toán LPA gần tuyến tính.

Nhiều đề xuất phát triển, cải tiến thuật toán lan truyền nhãn bằng cách thay đổi cách gán nhãn ban đầu, cách liên kết ngẫu nhiên bị phá vỡ và liệu một đỉnh có bao gồm chính nó trong việc tính toán nhãn thường xuyên nhất trong vùng lân cận.

Năm 2010, Liu và các cộng sự [66] đề xuất cải tiến thuật toán lan truyền nhãn ELPA dựa vào độ đo đơn thể nâng cao (Advanced modularity) để phát hiện các cấu trúc cộng đồng trong mạng hiệu quả hơn bằng cách kết hợp các lợi thế tự nhiên của cấu trúc cộng đồng. Năm 2012, Wu và các cộng sự [109] sử dụng phương pháp lan truyền đa nhãn cân bằng EMLPA (Balanced multi-label propagation) để phát hiện các cấu trúc cộng đồng gối nhau. Sau đó nhiều nhóm nghiên cứu tiếp tục cải tiến thuật toán lan truyền nhãn, như Wang và các cộng sự [104] đề xuất thuật toán lan truyền nhãn lai HLPA, sử dụng một chiến lược cập nhật lai để cải thiện hiệu quả quá trình lan truyền nhãn. Năm 2014, Zhang và các cộng sự [117] đề xuất thuật toán lan truyền nhãn dự báo chuyển đổi tỷ lệ. Năm 2018, Arab và Hasheminezhad [5] đề xuất thuật toán phát hiện cộng đồng hiệu quả với việc truyền nhãn bằng cách sử dụng tầm quan trọng của các đỉnh và trọng số của các cạnh.

Đồ thị mạng xã hội thường rất phức tạp, có số đỉnh, số cạnh rất lớn, nên công việc phát hiện các cấu trúc cộng đồng trên mạng xã hội đòi hỏi rất nhiều thời gian.

Vì vậy, mặc dù thuật toán lan truyền nhãn đã có độ phức tạp thời gian tính toán gần tuyến tính, tuy nhiên gần đây vẫn có rất nhiều các nghiên cứu tiếp tục cải tiến, phát triển thuật toán lan truyền nhãn nhằm phát hiện cộng đồng nhằm đạt được hiệu quả cao hơn nữa. Tuy nhiên, hầu hết những thuật toán cải tiến, phát triển thuật toán lan truyền nhãn nêu trên chưa đề cập đến việc rút gọn đồ thị mạng xã hội theo nguyên lý lan truyền nhãn có thể giảm thiểu được đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội giúp cho việc phát hiện cộng đồng mạng xã hội nhanh, hiệu quả tốt hơn.

Phần tiếp theo trình bày tính chất tương đương của lớp đỉnh theo nguyên lý lan truyền nhãn và đề xuất thuật toán thực hiện kết hợp những đỉnh tương đương (có cùng nhãn, chung nhãn) với nhau thành một đỉnh đại diện giúp cho giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội.

2.4.2. Rút gọn đồ thị mạng xã hội dựa vào nguyên lý lan truyền nhãn

Trên đồ thị mạng xã hội thường có khá nhiều đỉnh có nhãn giống với nhãn của một trong số các đỉnh lân cận, và nhãn của chúng luôn được cập nhật lại theo những đỉnh đó suốt trong quá trình lan truyền nhãn. Những đỉnh này tương đương với nhau theo nguyên lý lan truyền nhãn, luôn có cùng nhãn trong các bước lan truyền tiếp theo và sẽ tạo thành các lớp tương đương. Do vậy, các lớp đỉnh tương đương này có thể kết hợp được với nhau thành một đỉnh đại diện nhằm giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội. Đồng thời giải quyết được vấn đề đặt ra là phát hiện cộng đồng trên các mạng xã hội có kích thước rất lớn, phát triển không ngừng theo thời gian.

2.4.2.1. Lớp các đỉnh tương đương theo nguyên lý lan truyền nhãn

Mạng xã hội được biểu diễn dưới dạng đồ thị đơn, liên thông G = (V, E), V là tập các đỉnh và E là tập các cạnh. Đỉnh v liền kề (lân cận) với w nếu (v, w) ∈ E (hoặc (w, v) ∈ E). Giả sử đỉnh v có k đỉnh liền kề, ký hiệu N(v) = $\{v_1, v_2, ..., v_k\}$. Mỗi đỉnh liền kề v_j mang nhãn $L(v_j)$ biểu thị cho cộng đồng mà v_j thuộc về.

Phương pháp lan truyền nhãn thực hiện cập nhật lại nhãn của đỉnh v theo nhãn xuất hiện thường xuyên nhất của các đỉnh liền kề. Một cách hình thức, nhãn của đỉnh v được cập nhật theo nhãn của các đỉnh u liền kề như sau [85]:

$$L(v) = \underset{l}{argmax} \sum_{u \in N(v)} [L(u) == l] \quad (2.27)$$

Trong đó: L(u) ký hiệu nhãn hiện tại của đỉnh u, $L(u) == l$ tức là nhãn xuất hiện thường xuyên nhất của đỉnh u là l.

L(v) ký hiệu nhãn mới của đỉnh v, $N(v)$ là tập các đỉnh liền kề (lân cận) của đỉnh v

Công thức 2.27 xác định L(v) nhãn xuất hiện thường xuyên nhất (cực đại) trong tập các đỉnh lân cận N(v) của đỉnh v.

Tính chất 2.13. Nếu hai đỉnh u, v ∈ V có các tập các đỉnh liền kề (lân cận) giống nhau N(u) = N(v) thì chúng có cùng nhãn, nghĩa là L(u) = L(v).

Chứng minh. Suy ra trực tiếp từ công thức (2.27) ■

Nhãn của u, v được cập nhật lại theo nhãn của cùng một đỉnh xuất hiện thường xuyên nhất. Ví dụ đỉnh w ∈ N(u) = N(v).

Hệ quả 2.1.

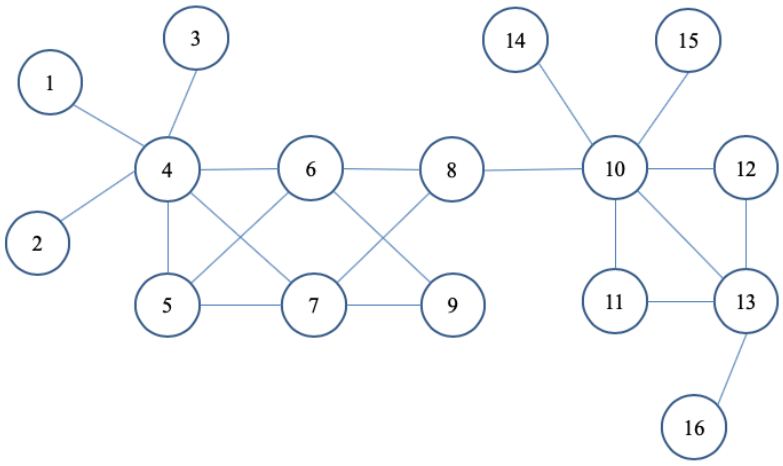
(i) Các đỉnh treo tương đương được cập nhật cùng một nhãn

(ii) Các đỉnh sườn tương đương được cập nhật cùng một nhãn

Như vậy các đỉnh treo và đỉnh sườn tương đương được trình bày ở trên sẽ có cùng nhãn với nhau.

Định nghĩa 2.6. Cho trước đồ thị vô hướng, liên thông G = (V, E). Hai đỉnh u, v ∈ V được gọi là hai đỉnh tương đồng trên G, ký hiệu là u ≈ v khi và chỉ khi N(u) = N(v).

Ví dụ 2.9. Xét đồ thị được cho như trong Hình 2.10.



Hình 2.10. Đồ thị mạng xã hội G

Dễ thấy, các đỉnh 1, đỉnh 2, đỉnh 3 tương đương với nhau, đỉnh 1 ≈ đỉnh 2 ≈ đỉnh 3 vì N(đỉnh 1) = N(đỉnh 2) = N(đỉnh 3) = {Đỉnh 4}. Tương tự như vậy, đỉnh 6 ≈ đỉnh 7 vì N(đỉnh 6) = N(đỉnh 7) = {đỉnh 4, đỉnh 5, đỉnh 8, đỉnh 9} hoặc đỉnh 14 ≈ đỉnh 15 vì N(đỉnh 14) = N(đỉnh 15) = {đỉnh 10}.

Quan hệ ≈ hiển nhiên là quan hệ tương đương, do vậy, những đỉnh tương đồng (tương đương) sẽ tạo thành các lớp tương đương. Tất cả các đỉnh trong lớp tương đương, có thể kết hợp lại thành một đỉnh đại diện để rút gọn số đỉnh và số cạnh trên đồ thị.

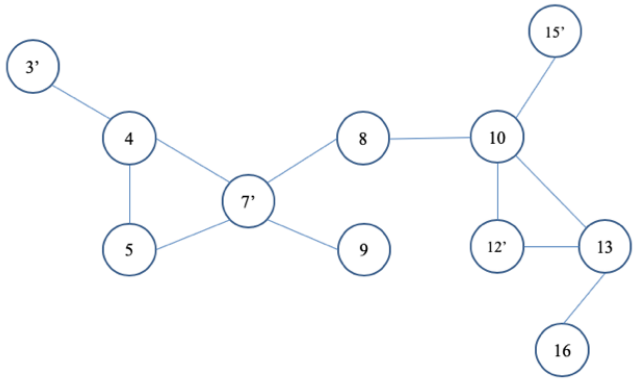
Cho trước đồ thị vô hướng, liên thông G = (V, E) và quan hệ ≈ xác định q lớp tương đương D_i, i = 1..q. Kết hợp những đỉnh tương đồng trong lớp D_i, $|D_i| > 2$, i = 1..q, thành một đỉnh đại diện D_i', để nhận được đồ thị rút gọn $G_1 = (V_1, E_1)$, trong đó:

(i) $V_1 = V - V_2 \cup \{D'_1, D'_2, \ldots, D'_q\}$, với $V_2 = D_1 \cup D_2 \cup \ldots \cup D_q$. (2.28)

(ii) $E_1 = E - \{(u, v) \mid u \in V_2, v \in N(u)\} \cup \{(v, D'_i) \mid i = 1..q, v \in N(u), \text{với } u \in D_i\}$ (2.29)

Theo nguyên lý phương pháp lan truyền nhãn, thì nhãn của các đỉnh trong mỗi lớp tương đương cũng sẽ được cập nhật lại theo nhãn của đỉnh đại diện khi quá trình lan truyền nhãn kết thúc.

Ví dụ 2.10. Kết hợp những đỉnh đồng nhất của đồ thị G trên Hình 2.10 như sau. Lớp tương đương gồm các đỉnh 1, đỉnh 2, và đỉnh 3 sẽ được đại diện bằng đỉnh 3', lớp tương đương gồm các đỉnh 6 và đỉnh 7 kết hợp với nhau và được đại diện bằng đỉnh 7'. Tương tự, lớp tương đương gồm các đỉnh 11 và đỉnh 12 được đại diện bằng đỉnh 12' và lớp tương đương gồm đỉnh 14 và đỉnh 15 được đại diện bằng đỉnh 15'. Khi đó đồ thị G ở Hình 2.10 sẽ được rút gọn như sau:



Hình 2.11. Đồ thị G_1 rút các đỉnh tương đương từ G

Đồ thị G có 16 đỉnh, 21 cạnh được rút gọn thành G_1 còn 11 đỉnh và 12 cạnh. Hiển nhiên, L(đỉnh 1) = L(đỉnh 2) = L(đỉnh 3) = L(đỉnh 3'), L(đỉnh 6) = L(đỉnh 7) = L(đỉnh 7'), L(đỉnh 14) = L(đỉnh 15) = L(đỉnh 15') và L(đỉnh 11) = L(đỉnh 12) = L(đỉnh 12'), $V_1 = V - V_2 \cup \{D'_1, D'_2, \ldots, D'_q\}$, với $V_2 = D_1 \cup D_2 \cup \ldots \cup D_q$.

2.4.2.2. Thuật toán kết hợp các đỉnh tương đồng tương đương trên đồ thị mạng xã hội

Trên cơ sở nghiên cứu các tính chất tương đương của các đỉnh theo phương pháp lan truyền nhãn, Mục này đề xuất thuật toán LREN (Label based Reduce Equivalence Nodes) thực hiện rút gọn đồ thị trên cơ sở kết hợp các đỉnh tương đồng tương đương thành đỉnh đại diện nhằm giảm thiểu không gian tính toán của đồ thị.

Thuật toán LREN (G)

Input: G = (V, E) - đồ thị ban đầu

Output: $G_1 = (V_1, E_1)$ - đồ thị thu được sau khi kết hợp các đỉnh tương đồng.

Thuật toán LREN gồm ba bước như sau:

Bước 1. Tìm tập những đỉnh đồng nhất của đồ thị mạng xã hội ban đầu

Bước 2. Tìm các lớp tương đương của các đỉnh tương đồng trên đồ thị mạng xã hội

Bước 3. Kết hợp các đỉnh tương đương thành đỉnh đại diện D'_j theo (2.28) và (2.29).

Giả mã thuật toán Label based Reduce Equivalence Nodes (LREN)

Input : G = (V, E)
Output: $G_1 = (V_1, E_1)$ - đồ thị thu được sau khi kết hợp các đỉnh tương đồng
V_D - tập những đỉnh tương đồng của đồ thị G
$V_1 = V$;
$E_1 = E$;

```
        S = ∅;              // Stack lưu các đỉnh tương đồng và tập đỉnh liền kề
// Bước 1. Tìm tập các đỉnh lân cận và lưu vào S
    for u ∈ V do { // Tìm tất cả các đỉnh tương đồng
    |   N[u] = Neighbor(G, u); // Tìm các đỉnh liền kề với u
    |   if (!Clique(N[u] && |N[u]| > 1) then { // Kiểm tra u có thể là đỉnh tương đồng
    |   |   S.push(u, N[u]);  // Lưu cặp (u, N[u]) vào S
    |   }
    }
// Bước 2. Tìm các lớp tương đương của các đỉnh tương đồng
    h = 1;
    (u, M) = S.pop();   // Lấy ra từng đỉnh và tập đỉnh liền kề của nó
    S[h] = {u}   // Lớp các đỉnh tương đồng
    N[h] = M;   // Tập các đỉnh liền kề với đỉnh tương đồng thuộc lớp S[h]
    while( S != ∅) do {
    |   (u, M) = S.pop();
    |   j = 1;
    |   loop = true;
    |   while (j <= h && loop) do {
    |   |   if(N[j] == M) then { // Kiểm tra những đỉnh sườn tương đương
    |   |   |   D[j] = D[j] ∪ {u};          // Đưa u vào lớp tương đương D[j]
    |   |   |   V1 = V1 - {u};   // Loại bỏ đỉnh u để sau đó thay bằng đỉnh đại diện
    |   |   |   for v ∈ N[j] do {
    |   |   |   |  E1 = E1 - {(u, v)}; // Loại bỏ cặp cạnh liền kề u, v
    |   |   |   }
    |   |   |   loop = false;
    |   |   }
    |   |   if (loop) then {
    |   |   |   h = h + 1;
    |   |   |   D[h] = {u};
    |   |   |   N[h] = M;
    |   |   }
    |   }
    }
    k = 0; // Chỉ xét những lớp có nhiều hơn hoặc bằng 2 đỉnh tương đồng
    for j = 1 to h do {
    |   if (|D[j]| ≥ 2) then {
    |   |   k = k + 1;
    |   |   D'[k] = D[j];
    |   |   N'[k] = N[j];
    |   }
    }

// Bước 3. Kết hợp các đỉnh tương tương đương ở lớp D[j] thành đỉnh tương đồng đại diện D'j
    VD = ∅;
    for j = 1 to k do {
    |   VD = VD ∪ {D'j};    // Bổ sung đỉnh đại diện lớp tương đương
    |   for v ∈ N'[j] do {
    |   |   E1 = E1 ∪ {(D'j, v)} // Bổ sung các cạnh liền kề của đỉnh đại diện
    |   }
    }
    V1 = V1 ∪ VD;
```

Thủ tục tính toán Neighbor(G, u): Tìm các đỉnh liền kề của u trong đồ thị G
Input: Đồ thị G = (V, E, W) và đỉnh u ∈ V
Output: N - tập các đỉnh liền của u trong đồ thị G

```
N = ∅;
for v ∈ V do {
    if ((u, v) ∈ E) then {
        N = N ∪ {v};
    }
}
return N;
```

Thủ tục tính toán Clique(G, N): Kiểm tra xem đồ thị con sinh bởi tập N trong đồ thị G có là đồ thị con đầy đủ hay không.
Input: Đồ thị G = (V, E, W) và tập đỉnh N ⊆ V
Output: **True** nếu đồ thị con sinh bởi tập đỉnh N trong đồ thị G_1 là clique, ngược lại **False**

```
for u ∈ N do {
    for v ∈ N - {u} do {
        if ((u, v) ∉ E) then {
            return false;
        }
    }
}
return true;
```

Độ phức tạp tính toán của thuật toán LREN

Thuật toán LREN (G) thực hiện qua ba bước với độ phức tạp như sau:

Bước 1. Có độ phức tạp tính toán là O(n * d), với n = |V| và d là độ phức tạp của thủ tục tính toán Neighbor(G, u), tìm các đỉnh lân cận của u.

Bước 2. Duyệt lần lượt các cặp (đỉnh, tập các đỉnh lân cận) được lấy ra từ S để tìm các lớp tương đương có độ phức tạp tính toán là O(n * k), với k là bậc của các đỉnh trên đồ thị.

Bước 3. Rút gọn h lớp tương đương nên có độ phức tạp tính toán sẽ là O(h * k), thông thường h << n.

Đối với những đồ thị mạng xã hội thường là dạng đồ thị có số các đỉnh lân cận (bậc của mỗi đỉnh) d = k << m, với d và m là hằng số, nên thuật toán LREN có độ phức tạp thời gian gần tuyến tính O(n).

2.5. Thực nghiệm và đánh giá

Việc đánh giá mức độ cải thiện về độ phức tạp và tính hiệu quả, chất lượng rút gọn của các thuật toán đề xuất được thực nghiệm trên các bộ dữ liệu mạng xã hội chuẩn [47], [60].

2.5.1. Bộ dữ liệu

Để thấy rõ hiệu quả của thuật toán đề xuất, nghiên cứu sinh thực hiện thực nghiệm trên ba bộ dữ liệu. Nhóm thứ nhất gồm các bộ dữ liệu Com-Amazon, com-Youtube và com-DBLP là các mạng xã hội lớn có trên nguồn dữ liệu đã được công bố công khai trên Stanford large network dataset collection [60].

- Co-product purchasing network (Com-Amazon) [60]: Các đỉnh đại diện cho các sản phẩm. Các đỉnh liền kề đại diện cho các sản phẩm thường được mua lại. Các cộng đồng trong mạng được xác định theo hệ thống phân loại sản phẩm. Hệ số phân cụm trung bình của mạng là 0.3967, số lượng cấu trúc cộng đồng thực là 75149 cộng đồng.
- Co-publishing network (Com-DBLP) [60]: Các đỉnh đại diện cho các tác giả. Các đỉnh liền kề đại diện cho các tác giả có ít nhất một ấn phẩm được chia sẻ. Các cộng đồng được xác định là tập hợp các tác giả đã xuất bản trong cùng một tạp chí hoặc hội nghị. Hệ số phân cụm trung bình của mạng là 0.6324, số lượng cấu trúc cộng đồng thực là 13477 cộng đồng.
- Com-Youtube [60]: Các đỉnh đại diện cho người dùng và các cạnh thể hiện mối quan hệ giữa hai người dùng. Các cộng đồng được xác định bởi tư cách thành viên trong các nhóm do người dùng tạo ra. Hệ số phân cụm trung bình của mạng là 0.0808, số lượng cấu trúc cộng đồng thực là 8385 cộng đồng.

Như vậy qua phân tích ba bộ dữ liệu mạng Com-DBLP, com-Amazon và com-Youtube ta thấy điểm khác biệt giữa các bộ dữ liệu như sau:

Thứ nhất hệ số phân cụm của mạng Com-DBLP có giá trị lớn nhất 0.6324, sau đó đến mạng Com-Amazon là 0.3967 và cuối cùng nhỏ nhất mạng com-Youtube là 0.0808. Hệ số phân cụm khác nhau phản ánh cấu trúc các mạng khác nhau.

Thứ hai xem xét tỉ lệ số cạnh - số đỉnh trên cả ba mạng là khá tương đồng nhau, đối với mạng com-Amazon là 2.76, mạng com-DBLP là 3.31 và mạng com-Youtube là 2.63.

Thứ ba quy mô cộng đồng trung bình trong mạng com-DBLP là 53.41 lớn hơn so với hai mạng còn lại, gấp 2.75 lần mạng com-Amazon là 19.38 và gấp 3.96 lần mạng com-Youtube là 13.50

Thứ tư cộng đồng trong mỗi bộ dữ liệu là khác nhau. Cộng đồng ở mạng com-Amazon là danh mục các sản phẩm, đối với mạng com-DBLP là nơi xuất bản, và mạng com-Youtube là một nhóm người dùng.

Cuối cùng, quy mô thành viên trung bình trong mạng khác nhau. Trung bình, một sản phẩm của com-Amazon được phân vào 8,74 danh mục sản phẩm khác nhau; một tác giả trong com-DBLP xuất bản tới 1,69 địa điểm và chỉ có một trong số mười người dùng com-YouTube tham gia nhóm sở thích.

Như vậy ba bộ dữ liệu được đưa vào thực nghiệm là các bộ dữ liệu có tính đa dạng từ cấu trúc mạng đến tính chất của mạng, đảm bảo được yếu tố khách quan trong thực nghiệm.

Bảng 2.2. Bảng các bộ dữ liệu thuộc nhóm thứ nhất

Stt	Bộ dữ liệu thực nghiệm	Số đỉnh của đồ thị mạng xã hội	Số cạnh của đồ thị mạng xã hội	Số lượng cộng đồng thực tế công bố
1	Com-DBLP	317080	1049866	13477
2	Com-Amazon	334863	925872	75149
3	Com-Youtube	1134890	2987624	8385

2.5.2. Cài đặt thực nghiệm

Kịch bản thực nghiệm

Các thuật toán thực nghiệm được luận án thực hiện riêng lẻ với từng bộ dữ liệu và lúc này trên máy tính chỉ thực hiện duy nhất một chương trình.

Môi trường thực nghiệm là máy tính PC với cấu hình Intel™ Core™ i7-9700CPU @4.70 GHz, 8 GB RAM, sử dụng hệ điều hành Windows 10. Công cụ lập trình thực hiện thuật toán là ngôn ngữ lập trình Python.

2.5.3. Kết quả thực nghiệm

Số lượng đỉnh và số lượng cạnh của đồ thị rút gọn sau khi thực hiện thuật toán REG được thể hiện trong Bảng 2.3.

Bảng 2.3. Số lượng đỉnh và cạnh của đồ thị mạng xã hội rút gọn bởi REG

Stt	Bộ dữ liệu thực nghiệm	Số lượng đỉnh của đồ thị ban đầu	Số lượng cạnh của đồ thị ban đầu	Số lượng đỉnh của đồ thị rút gọn	Số lượng cạnh của đồ thị rút gọn
1	Com-Amazon	334863	925872	300271	703533
2	Com-DBLP	317080	1049866	271484	773226
3	Com-Youtube	1134890	2987624	852189	2095221

Qua số liệu tại Bảng 2.3 thì số lượng đỉnh và cạnh được giảm sau khi thực hiện rút gọn đồ thị REG là khá lớn và lần lượt là 34592 đỉnh và 222339 cạnh đối với mạng Com-DBLP, 45596 đỉnh và 276640 cạnh đối với mạng Com-Amazon, 282701 đỉnh và 892403 cạnh đối với mạng Com-Youtube. Như vậy kết quả thực nghiệm cho thấy rằng kích thước của mạng xã hội càng lớn và cấu trúc của mạng xã hội xuất hiện nhiều các đỉnh tương đương theo độ đo trung tâm trung gian sẽ quyết định đến số lượng đỉnh và cạnh giảm được.

Bảng 2.4. Tỷ lệ rút gọn đồ thị bởi REG

Stt	Bộ dữ liệu thực nghiệm	Số lượng cạnh của đồ thị ban đầu	Số lượng cạnh của đồ thị rút gọn	Tỉ lệ rút gọn đồ thị
1	Com-Amazon	925872	703533	0.240
2	Com-DBLP	1049866	773226	0.264
3	Com-Youtube	2987624	2095221	0.299

Qua số liệu ở Bảng 2.4 cho thấy tỉ lệ rút gọn đồ thị mạng xã hội là khá lớn lần lượt là 0.240, 0.264 và 0.299 đối với các mạng Com-DBLP, Com-Amazon và Com-

Youtube. Như vậy, một nhận xét quan trọng đã được khẳng định rằng hiệu suất rút gọn đồ thị tăng khi quy mô của mạng xã hội tăng lên, đồng thời giá trị số lượng đỉnh và cạnh rút gọn được có ý nghĩa quan trọng đối với bài toán phát hiện cộng đồng trên mạng xã hội. Tiếp theo, số lượng đỉnh và cạnh của đồ thị rút gọn sau khi thực hiện thuật toán LREN được thể hiện trong Bảng 2.5.

Bảng 2.5. Số lượng đỉnh và cạnh của đồ thị mạng xã hội rút gọn bởi LREN

Stt	Bộ dữ liệu thực nghiệm	Số lượng đỉnh của đồ thị ban đầu	Số lượng cạnh của đồ thị ban đầu	Số lượng đỉnh của đồ thị rút gọn	Số lượng cạnh của đồ thị rút gọn
1	Com- Amazon	334863	925872	301892	704251
2	Com-DBLP	317080	1049866	272994	775148
3	Com-Youtube	1134890	2987624	853874	2116447

Qua số liệu tại Bảng 2.5 cho thấy số lượng đỉnh và cạnh được giảm sau khi thực hiện thuật toán rút gọn đồ thị theo nguyên lý lan truyền nhãn LREN là khá lớn và các giá trị lần lượt là 32971 đỉnh và 221621 cạnh đối với mạng Com-DBLP, 44086 đỉnh và 274718 cạnh đối với mạng Com-Amazon, 281016 đỉnh và 871177 cạnh đối với mạng Com-Youtube.

Bảng 2.6. Tỷ lệ rút gọn đồ thị bởi LREN

Stt	Bộ dữ liệu thực nghiệm	Số lượng cạnh của đồ thị ban đầu	Số lượng cạnh của đồ thị rút gọn	Tỉ lệ rút gọn đồ thị
1	Com-Amazon	925872	704251	0.239
2	Com-DBLP	1049866	775148	0.262
3	Com-Youtube	2987624	2116447	0.292

Qua số liệu tại Bảng 2.6 cho thấy tỉ lệ rút gọn đồ thị mạng xã hội của thuật toán LREN là khá lớn và lần lượt là 0.239, 0.262 và 0.292 đối với các mạng Com-DBLP, Com-Amazon và Com-Youtube. Như vậy kết quả thực nghiệm tại các bảng 2.3, bảng 2.4, bảng 2.5 và bảng 2.6 khẳng định rằng kích thước của mạng xã hội càng lớn và

cấu trúc của mạng xã hội xuất hiện nhiều các đỉnh tương đương theo độ đo trung tâm trung gian và nguyên lý lan truyền nhãn quyết định đến số lượng đỉnh và cạnh giảm được.

2.6. Kết luận chương 2

Chương 2 trình bày các tính chất của các đỉnh tương đương theo độ đo trung tâm trung gian và phương pháp kết hợp các lớp đỉnh tương đương có cùng độ đo trung tâm trung gian để rút gọn đồ thị mạng xã hội. Đồng thời trong chương 2 cũng trình bày phương pháp kết hợp các lớp đỉnh tương đương theo nguyên lý lan truyền nhãn để rút gọn đồ thị mạng xã hội. Chương này trình bày các kết quả chính như sau:

- Đề xuất thuật toán REG thực hiện rút gọn đồ thị mạng xã hội ban đầu dựa vào các lớp đỉnh tương đương theo độ đo trung tâm trung gian nhưng vẫn bảo toàn giá trị độ đo trung tâm trung gian của đồ thị. Kết quả này được công bố trong công trình [CT1], [CT4].
- Đề xuất thuật toán rút gọn đồ thị LREN thực hiện kết hợp những đỉnh tương đương với nhau theo tiêu chí lan truyền nhãn thành đỉnh đại diện nhằm giảm thiểu số đỉnh, cạnh của đồ thị khá nhiều và qua đó giảm độ phức tạp tính toán của các thuật toán phát hiện cấu trúc cộng đồng trên mạng xã hội. Kết quả này được công bố trong công trình [CT3].

- Đồng thời trong chương này cũng tiến hành thực nghiệm các thuật toán trên các bộ dữ liệu thực nghiệm từ kho dữ liệu mạng xã hội lớn nhằm đánh giá tính hiệu quả của các thuật toán đề xuất.

- Thuật toán rút gọn đồ thị mạng xã hội dựa trên các lớp đỉnh tương đương về độ đo trung tâm trung gian đã cải tiến hiệu quả thời gian tính toán độ đo trung tâm trung gian và bảo toàn được giá trị độ đo trung tâm trung gian trên đồ thị sau khi rút gọn. Thuật toán rút gọn đồ thị mạng xã hội dựa theo nguyên lý lan truyền nhãn đã chứng minh được hiệu quả. Chính vì vậy, luận án tiếp tục đề xuất áp dụng thuật toán rút gọn đồ thị mạng xã hội để cải tiến thuật toán phát hiện cộng đồng trên mạng xã hội nhanh, hiệu quả cao hơn.

CHƯƠNG 3. ÁP DỤNG THUẬT TOÁN RÚT GỌN ĐỒ THỊ ĐỂ PHÁT HIỆN CỘNG ĐỒNG TRÊN MẠNG XÃ HỘI

3.1. Giới thiệu

Do tính chất của mạng xã hội có cấu trúc khá tự do và kích thước lớn, không ngừng phát triển theo thời gian, vì vậy hầu hết các thuật toán phát hiện cộng đồng truyền thống mất rất nhiều thời gian và chưa thực sự hiệu quả. Một trong những cách tiếp cận để khắc phục được hạn chế trên là sử dụng phương pháp rút gọn đồ thị mạng xã hội và vẫn bảo toàn được các tính chất của cộng đồng sau khi rút gọn được đề xuất tại chương 2 nhằm mục đích giảm thiểu thời gian tính toán. Chương 3 của luận án trình bày (i) Đề xuất thuật toán cải tiến thời gian tính độ đo trung tâm trung gian trên đồ thị mạng xã hội, (ii) Đề xuất phát triển thuật toán phát hiện cộng đồng mạng xã hội trên đồ thị rút gọn dựa vào độ đo trung tâm trung gian, (iii) Đề xuất phát triển thuật toán lan truyền nhãn trên đồ thị rút gọn để phát hiện cộng đồng nhanh, hiệu quả mà không yêu cầu tối ưu hóa hàm mục tiêu cũng như thông tin dự đoán về các cộng đồng. Điều này hoàn toàn phù hợp với tính chất của mạng xã hội là hầu hết không thể dự đoán trước được số lượng cộng đồng đang tồn tại và cộng đồng thì thường xuyên thay đổi theo thời gian. Kết quả thực nghiệm trên các bộ dữ liệu mẫu khẳng định tính hiệu quả của thuật toán đề xuất, thời gian thực hiện của thuật toán đề xuất giảm thiểu đáng kể so với các thuật toán đã công bố trước đó. Kết quả nghiên cứu ở chương này được công bố trong công trình số [CT2], [CT3].

Xuất phát từ ý tưởng của phương pháp phát hiện cộng đồng dựa vào độ đo trung tâm trung gian, nghiên cứu sinh nhận thấy trên đồ thị mạng xã hội có khá nhiều đỉnh tương đương với nhau theo cấu trúc có cùng độ đo trung tâm trung gian, chúng tạo thành các lớp tương đương và có thể kết hợp chúng lại với nhau thành một đỉnh đại diện duy nhất cho cả lớp đỉnh. Nhờ vậy có thể giảm thiểu được đáng kể số đỉnh và cạnh của đồ thị mạng xã hội ban đầu, tiết kiệm được chi phí tính toán mà lại không ảnh hưởng đến cấu trúc của đồ thị mạng xã hội ban đầu. Vì vậy nghiên cứu sinh đề xuất áp dụng thuật toán rút gọn đồ thị mạng xã hội dựa vào độ đo trung tâm trung

gian để cải tiến thời gian tính toán của thuật toán tính độ đo trung tâm trung gian đồng thời cải tiến nhóm thuật toán phát hiện cộng đồng mạng xã hội dựa vào độ đo trung tâm trung gian nhanh và hiệu quả hơn.

Phát hiện cộng đồng trên mạng xã hội là một nhiệm vụ quan trọng hàng đầu trong phân tích mạng xã hội. Để giải quyết nhiệm vụ này, nhiều thuật toán phát hiện cộng đồng trên mạng xã hội đã được đề xuất. Tuy nhiên, các thuật toán này hầu hết chưa đạt hiệu quả trong việc phát hiện cộng đồng trên các mạng xã hội quy mô rất lớn do độ phức tạp về thời gian và không gian tính toán. Để giải quyết thách thức đặt ra, chương 2 luận án đã đề xuất các thuật toán rút gọn đồ thị dựa vào các lớp đỉnh tương đương về độ đo trung tâm trung gian và theo nguyên lý lan truyền nhãn. Đồng thời, tiến hành thực nghiệm trên các mạng xã hội khác nhau cùng với độ đo tỷ lệ rút gọn đồ thị đã chứng minh tính ưu việt và hiệu quả của thuật toán rút gọn đồ thị đề xuất. Chương 3 luận án tiếp tục áp dụng các thuật toán rút gọn đồ thị mạng xã hội đã được đề xuất ở chương 2 để thực hiện phát hiện các cộng đồng mạng xã hội hiệu quả. Đồng thời trong chương 3 luận án cũng thực hiện các thực nghiệm trên các mạng xã hội khác nhau và sử dụng các độ đo để đánh giá tính ưu việt và hiệu quả của các thuật toán phát hiện cộng đồng trên mạng xã hội đề xuất.

3.2. Thuật toán tính nhanh độ đo trung tâm trung gian trên đồ thị mạng xã hội rút gọn

Theo Jamour và cộng sự (năm 2018) [48], Nakajima và cộng sự (năm 2020) [74] đều khẳng định rằng thuật toán tính độ đo trung tâm trung gian của Brandes [18], [19] đề xuất đến nay vẫn là một trong những thuật toán nhanh nhất để tính toán độ đo trung tâm trung gian trên đồ thị kích thước lớn. Vì vậy việc phát triển dựa vào thuật toán gốc Brandes vẫn đạt được những hiệu quả nhất định. Dựa vào phương pháp tính độ đo trung tâm trung gian C_B theo kỹ thuật tích lũy phụ thuộc của Brandes [18], [19] và sử dụng các định nghĩa, tính chất nêu ở Mục 2.1 của chương 2, luận án đề xuất thuật toán FBC tính nhanh độ đo trung tâm trung gian C_B của các cạnh trên đồ thị mạng xã hội.

3.2.1. Duyệt đồ thị theo chiều rộng

Để thực hiện thuật toán tính độ đo trung tâm trung gian của các đỉnh trên đồ thị một cách hiệu quả, người ta thường sử dụng phương pháp duyệt theo chiều rộng BFS (Breadth-First Search) [55]. Thuật toán duyệt theo chiều rộng tìm kiếm các đường đi ngắn nhất từ đỉnh gốc qua các cạnh tới tất cả các đỉnh khác trong đồ thị. Các cạnh giữa các mức của quá trình duyệt BFS bắt đầu từ đỉnh gốc X sẽ tạo thành đồ thị định hướng, phi chu trình, được gọi DAG_X. Với mỗi đỉnh v trên DAG_x, khoảng cách của đỉnh v đến gốc x, được ký hiệu là d_v, là số các cạnh trên đường đường đi ngắn nhất từ x đến v.

Thuật toán BFS

Input: $G = (V, E), x \in V$

Output: Đồ thị duyệt theo chiều rộng BFS: DAG_x

Bước 1. Duyệt theo chiều rộng BFS bắt đầu từ đỉnh $x \in V$.

Bước 2. Gắn nhãn cho từng đỉnh bằng số các đường đi ngắn nhất đi từ gốc x tới chúng. Đỉnh gốc bắt đầu được gắn nhãn bằng 1. Sau đó thực hiện từ trên xuống, nhãn của các đỉnh ở mức tiếp theo bằng tổng số nhãn của các đỉnh cha của chúng.

Bước 3. Tính số các đường đi ngắn nhất từ gốc trên DAG_x

- Bắt đầu từ đỉnh gốc x với khoảng cách cho trước $d_x = 0$;
- Đối với mỗi đỉnh i liền kề với x (với $(x, i) \in E$) ta thực hiện tính khoảng cách $d_i = d_x + 1$.
- Lặp lại bước 3 cho đến khi không còn lại đỉnh nào là không được gắn nhãn.

3.2.2. Thuật toán tính nhanh độ đo trung tâm trung gian

Nghiên cứu sinh đề xuất thuật toán tính nhanh độ đo trung tâm trung gian **FBC** (Fast algorithm for Betweenness Centrality) trên đồ thị mạng xã hội. Ý tưởng của thuật toán đề xuất là thay vì thực hiện tính toán độ đo trung tâm trung gian trên đồ thị mạng xã hội ban đầu như thuật toán gốc Brandes [19], thuật toán đề xuất FBC thực hiện rút gọn đồ thị mạng xã hội ban đầu nhằm giảm thiểu không gian tính toán nhưng vẫn bảo toàn được giá trị độ đo trung tâm trung gian và thực hiện tính toán độ đo trung tâm trung gian trên đồ thị mạng xã hội rút gọn.

Thuật toán FBC (Fast algorithm for Betweenness Centrality)

Input: Đồ thị mạng xã hội G = (V, E)

Output: Độ đo trung tâm trung gian của các cạnh trên đồ thị mạng xã hội.

Thuật toán đề xuất FBC bao gồm bốn bước như sau:

Bước 1. Thực hiện thuật toán REG (G) đã nêu ở Mục 2.3 thực hiện rút gọn các lớp đỉnh treo và đỉnh sườn tương đương về độ đo trung tâm trung gian, chuyển đồ thị mạng xã hội ban đầu G = (V, E) về đồ thị mạng xã hội rút gọn $G_1 = (V_1, E_1)$.

Bước 2. Khởi tạo các giá trị cho mảng $C_B[e] = 0$, $e \in E_2$, stack $S = \varnothing$, queue $Q = \varnothing$, bốn mảng bổ sung Pr_x, Po_x, δ và d. Mảng δ xác định tỷ số đường đi ngắn nhất từ gốc x tới mỗi đỉnh trên DAG_x, mảng d đo khoảng cách của mỗi đỉnh từ gốc x. Ban đầu, khoảng cách của các đỉnh và gốc đều gán bằng -1. Mảng Pr_x, là danh sách các đỉnh cha liên kết với mỗi đỉnh v, $Po_s[v]$ chứa những đỉnh con ở dưới v trong lần duyệt theo chiều rộng BFS từ gốc x. V_C là tập các đỉnh treo, V_S là tập các đỉnh sườn của G_1.

Bước 3. Duyệt theo chiều rộng BFS từ gốc x để tìm những đường đi ngắn nhất tới tất cả các đỉnh khác. Trong bước này, mỗi phần tử được đặt vào một hàng đợi khi nó được tìm thấy. Khi duyệt theo chiều rộng, khoảng cách từ gốc x tới từng đỉnh v được tính. Với mỗi đỉnh v được tìm thấy trong lần duyệt BFS sẽ tương ứng với hai danh sách các đỉnh cha, đỉnh con liền kề v và δ[t] là số đường đi ngắn nhất đi từ x đến t.

Bước 4. Tính độ đo trung tâm trung gian C_B của các cạnh theo kỹ thuật tích lũy của Brandes. Với mỗi DAG_x, $x \in V2$, tính độ đo trung tâm trung gian của các cạnh trên DAG_x, sau đó cộng dồn vào những cạnh đã được tính trên những DAG đã được duyệt trước đó cho độ đo trung tâm trung gian của các cạnh trên toàn đồ thị mạng xã hội.

Giả mã thuật toán Fast Algorithm for Betweenness Centrality (FBC)

```
Input:   + G = (V, E) đồ thị mạng xã hội
Output: + C_B(e), e ∈ E
// Bước 1.
        G1 = RED (G);  // Kết hợp các đỉnh treo, sườn tương đương, G1 = (V1, E1)
        V_C, V_S - tập các đỉnh treo, đỉnh sườn của G1
// Bước 2. Khởi tạo các giá trị
          for all e ∈ E1  do {
            |   C_B[e] = 0; // Khởi  tạo giá trị độ đo trung tâm trung gian ban đầu là 0
           }
          V2 = V1 - V_C;                  // Tập các đỉnh không phải là đỉnh treo
```

```
for x ∈ V2 do { / /Duyệt BFS bắt đầu từ đỉnh x – không phải đỉnh treo
    S = ∅;  //empty stack;
    Q = ∅;  //empty queue;
    for w ∈ V1 do {
        Prx[w] = ∅; // Danh sách các tiền tố (đỉnh cha) liền kề với w trên DAGx
        Pox[w] = ∅; // Danh sách các hậu tố (đỉnh con) liền kề với w trên DAGx
        δ[w] = 0;     // Trọng số của đỉnh w
        d[w] = -1;   // Khoảng cách từ x đến w, ban đầu là -1 (chưa được duyệt)
    }
// Bước 3. Duyệt BFS từ gốc x để tìm những đường đi ngắn nhất tới tất cả các đỉnh khác
    δ[x] = 1;         // Số đường đi ngắn nhất từ x
    d[x] = 0;         // Khoảng cách đường đi ngắn nhất đi từ x
    Q.enqueue(x);                    //Nạp x vào hàng đợi Q
    while (Q != ∅) do{        // Duyệt theo chiều rộng BFS
        v  =  Q.removefront();  // Lấy ra phần tử đầu hàng đợi Q và gán cho v
        S. push(v);               // Đưa v vào S (stack)
        for w ∈ N[v] do{              // Với mọi đỉnh w liền kề với v, ban đầu là x
            if (d[w] < 0) then {        // Khi đỉnh w chưa được duyệt - gặp lần đầu
                Q.enqueue(w);   // Đưa w vào hàng đợi Q
                d[w] = d[v] + 1; // Tăng khoảng cách đường đi ngắn nhất lên 1
                δ[w] = δ[v];
            }
            // Khi đường đi ngắn nhất từ x tới w đi qua v trên đồ thị DAGx
            if (d[w] == d[v] + 1) then { // Khi w là con của v trên DAGx
                δ [w] = δ [w] + δ [v];
                Prx[w].append(v);     // Bổ sung v vào danh sách Prx[w]
                Pox[v].append(w);     // Bổ sung v vào danh sách Pox[w]
            }
        }
// Bước 4. Tính tích lũy độ đo trung tâm trung gian CB của các cạnh trên DAGx
    while (S != ∅) do { // S chứa các đỉnh theo thứ tự không giảm theo khoảng cách từ x
        w = S.pop();             // Lấy phần tử ở cuối Stack S và gán cho w
        if (Pox[w] == ∅) then { // w là là
            v = Prx[w] ∩ N [w]; // v là cha của w
            e = (v, w); // e ∈ E1
            if (w ∈ VC) then {  // Nếu w là lá và là đỉnh treo
                τ[e] = W1(w); // Tính độ đo trung tâm trung gian của cạnh e trên DAGx
            }else  if (w ∈ VS ) then { // Nếu w là đỉnh sườn và là lá
                τ[e]=δ [v]/δ [w]* W1(w) + 1/|N[w]|;
            } else { // w lá và không phải đỉnh treo, không phải đỉnh sườn
                τ[e] = δ [v] / δ [w];   Tính độ đo trung tâm trung gian cạnh e trên DAGx
            }
            CB[e] = CB[e] + τ[e];// Cộng dồn độ đo trung tâm trung gian cạnh e
        } // if(Pox[w] == ∅)
        else {// Khi w không phải là lá trên DAGx
            m = 0;
            for all k ∈ Pox[w] do{
                e = (w, k);              // e ∈ E2 và τ[e] đã được tính ở bước trước
                m = m + τ[e]; // Tổng độ đo trung tâm trung gian các cạnh liên thuộc w
            }
```

```
        for all v ∈ Prx[w] do {
            e = (v, w);  // e ∈ E2
            if (v == x && v ∈ VS) then { // Khi gốc x là đỉnh sườn
                τ[e] = ((m + 1) * δ [v] / δ [w] + 1 / |N(x)|) * W2(x);
            }else{
                τ[e] = (m + 1) * δ [v] / δ [w];
            }
            CB(e) = CB(e) + τ[e];
        }
    } // else Khi w không phải là lá
  }   // while (S != ∅) – khi w chưa phải là gốc
}     // for x ∈ V2
```

Độ phức tạp của thuật toán FBC

Kích thước bộ nhớ của stack, queue và các mảng σ và d là O(|V2|), nghĩa là cỡ của các cấu trúc bổ sung được giới hạn bằng số đỉnh V2 của đồ thị. Bộ nhớ cần thiết cho mảng liên kết được giới hạn bởi số cạnh E_2, đó là O(|E_1|). Bởi mỗi lần duyệt BFS được tính độc lập, chỉ cần duy trì một bản copy của những cấu trúc này. Độ phức tạp tính toán của việc duyệt cây BFS là O(|V2| + |E_1|) và của việc tích lũy (accumulation) cũng vào khoảng O(|V2| + |E_1|), số các bước cực đại được xác định bởi số đỉnh cha là O(|E_1|), và số đỉnh con tương ứng là O(|V2|). Vậy, độ phức tạp của thuật toán sẽ là O(|V2|^2 +|V2| * |E_1|). Trong trường hợp |E_1| > |V2|, thì độ phức tạp của thuật toán sẽ là O(|V2| * |E_1|). Thuật toán của Brandes [19] có độ phức tạp là O(|V|^2 +|V| * |E|), do vậy thuật toán cải thiện nhanh hơn, hiệu quả, bởi vì thông thường thì |V2| < |V| và |E_1| < |E|.

3.3. Thuật toán phát hiện cộng đồng trên đồ thị rút gọn dựa vào độ đo trung tâm trung gian

Thuật toán phát hiện cấu trúc cộng đồng trên đồ thị dựa vào độ đo trung tâm trung gian điển hình và phổ biến nhất chính là thuật toán GN (Girvan-Newman) [76]. Dựa vào ý tưởng của Girvan-Newman, nghiên cứu sinh đề xuất phát triển thuật toán CDAB (Community Detection Algorithm based on Betweenness) phát hiện cấu trúc cộng đồng trên đồ thị rút gọn dựa vào độ đo trung tâm trung gian. Xuất phát từ độ phức tạp thời gian tính toán của thuật toán GN trên đồ thị mạng xã hội rất lớn, nghiên

cứu sinh thực hiện đề xuất cải tiến thời gian tính toán bằng cách giảm thiểu thời gian tính toán độ đo trung tâm trung gian của các cạnh trên đồ thị.

Thuật toán đề xuất CDAB gồm các bước như sau:

Input: Đồ thị mạng xã hội G = (V, E)

Output: Tập các cộng đồng mạng xã hội.

Bước 1. Đề xuất thực hiện tính độ đo trung tâm trung gian của tất cả các cạnh trong mạng theo thuật toán tính nhanh độ đo trung tâm trung gian FBC đề xuất ở Mục 3.2.

Bước 2. Tìm những cạnh có độ đo trung tâm trung gian lớn nhất và loại bỏ chúng,

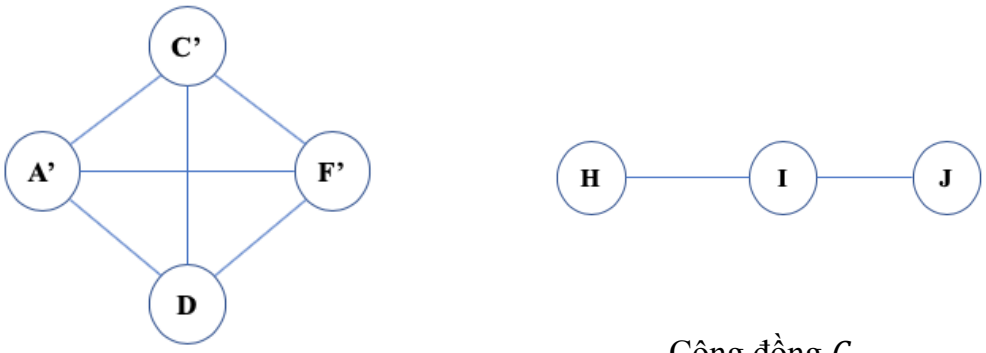
Bước 3. Đề xuất thực hiện tính lại độ đo trung tâm trung gian của tất cả các cạnh trong các thành phần còn lại của mạng theo thuật toán tính nhanh độ đo trung tâm trung gian FBC đã được đề xuất ở Mục 3.2.

Bước 4. Lặp lại từ bước 2 cho đến khi đến khi không có cạnh nào vượt qua ngưỡng của độ đo trung tâm trung gian cho trước hoặc không còn cạnh trung gian.

Như vậy thuật toán đề xuất CDAB thực hiện cải tiến so với thuật toán gốc GN ở Bước 1 và Bước 3 khi sử dụng thuật toán tính nhanh độ đo trung tâm trung gian FBC nhằm giảm thiểu thời gian tính toán của thuật toán phát hiện cộng đồng mạng xã hội.

Ví dụ 3.1. Minh hoạ thuật toán CDAB

- Cho đồ thị mạng xã hội Kite ban đầu như Hình 2.8
- Thực hiện thuật toán REG rút gọn các đỉnh tương đương theo độ đo trung tâm trung gian ta thu được đồ thị mạng xã hội Kite rút gọn như Hình 2.9
- Thực hiện thuật toán GN trên đồ thị mạng xã hội Kite rút gọn ta thu được hai cộng đồng như Hình 3.1



Cộng đồng C_1
gồm tập các đỉnh {D, A', C', F'}

Cộng đồng C_2
gồm tập các đỉnh {H, I, J}

Hình 3.1. Các cộng đồng của đồ thị mạng xã hội Kite

Trong đó A' = {A, B}, C' = {C, E}, F' = {F, G}. Vậy các cộng đồng của đồ thị ban đầu ở Hình 2.8 là $C_1 = \{D, A, B, C, D, E, F, G\}$ và $C_2 = \{H, I, J\}$.

Độ phức tạp thời gian tính toán của thuật toán CDAB

Đối với đồ thị liên thông, vô hướng và không trọng số G = (V, E) với $m = |E|, n = |V|$. Đồ thị ban đầu G = (V, E) sau khi rút gọn đồ thị là $G_1 = (V_1, E_1)$ với $m_1 = |E|, n_1 = |V|$ trong đó $m_1 < m$, $n_1 < n$. Độ phức tạp của thuật toán CDAB là $O({m_1}^2 n_1)$ và đối với trường hợp đồ thị thưa là $O({n_1}^3)$.

Độ phức tạp của thuật toán Girvan - Newman là m cạnh cần loại bỏ với mỗi bước lặp có độ phức tạp O(mn) cần thời gian là $O(m^2 n)$ và đối với trường hợp đồ thị thưa là $O(n^3)$. Vì số đỉnh của đồ thị lớn thường nhỏ hơn số cạnh rất nhiều, nghĩa là n << m, nên độ phức tạp của thuật toán CDAB là nhỏ hơn so với thuật toán GN.

Xuất phát từ ý tưởng nhóm thuật toán phát hiện cộng đồng dựa vào nguyên lý lan truyền nhãn, nghiên cứu sinh nhận thấy trên đồ thị mạng xã hội có khá nhiều đỉnh có nhãn giống với nhãn (trong cùng một cộng đồng) của một trong số các đỉnh lân cận, và nhãn của chúng luôn được cập nhật lại theo những đỉnh đó suốt trong quá trình lan truyền nhãn. Những đỉnh này tương đương với nhau theo cấu trúc, luôn có cùng nhãn trong các bước lan truyền nhãn, sẽ tạo thành các lớp tương đương và do vậy, có thể kết hợp chúng với nhau thành một đỉnh đại diện duy nhất cho cả lớp đỉnh nhằm giảm thiểu đáng kể số đỉnh và số cạnh của đồ thị mạng xã hội ban đầu mà không

ảnh hưởng đến cấu trúc của đồ thị mạng xã hội ban đầu. Vì vậy, nghiên cứu sinh đề xuất thuật toán rút gọn đồ thị mạng xã hội dựa vào nguyên lý lan truyền nhãn và áp dụng vào nhóm thuật toán phát hiện cộng đồng dựa vào lan truyền nhãn để phát hiện cộng đồng mạng xã hội nhanh và hiệu quả hơn.

3.4. Thuật toán lan truyền nhãn phát hiện cộng đồng trên đồ thị mạng xã hội rút gọn

Theo phương pháp lan truyền nhãn, thì nhãn của các đỉnh trong mỗi lớp tương đương cũng sẽ được cập nhật lại theo nhãn của đỉnh đại diện khi quá trình lan truyền nhãn kết thúc. Nghiên cứu sinh đề xuất thuật toán **LPAA (Label Propagation Algorithm on Abridged graph)** lan truyền nhãn phát triển trên đồ thị rút gọn.

Input: Đồ thị vô hướng, liên thông G = (V, E)

Output: Các cấu trúc cộng đồng trên đồ thị mạng xã hội

Thuật toán thực hiện qua 2 bước:

Bước 1. Sử dụng thuật toán đề xuất LREN (G) thực hiện tìm các đỉnh đồng nhất tương đương của đồ thị G = (V, E) và rút gọn các đỉnh tương đương thành đồ thị $G_1 = (V_1, E_1)$.

Bước 2. Thực hiện thuật toán lan truyền nhãn trên đồ thị rút gọn G_1 để phát hiện những đỉnh có cùng nhãn tạo thành các cấu trúc cộng đồng mạng xã hội.

Thuật toán lan truyền nhãn phát triển trên đồ thị rút gọn thực hiện lặp lại qua nhiều bước. Mỗi bước lặp nhãn của các đỉnh trên đồ thị sẽ được cập nhật lại theo nhãn của đỉnh lân cận xuất hiện thường xuyên nhất theo công thức tính (2.28) và (2.29).

Điều kiện dừng của thuật toán: kiểm tra xem nhãn của các đỉnh ở bước hiện tại so với nhãn của các đỉnh ở bước trước, nếu không có thay đổi nhãn xảy ra thì thuật toán dừng (bước tiếp theo sẽ không có sự thay đổi bất kỳ nhãn nào).

Giả mã thuật toán LPAA

```
Input: Đồ thị vô hướng, liên thông G = (V, E)
Output: Các cộng đồng trên mạng xã hội
LREN(G);    //Kết quả là đồ thị rút gọn G1 = (V1, E1)
i = 0; //Lần lặp thứ i
for     v ∈ V1 do //Khởi tạo nhãn của G1
|       L(i, v) = v;
while (Chưa thỏa Điều kiện dừng thuật toán) do {
```

```
        i = i + 1;        //Lan truyền nhãn ở bước tiếp theo
        for v ∈ V1 do{
            N(v) = Neighbor(G, v); //Tìm các nút liền kề với v
            //Xếp các nút lân cận của v theo thứ tự ngẫu nhiên thành mảng A[]
            A[k] = N(v)[k]; // k = |N(v)|
                //Đếm số lần xuất hiện thường xuyên của nhãn trong A[]
            j = 0; max = 1; //max số lần xuất hiện cực đại
            for (j = 1; j < k; j++){
                dem = 0;
                //Đếm số lần xuất hiện thường xuyên của nhãn A[j]
                for (l = 0; l < k; l++)
                    if (L(i-1, A[l]) == L(i-1, A[j]))
                        dem = dem +1;
                if (dem > max)
                    max = dem;
            }
            //Gắn nhãn v theo nút trong N(v) xuất hiện thường xuyên nhất là A[j]
            L(i, v) = L(i-1, A[j]);
        }//Thỏa mãn <Điều kiện dừng lan truyền nhãn>
} // return L(i, v) với mọi nút v ∈ V1 ở lần lặp i cuối cùng.
```

Thủ tục tính toán Neighbor(G, u): Tìm các đỉnh liền kề của u trong đồ thị G
Input: Đồ thị G = (V, E, W) và đỉnh u ∈ V
Output: N - tập các đỉnh liền của u trong đồ thị G

```
        N = ∅;
        for v ∈ V do {
            if ((u, v) ∈ E) then {
            N = N ∪ {v};
            }
        }
        return N;
```

Độ phức tạp tính toán của thuật toán LPAA

Khi thuật toán kết thúc thì những đỉnh có cùng nhãn sẽ ở trong cùng một cộng đồng của mạng xã hội. Những đỉnh trong mỗi lớp tương đương được xác định trong giai đoạn 1 có nhãn trùng với nhãn của đỉnh đại diện, do vậy chúng cũng sẽ cùng cộng đồng với đỉnh đại diện.

Thuật toán LREN (G) có độ phức tạp thời gian gần tuyến tính *O(n)* và thuật toán lan truyền nhãn trên đồ thị mạng xã hội cũng có độ phức tạp tính toán gần tuyến tính, do vậy thuật toán LPAA cũng có độ phức tạp tính toán là gần tuyến tính *O(n)*, với n = |v|.

Các đồ thị mạng xã hội thường có nhiều đỉnh tương đương với nhau theo cấu trúc, có cùng nhãn theo phương pháp lan truyền nhãn. Do vậy, việc kết hợp những đỉnh tương đương với nhau thành đỉnh đại diện sẽ giúp cho việc giảm thiểu số đỉnh và số cạnh của đồ thị khá nhiều, nhằm giảm thời gian tính toán của các thuật toán phát hiện cấu trúc cộng đồng trên mạng xã hội. Thuật toán đề xuất LPAA được phát triển trên đồ thị mạng xã hội rút gọn khá hiệu quả qua đánh giá thực nghiệm và có độ phức tạp tính toán là gần tuyến tính.

3.5. Thực nghiệm và đánh giá

Để thấy rõ hiệu quả của thuật toán đề xuất, nghiên cứu sinh thực hiện tiến hành thực nghiệm trên cùng các bộ dữ liệu được giới thiệu trong Mục 2.5.1 ở Chương 2. Nghiên cứu sinh tiến hành thực nghiệm trên nhóm dữ liệu này việc so sánh thuật toán đề xuất tính nhanh độ đo trung tâm trung gian FBC với thuật toán gốc Brandes [19], công cụ tính độ đo trung tâm trung gian tiêu biểu gần đây NetworKit [98] và thuật toán đề xuất phát hiện cộng đồng trên mạng xã hội CDAB với thuật toán gốc GN nhằm khẳng định sự vượt trội, tính hiệu quả của thuật toán đề xuất về thời gian thực hiện và chất lượng phát hiện cộng đồng mạng xã hội.

Nhóm thứ hai gồm các bộ dữ liệu gồm Zachary Karate Club và Dolphin social network được công bố trên The Koblenz network collection [47]. Nghiên cứu sinh tiến hành thực nghiệm trên nhóm dữ liệu này việc so sánh thuật toán đề xuất CDAB với thuật toán cải tiến thuật toán gốc GN tiên tiến nhất gần đây (năm 2018) là thuật toán MAA [6]. Thuật toán MAA đã công bố các kết quả nghiên cứu liên quan đến các dữ liệu thuộc nhóm thứ hai này. Vì vậy nhằm mục đích kết quả so sánh thuật toán khách quan, tin cậy thì nghiên cứu sinh thực nghiệm thuật toán đề xuất CDAB trên nhóm dữ liệu thứ hai và so sánh kết quả với các kết quả MAA đã công bố.

- Zachary karate club [47]: Mạng lưới câu lạc bộ Karate nổi tiếng của Zachary là một bộ dữ liệu chuẩn để phát hiện các cấu trúc cộng đồng. Zachary quan sát 34 thành viên của một câu lạc bộ Karate ở Hoa Kỳ trong hai năm. Do sự bất đồng giữa quản trị viên và người hướng dẫn của câu lạc bộ, một câu lạc bộ mới được thành lập bởi người hướng dẫn bằng cách lấy khoảng một nửa số thành viên câu

lạc bộ ban đầu. Cạnh giữa các đỉnh (thành viên) của mạng này thể hiện mối quan hệ hay sự tương tác xã hội giữa các thành viên bên ngoài câu lạc bộ. Hai cấu trúc cộng đồng ban đầu này được chỉ định với các hình dạng hình tròn và hình vuông.

- Dolphin social network [47]: Mạng cá heo cho thấy mối liên hệ thường xuyên giữa 62 con cá heo sống ở New Zealand. Các đỉnh là cá heo và các cạnh giữa các đỉnh cho ta thấy mối quan hệ giữa các con cá heo tương ứng với nhau.

Bảng 3.1. Bảng các bộ dữ liệu thuộc nhóm thứ hai

Stt	Bộ dữ liệu thực nghiệm	Số đỉnh đồ thị mạng xã hội	Số cạnh đồ thị mạng xã hội	Số cộng đồng thực
1	Zachary Karate Club	34	78	2
2	Dolphin social network	62	159	2

3.5.1. Cài đặt thực nghiệm

3.5.1.1. Độ đo

Độ đo đánh giá hiệu quả của thuật toán đề xuất so với phương pháp khác nghiên cứu sinh sử dụng độ đo F-measure, độ đo đơn thể mô đun Q và độ đo thông tin tương hỗ chuẩn NMI để đánh giá độ chính xác của thuật toán phát hiện cấu trúc cộng đồng trên mạng xã hội. Chi tiết các độ đo F-measure, độ đo đơn thể mô đun Q và độ đo thông tin tương hỗ chuẩn NMI được trình bày trong Mục 1.4 ở chương 1.

Nhiệm vụ chính của bài toán phát hiện cấu trúc cộng đồng trên đồ thị mạng xã hội là dự đoán số cấu trúc và các thành viên trong cấu trúc cộng đồng đó. Để đánh giá hiệu quả của thuật toán phát hiện cấu trúc cộng đồng trên đồ thị mạng xã hội các độ đo được sử dụng thường là độ đo dựa trên phép tính toán cặp (pair-counting), độ đo dựa trên độ trùng cặp (set-matching based) và độ đo dựa trên lý thuyết thông tin (information theoretic). Một số độ đo phổ biến thường được sử dụng để đánh giá chất lượng cộng đồng mạng xã hội bao gồm: Độ đo Rand, độ đo Rand điều chỉnh, độ đo tương tự Jaccard, độ đo đơn thể mô đun Q, độ đo F-measure và độ đo thông tin tương hỗ chuẩn NMI (Normal mutual information).

Tuy nhiên, để đánh giá tính hiệu quả, chất lượng phát hiện cộng đồng của thuật toán phát hiện cấu trúc cộng đồng trên mạng xã hội được đề xuất với các phương pháp khác luận án sử dụng các độ đo: độ đo F-measure, độ đo đơn thể mô đun Q và độ đo thông tin tương hỗ chuẩn NMI. Luận án sử dụng các độ đo này vì ngoài yếu tố đây là các độ đo rất phổ biến, thông dụng để đánh giá hiệu quả, chất lượng phát hiện cộng đồng mạng xã hội [64], [88], [111], [112], [113]. Ngoài ra thêm một yếu tố nữa là để đảm bảo yếu tố tin cậy, khách quan các thuật toán được luận án so sánh đã công bố kết quả nghiên cứu liên quan đến nhóm các độ đo này. Chi tiết độ đo F-measure, độ đo đơn thể mô đun Q và độ đo NMI đã được trình bày chi tiết trong Mục 1.4 tại chương 1 của luận án.

3.5.1.2. Phương pháp thực nghiệm

Để so sánh, đánh giá độ phức tạp tính toán và hiệu quả của thuật toán đề xuất FBC tính nhanh độ đo trung tâm trung gian và thuật toán đề xuất CDAB và LPAA phát hiện nhanh các cộng đồng trên đồ thị mạng xã hội rút gọn, luận án cài đặt chương trình và thực nghiệm trên những bộ dữ liệu nêu trên của các thuật toán đề xuất với thuật toán gốc là thuật toán Brandes [19], thuật toán GN [76], thuật toán LPA [85], đồng thời so sánh với công cụ tính độ đo trung tâm trung gian tiêu biểu gần đây (năm 2016) là NetworKit [98], thuật toán cải tiến GN tiên tiến gần đây (năm 2018) là thuật toán MAA [6], thuật toán cải tiến LPA tiên tiến gần đây (năm 2018) là thuật toán OLP [82].

Sau khi thực hiện đo thời gian, luận án thực hiện tính các độ đo về chất lượng của thuật toán phát hiện cấu trúc cộng đồng mạng xã hội là độ đo F-measure, độ đo đơn thể mô đun Q và độ đo NMI của lần lượt từng thuật toán với từng bộ dữ liệu.

3.5.1.3. Kịch bản thực nghiệm

Các thuật toán thực nghiệm được luận án thực hiện riêng lẻ với từng bộ dữ liệu và lúc này trên máy tính chỉ thực hiện duy nhất một chương trình.

Luận án thực hiện thực nghiệm lần lượt thuật toán tính độ đo trung tâm trung gian của thuật toán gốc là thuật toán Brandes [19], công cụ tính độ đo trung tâm trung

gian tiêu biểu gần đây NetworKit [98] và thuật toán đề xuất FBC. Thời gian đo bắt đầu từ lúc thuật toán bắt đầu chạy cho đến khi thuật toán dừng.

Tiếp theo, các thực nghiệm thuật toán phát hiện cấu trúc cộng đồng trên mạng xã hội là thuật toán GN và thuật toán đề xuất CDAB được lần lượt thực hiện. Thời gian của thuật toán được tính bắt đầu từ lúc thuật toán chạy đến lúc thuật toán dừng trên từng bộ dữ liệu. Trong đó, thời gian đo của thuật toán đề xuất CDAB đã bao gồm thời gian rút gọn các lớp đỉnh tương đương theo độ đo trung tâm trung gian của đồ thị mạng xã hội.

Luận án tiếp tục thực hiện thực nghiệm lần lượt hai thuật toán phát hiện cấu trúc cộng đồng trên mạng xã hội là thuật toán LPA và thuật toán đề xuất LPAA. Thời gian của thuật toán được tính bắt đầu từ lúc thuật toán chạy đến lúc thuật toán dừng trên từng bộ dữ liệu. Trong đó, thời gian đo của thuật toán đề xuất LPAA đã bao gồm thời gian rút gọn các lớp đỉnh tương đương theo nguyên lý lan truyền nhãn của đồ thị mạng xã hội.

Môi trường thực nghiệm là máy tính PC với cấu hình Intel™ Core™ i7-9700CPU @4.70 GHz, 8 GB RAM, sử dụng hệ điều hành Windows 10. Công cụ lập trình thực hiện thuật toán là ngôn ngữ lập trình Python.

3.5.2. Đánh giá kết quả

Kết quả thực nghiệm bao gồm:

- Kết quả thực nghiệm đánh giá hiệu quả của thuật toán đề xuất tính nhanh độ đo trung tâm trung gian FBC với thuật toán gốc Brandes [19], công cụ tính độ đo trung tâm trung gian tiêu biểu gần đây NetworKit [98].
- Kết quả thực nghiệm đánh giá hiệu quả của thuật toán đề xuất phát hiện cấu trúc cộng đồng trên mạng xã hội CDAB với thuật toán gốc điển hình GN [76] và với thuật toán cải tiến GN tiên tiến gần đây (năm 2018) là MAA [6].
- Kết quả thực nghiệm đánh giá hiệu quả của thuật toán đề xuất phát hiện cấu trúc cộng đồng trên mạng xã hội LPAA với thuật toán gốc điển hình LPA [85] và với thuật toán cải tiến LPA tiên tiến gần đây (năm 2018) là OLP [82].

3.5.2.1. Kết quả thực nghiệm đánh giá độ phức tạp tính toán thuật toán FBC

Hiệu suất của thuật toán tính độ đo trung tâm trung gian đề xuất FBC được so sánh với thuật toán gốc Brandes [19] trong các mạng xã hội lớn [60] bằng cách sử dụng số liệu về thời gian tính toán (giây).

Bảng 3.2. Bảng thời gian tính toán độ đo trung tâm trung gian của thuật toán đề xuất FBC với thuật toán Brandes trên đồ thị mạng xã hội

Thời gian: Giây

Stt	Bộ dữ liệu thực nghiệm	Thuật toán Brandes [19]	Thuật toán đề xuất FBC
1	Com-DBLP	5849	**2105**
2	Com-Amazon	1043	**263**
3	Com-Youtube	11377	**3859**

Qua số liệu Bảng 3.2 về kết quả thực nghiệm cho thấy thời gian thực hiện của thuật toán đề xuất FBC cho thời gian tính toán vượt trội so với thuật toán tính độ đo trung tâm trung gian của Brandes trên tất cả các mạng Com-DBLP, com-Amazon, và com-Youtube. Đối với mạng có kích thước càng lớn thì thời gian thực hiện càng giảm càng lớn. Thời gian thực hiện của thuật toán đề xuất FBC so với thuật toán gốc của Brandes trên mạng com-Youtube giảm 7518 giây, với mạng com-DBLP giảm 3744 giây và với mạng com-Amazon giảm 780 giây.

Hiệu suất của thuật toán tính độ đo trung tâm trung gian đề xuất FBC tiếp tục được so sánh với công cụ tính độ đo trung tâm trung gian tiêu biểu gần đây NetworKit [98] trong các mạng xã hội lớn [60] bằng cách sử dụng số liệu về thời gian (giây).

Bảng 3.3. Bảng thời gian tính toán độ đo trung tâm trung gian của thuật toán đề xuất FBC với NetworkKit trên đồ thị mạng xã hội

Thời gian: Giây

Stt	Bộ dữ liệu thực nghiệm	NetworKit [98]	Thuật toán đề xuất FBC
1	Com-DBLP	4823	**2105**
2	Com-Amazon	542	**263**
3	Com-Youtube	7695	**3859**

Qua số liệu Bảng 3.3 về kết quả thực nghiệm cho thấy thời gian thực hiện của thuật toán đề xuất FBC cho thời gian tính toán vượt trội so với công cụ tính độ đo trung tâm trung gian tiêu biểu gần đây NetworKit trên tất cả các mạng Com-DBLP, com-Amazon, và com-Youtube. Đối với mạng có kích thước càng lớn thì thời gian thực hiện càng giảm càng nhiều. Thời gian thực hiện của thuật toán đề xuất FBC so với NetworKit trên mạng com-Youtube giảm 3836 giây, với mạng com-DBLP giảm 2718 giây và với mạng com-Amazon giảm 279 giây.

Như vậy, thuật toán đề xuất FBC giúp cho việc giảm thời gian tính toán độ đo trung tâm trung gian của các cạnh trên mạng xã hội nhưng vẫn bảo toàn được giá trị độ đo trung tâm trung gian và sử dụng thuật toán FBC vào nhóm thuật toán phát hiện cấu trúc cộng đồng dựa vào độ đo trung tâm trung gian để phát hiện cấu trúc cộng đồng trên mạng xã hội nhanh và hiệu quả hơn.

3.5.2.2. Kết quả thực nghiệm đánh giá độ phức tạp tính toán của thuật toán CDAB.

Số lượng cộng đồng được phát hiện bởi thuật toán đề xuất CDAB và LPAA được so sánh với số lượng cộng đồng được phát hiện bởi thuật toán gốc trong các mạng xã hội. Kết quả được trình bày trong Bảng 3.4.

Bảng 3.4. Số cộng đồng phát hiện bởi thuật toán GN, CDAB, LPA và LPAA

Đơn vị tính: Cộng đồng

Stt	Bộ dữ liệu thực nghiệm	Số cộng đồng phát hiện bởi thuật toán GN [76]	Số cộng đồng phát hiện bởi thuật toán CDAB	Số cộng đồng phát hiện bởi thuật toán LPA [85]	Số cộng đồng phát hiện bởi thuật toán LPAA
1	Com-DBLP	13141	13141	12768	12768
2	Com-Amazon	19246	19246	18460	18460
3	Com-Youtube	7933	7933	8138	8138

Qua số liệu Bảng 3.4 ta thấy số lượng cộng đồng được phát hiện bởi thuật toán GN và thuật toán đề xuất CDAB là như nhau và lần lượt đạt 97.5%, 97%

và 94.6% so với số lượng cộng đồng thực có trong các mạng xã hội com-DBLP, com-Amazon và com-Youtube được công bố. Như vậy thuật toán đề xuất CDAB bảo toàn số lượng cộng đồng phát hiện so với thuật toán gốc GN.

Đồng thời ta cũng thấy số lượng cộng đồng được phát hiện bởi thuật toán LPA và thuật toán đề xuất LPAA là như nhau và lần lượt đạt 94.7%, 93.1% và 97.1% so với số lượng cộng đồng thực có trong các mạng xã hội com-DBLP, com-Amazon và com-Youtube được công bố. Như vậy thuật toán đề xuất LPAA bảo toàn số lượng cộng đồng phát hiện được so với thuật toán gốc LPA trong các mạng xã hội.

Hiệu suất của thuật toán đề xuất phát hiện cấu trúc cộng đồng CDAB, LPAA tiếp tục được kiểm chứng thông qua việc so sánh với thuật toán gốc GN [76], LPA [85] trong các mạng xã hội bằng cách sử dụng số liệu về thời gian thực hiện (giây).

Bảng 3.5. Kết quả so sánh thuật toán GN, CDAB, LPA và LPAA về thời gian thực hiện

Đơn vị tính: Giây

Stt	Bộ dữ liệu thực nghiệm	Thuật toán GN [76]	Thuật toán đề xuất CDAB	Thuật toán LPA [85]	Thuật toán đề xuất LPAA
1	Com-DBLP	26325	10922	645	**269**
2	Com-Amazon	5046	2279	245	**168**
3	Com-Youtube	57218	24182	1275	**592**

Hiệu suất của thuật toán phát hiện cấu trúc cộng đồng CDAB và LPAA được khẳng định thông qua việc so sánh thời gian thực hiện với thuật toán gốc GN và LPA trong các mạng xã hội bằng cách sử dụng số liệu về thời gian tính toán là giây.

Qua số liệu Bảng 3.5 về kết quả thực nghiệm cho thấy thời gian thực hiện của thuật toán đề xuất CDAB cho thời gian tính toán vượt trội so với thuật toán GN [76] trên tất cả các mạng Com-DBLP, com-Amazon, và com-Youtube. Đối với mạng có

kích thước càng lớn thì thời gian thực hiện giảm càng nhiều. Thời gian thực hiện của thuật toán đề xuất CDAB với mạng com-Youtube giảm 33036 giây, với mạng com-DBLP giảm 15403 giây và mạng com-Amazon giảm 2767 giây.

Đồng thời kết quả thực nghiệm cho thấy thời gian thực hiện của thuật toán đề xuất LPAA phát hiện cộng đồng trên đồ thị rút gọn dựa theo nguyên lý lan truyền nhãn cho thời gian tính toán nhanh hơn so với thuật toán phát hiện cấu trúc cộng đồng phổ biến LPA. Hiệu quả thể hiện rõ trên các mạng có kích thước càng lớn thì thời gian thực hiện càng giảm nhiều. Thời gian thực hiện của thuật toán đề xuất LPAA với mạng com-Youtube giảm 683 giây, với mạng com-DBLP giảm 376 giây và với mạng com-Amazon giảm 77 giây.

Như vậy qua kết quả thực nghiệm khẳng định thuật toán đề xuất CDAB và LPAA phát hiện cộng đồng trên đồ thị rút gọn dựa vào độ đo trung tâm trung gian và nguyên lý lan truyền nhãn cho thời gian tính toán vượt trội hơn so với thuật toán gốc GN và LPA.

3.5.2.3. Kết quả thực nghiệm đánh giá độ chính xác và chất lượng các cộng đồng của thuật toán đề xuất CDAB, LPAA phát hiện cộng đồng trên mạng xã hội

Độ chính xác và chất lượng các cộng đồng của các thuật toán phát hiện cộng đồng trên mạng xã hội được đánh giá thông qua độ đo đơn thể mô đun Q, độ đo F-measure và độ đo thông tin tương hỗ NMI.

Bảng 3.6. Bảng kết quả so sánh thuật toán GN, CDAB, LPA, LPAA về chất lượng cộng đồng thông qua độ đo đơn thể mô đun Q

Stt	Bộ dữ liệu thực nghiệm	Thuật toán GN [76]	Thuật toán đề xuất CDAB	Thuật toán LPA [85]	Thuật toán đề xuất LPAA
1	Com-DBLP	0.662	**0.734**	0.671	0.721
2	Com-Amazon	0.734	**0.876**	0.786	0.825
3	Com-Youtube	0.682	**0.821**	0.512	0.659

Từ số liệu Bảng 3.6, giá trị độ đo đơn thể mô đun của thuật toán CDAB lần lượt là 0.734, 0.876, 0.821 và đối với thuật toán GN lần lượt là 0.662, 0.734, 0.682. Ta

thấy rằng thuật toán CDAB đạt được hiệu suất tốt nhất trong tất cả các mạng com-DBLP, com-Amazon và com-Youtube. Giá trị độ đo mô đun Q của thuật toán đề xuất LPAA trên các mạng com-DBLP, com-Amazon, com-Youtube lần lượt là 0.721, 0.825, 0.659 và của thuật toán LPA trên các mạng lần lượt là 0.671, 0.786, 0.512. Qua các số liệu độ đo đơn thể mô đun Q trong các bộ dữ liệu này cho ta thấy rằng phương pháp được đề xuất CDAB và LPAA vượt trội hơn so với thuật toán gốc GN và LPA trên tất cả các bộ dữ liệu thực nghiệm.

Độ đo thông tin tương hỗ NMI là một số liệu so sánh tiếp theo để đánh giá tỷ lệ chính xác được tìm thấy trong cộng đồng đã biết. Giá trị NMI càng cao đồng nghĩa tỉ lệ chính xác càng lớn, ít sai khác.

Bảng 3.7. Bảng kết quả so sánh thuật toán GN, CDAB, LPA và LPAA về chất lượng cộng đồng NMI

Stt	Bộ dữ liệu thực nghiệm	Thuật toán GN [76]	Thuật toán đề xuất CDAB	Thuật toán LPA [85]	Thuật toán đề xuất LPAA
1	Com-DBLP	0.136	0.197	0.367	**0.443**
2	Com-Amazon	0.146	0.215	0.356	**0.452**
3	Com-Youtube	0.062	**0.067**	0.033	0.041

Từ số liệu Bảng 3.7, giá trị độ đo NMI của thuật toán CDAB lần lượt là 0.197, 0.215, 0.067 và đối với thuật toán GN lần lượt là 0.136, 0.146, 0.062. Ta thấy rằng thuật toán CDAB đạt được hiệu suất tốt nhất trong tất cả các mạng com-DBLP, com-Amazon và com-Youtube.

Đồng thời qua Bảng 3.7 cũng cho thấy giá trị NMI của thuật toán đề xuất LPAA lần lượt trên các bộ dữ liệu thực nghiệm Com-DBLP, com-Amazon, com-Youtube lần lượt là 0.443, 0.452, 0.041 còn của thuật toán LPA đối với các bộ dữ liệu lần lượt là 0.367, 0.356, 0.033. Như vậy giá trị độ đo NMI của thuật toán đề xuất LPAA lớn hơn của thuật toán LPA trong mọi bộ dữ liệu thực nghiệm.

Điều đó khẳng định chất lượng phát hiện cấu trúc cộng đồng của thuật toán CDAB và LPAA là tốt hơn so với thuật toán gốc GN và LPA.

Bảng 3.8. Bảng kết quả so sánh thuật toán GN, CDAB, LPA và LPAA về chất lượng cộng đồng F-Measure

Stt	Bộ dữ liệu thực nghiệm	Thuật toán GN [76]	Thuật toán đề xuất CDAB	Thuật toán LPA [85]	Thuật toán đề xuất LPAA
1	Com-DBLP	0.542	0.686	0.725	**0.786**
2	Com-Amazon	0.614	0.758	0.803	**0.858**
3	Com-Youtube	0.041	**0.057**	0.016	0.028

Từ số liệu Bảng 3.8, giá trị độ đo F-measure của thuật toán CDAB lần lượt là 0.686, 0.758, 0.057 và đối với thuật toán GN lần lượt là 0.542, 0.614, 0.041. Ta thấy rằng thuật toán CDAB đạt được hiệu suất tốt hơn trong tất cả các mạng com-DBLP, com-Amazon và com-Youtube. Điều đó khẳng định rằng chất lượng phát hiện cấu trúc cộng đồng của thuật toán CDAB là tốt hơn so với thuật toán GN.

Đồng thời qua số liệu Bảng 3.8 cho thấy giá trị độ đo F-measure của thuật toán đề xuất LPAA lần lượt trên các bộ dữ liệu thực nghiệm Com-DBLP, com-Amazon, com-Youtube lần lượt là 0.786, 0.858, 0.028 còn của thuật toán LPA đối với các bộ dữ liệu lần lượt là 0.725, 0.803, 0.016. Như vậy giá trị độ đo F-measure của thuật toán đề xuất LPAA lớn hơn của thuật toán LPA trong mọi bộ dữ liệu thực nghiệm. Điều đó khẳng định chất lượng phát hiện cấu trúc cộng đồng của thuật toán LPAA là tốt hơn so với thuật toán gốc LPA. Như vậy, qua số liệu các Bảng 3.5, 3.6, 3.7 và 3.8 về thời gian thực hiện thuật toán và các độ đo chất lượng cộng đồng đã khẳng định rằng phương pháp được đề xuất CDAB, LPAA có thời gian thực hiện nhanh hơn và hiệu quả hơn so với thuật toán gốc GN, LPA và thường phát hiện các cộng đồng có chất lượng tốt hơn các cộng đồng được phát hiện bởi GN, LPA.

Nhằm mục đích đánh giá hiệu quả của phương pháp đề xuất CDAB được thể hiện thông qua việc so sánh với thuật toán tiên tiến nhất cải tiến thuật toán GN hiện nay là thuật toán MAA [6]. Thuật toán đề xuất CDAB được so sánh với thuật toán MAA trên lần lượt các bộ dữ liệu: Zachary Karate Club và Dolphin Social Network.

Độ đo đơn thể mô đun là độ đo tiêu chuẩn quan trọng để đo chất lượng của thuật toán phát hiện cấu trúc cộng đồng. Trong thực nghiệm này, nghiên cứu sinh chủ yếu đánh giá thuật toán đề xuất CDAB so với thuật toán MAA ở khía cạnh ảnh hưởng của độ đo đơn thể mô đun Q. Sau khi chạy tương ứng các thuật toán trong các tập dữ liệu mạng thực bao gồm Zachary Karate Club, Dolphin Social Network và tính toán độ đo đơn thể mô đun phát hiện cộng đồng.

Bảng 3.9. Kết quả so sánh thuật toán CDAB, MAA về chất lượng cộng đồng thông qua độ đo đơn thể mô đun Q

Stt	Bộ dữ liệu thực nghiệm	Thuật toán MAA [6]	Thuật toán đề xuất CDAB
1	Zachary Karate Club	0.1128	**0.3715**
2	Dolphin Social Network	0.1543	**0.3787**

Từ số liệu Bảng 3.9, giá trị độ đo đơn thể mô đun của thuật toán đề xuất CDAB lần lượt là 0.3715, 0.3787 và đối với thuật toán MAA lần lượt là 0.1128, 0.1543. Như vậy qua giá trị độ đo đơn thể mô đun Q đã khẳng định thuật toán đề xuất CDAB đạt được hiệu suất tốt hơn thuật toán MAA trong các mạng Zachary Karate Club và Dolphin Social Network. Kết quả thực nghiệm khẳng định rằng phương pháp được đề xuất CDAB hiệu quả hơn so với thuật toán MAA và thường phát hiện các cộng đồng có chất lượng tốt hơn các cộng đồng được phát hiện bởi MAA.

Như vậy, qua các kết quả thực nghiệm so sánh thuật toán đề xuất CDAB với thuật toán gốc GN và thuật toán cải tiến GN gần đây là MAA đã khẳng định rằng kết quả của phương pháp được đề xuất là đáng tin cậy và hiệu quả.

Thuật toán đề xuất CDAB đạt hiệu quả cao trong những trường hợp mạng xã hội có kích thước lớn do số lượng các đỉnh tương đương theo độ đo trung tâm trung gian nhiều, có thể kết hợp rút gọn được nhiều đỉnh và cạnh trên đồ thị mạng xã hội. Kết quả thực nghiệm đã khẳng định tính hiệu quả của thuật toán CDAB.

Tuy nhiên hạn chế của thuật toán CDAB còn độ phức tạp về thời gian tính toán và hiệu quả rất thấp trong trường hợp những mạng xã hội nhỏ do số lượng đỉnh tương

đương theo độ đo trung tâm trung gian ít, lại mất thêm thời gian rút gọn đồ thị dẫn đến việc chênh lệch thời gian tính toán so với thuật toán gốc là nhỏ.

Ngoài ra vì thuật toán CDAB là cải tiến từ thuật toán gốc GN nên thuật toán CDAB vẫn gặp phải một số hạn chế của thuật toán GN như vẫn sử dụng phương pháp loại trừ dần đến khi không có cạnh nào vượt qua ngưỡng của độ đo trung tâm trung gian cao, vì vậy nên số lượng cộng đồng không kiểm soát trước được. Bên cạnh đó, thuật toán cũng sử dụng nhiều phép phân vùng, khó có thể xác định được phép phân vùng nào mang lại hiệu quả tốt nhất.

Nhằm mục đích đánh giá hiệu quả của phương pháp đề xuất được thể hiện thông qua việc đánh giá chất lượng và tốc độ với thuật toán tiên tiến hiện nay cải tiến thuật toán LPA là thuật toán OLP [82].

Bảng 3.10. Kết quả so sánh thuật toán LPAA, OLP về chất lượng cộng đồng NMI

Stt	Bộ dữ liệu thực nghiệm	Thuật toán OLP [82]	Thuật toán đề xuất LPAA
1	Zachary Karate Club	0.7605	**0.8421**
2	Dolphin Social Network	0.7605	**0.9042**

Qua số liệu Bảng 3.10 cho thấy giá trị NMI của thuật toán đề xuất LPAA lần lượt trên các bộ dữ liệu thực nghiệm Zachary Karate Club, Dolphin Social Network, lần lượt là 0.8421, 0.9042 còn của thuật toán OLP đối với các bộ dữ liệu lần lượt là 0.7605, 0.7605. Như vậy giá trị độ đo NMI của thuật toán đề xuất LPAA lớn hơn của thuật toán OLP trong tất cả các bộ dữ liệu thực nghiệm. Điều đó khẳng định chất lượng phát hiện cấu trúc cộng đồng của thuật toán LPAA là tốt hơn so với thuật toán OLP [86].

Như vậy, qua các kết quả thực nghiệm so sánh thuật toán đề xuất LPAA với thuật toán gốc LPA và thuật toán cải tiến LPA gần đây là OLP đã khẳng định rằng kết quả của phương pháp được đề xuất là đáng tin cậy, hiệu quả hơn.

Thuật toán đề xuất LPAA đạt hiệu quả cao trong những trường hợp mạng xã hội có kích thước lớn do số lượng các đỉnh tương đương theo nguyên lý lan truyền

nhãn lớn, có thể kết hợp rút gọn được nhiều đỉnh và cạnh trên đồ thị mạng xã hội. Đặc biệt thuật toán đề xuất LPAA dễ thực hiện song song để phân tích, phát hiện nhanh, hiệu quả các cấu trúc cộng đồng trên mạng xã hội lớn, phức tạp.

Tuy nhiên hạn chế của thuật toán LPAA là hiệu quả thấp trong trường hợp những mạng xã hội nhỏ do số lượng đỉnh tương đương theo nguyên lý lan truyền nhãn khá ít dẫn đến việc chênh lệch thời gian so với thuật toán gốc không đáng kể.

Ngoài ra thuật toán LPAA vẫn gặp phải một số hạn chế của thuật toán gốc LPA như tính ngẫu nhiên của nó, bao gồm nhãn ban đầu ngẫu nhiên, thứ tự cập nhật nhãn ngẫu nhiên và chọn ngẫu nhiên một trong các nhãn tối đa làm nhãn của đỉnh khi nhãn tối đa không phải là duy nhất.

3.6. Kết luận chương 3

Chương 3 trình bày kết quả áp dụng các thuật toán rút gọn đồ thị mạng xã hội được đề xuất ở chương 2 vào phát triển các thuật toán phát hiện cộng đồng trên mạng xã hội dựa vào độ đo trung tâm trung gian và nguyên lý lan truyền nhãn. Chương này trình bày các kết quả chính như sau:

- Đề xuất thuật toán FBC cải tiến thời gian tính độ đo trung tâm trung gian của các đỉnh, cạnh trên đồ thị mạng xã hội sử dụng thuật toán REG rút gọn đồ thị dựa trên lớp đỉnh tương đương dựa vào độ đo trung tâm trung gian. Bằng lý thuyết và thực nghiệm trên các mạng xã hội luận án đã khẳng định tính hiệu quả của thuật toán đề xuất FBC.

- Đề xuất thuật toán CDAB phát hiện nhanh cộng đồng mạng xã hội trên cơ sở rút gọn đồ thị dựa vào độ đo trung tâm trung gian. Bằng lý thuyết và thực nghiệm trên các mạng xã hội luận án đã khẳng định tính hiệu quả của thuật toán đề xuất CDAB.

- Đề xuất thuật toán LPAA phát hiện nhanh cộng đồng mạng xã hội trên cơ sở rút gọn đồ thị dựa vào nguyên lý lan truyền nhãn. Đặc biệt thuật toán đề xuất LPAA dễ dàng thực hiện song song để phân tích, phát hiện nhanh, hiệu quả các cấu trúc cộng đồng trên mạng xã hội lớn, phức tạp. Bằng lý thuyết và thực nghiệm trên các mạng xã hội luận án đã khẳng định tính hiệu quả của thuật toán đề xuất LPAA.

- Các kết quả nghiên cứu trong chương 3 được công bố trong công trình [CT2], [CT3].

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

I. Kết quả đạt được của luận án

Mạng xã hội là một cấu trúc xã hội được tạo ra từ rất nhiều thực thể, tác nhân hay các tổ chức liên kết với nhau theo một hay nhiều mối quan hệ và thường được biểu diễn bởi đồ thị mạng xã hội. Cộng đồng mạng xã hội là một nhóm các thực thể có các tính chất tương tự nhau liên kết chặt chẽ với nhau và cùng đóng góp vai trò nhất định trong mạng xã hội.

Mục đích của luận án là nghiên cứu phát triển các thuật toán rút gọn đồ thị mạng xã hội dựa vào các lớp đỉnh tương đương theo độ đo trung tâm trung gian và nguyên lý lan truyền nhãn để áp dụng phát triển các thuật toán phát hiện cộng đồng trên mạng xã hội. Tính hiệu quả của thuật toán đề xuất được thực nghiệm trên các mạng xã hội thực. Kết quả thực nghiệm cho thấy thuật toán được đề xuất có hiệu suất tốt hơn trong việc phát hiện các cộng đồng mạng xã hội so với các thuật toán khác về chất lượng cộng đồng và thời gian thực hiện thuật toán. **Các kết quả chính của luận án:**

1. Trình bày một số định nghĩa, đề xuất một số các tính chất, hệ quả của các lớp đỉnh tương đương theo độ đo trung tâm trung gian trên đồ thị mạng xã hội. Từ đó, đề xuất thuật toán **REG** thực hiện rút gọn đồ thị dựa vào lớp tương đương của các đỉnh theo độ đo trung tâm trung gian. Đồng thời luận án cũng đề xuất thuật toán **FBC** cải tiến thời gian tính độ đo trung tâm trung gian trên đồ thị mạng xã hội rút gọn. Bằng lý thuyết và thực nghiệm trên các mạng xã hội luận án đã khẳng định tính hiệu quả của thuật toán đề xuất và độ phức tạp thời gian tính độ đo trung tâm trung gian trên đồ thị mạng xã hội giảm rõ rệt.

2. Phát triển thuật toán **CDAB** phát hiện nhanh các cộng đồng mạng xã hội trên cơ sở rút gọn đồ thị theo độ đo trung tâm trung gian. Bằng lý thuyết và thực nghiệm trên các mạng xã hội cũng như so sánh với thuật toán MAA mới gần đây nhất liên quan đến thuật toán được đề xuất luận án đã khẳng định tính hiệu quả của thuật toán đề xuất và thời gian phát hiện cộng đồng trên mạng xã hội giảm rõ rệt.

3. Đề xuất thuật toán **LREN** rút gọn đồ thị dựa vào lớp tương đương theo nguyên lý lan truyền nhãn và áp dụng để phát triển thuật toán lan truyền nhãn **LPAA** phát hiện

cấu trúc cộng đồng mạng xã hội trên cơ sở rút gọn đồ thị theo nguyên lý lan truyền nhãn. Bằng lý thuyết và thực nghiệm trên các mạng xã hội cũng như so sánh với thuật toán OLP mới gần đây nhất liên quan đến thuật toán được đề xuất luận án đã khẳng định tính hiệu quả của thuật toán đề xuất và thời gian phát hiện cộng đồng trên mạng xã hội giảm rõ rệt.

II. Hướng phát triển của luận án

Trong quá trình nghiên cứu lý thuyết và tiến hành các thực nghiệm về phân tích, phát hiện cấu trúc cộng đồng mạng xã hội, hướng phát triển tiếp theo của đề tài như sau:

1. Như chúng ta đã biết dữ liệu mạng xã hội (Social Data) có kích thước vô cùng lớn, khả năng phát triển nhanh, khó thu thập và phân tích với các công cụ thống kê hay ứng dụng cơ sở dữ liệu truyền thống. Vì vậy, việc tiếp tục thực hiện các nghiên cứu tiên tiến về công nghệ dữ liệu lớn (Big Data) sẽ giải quyết được các công việc hiện còn đang gặp nhiều khó khăn, thách thức như: phân tích, xử lý, phát hiện các cấu trúc cộng đồng mạng xã hội trên những mạng xã hội siêu lớn.

2. Hiện nay, luận án đã thực hiện đề xuất cải tiến phương pháp phát hiện cộng đồng mạng xã hội dựa trên thuật toán điển hình GN. Nghiên cứu sinh nhận thấy có thể thực hiện tiếp tục các nghiên cứu phát triển những thuật toán tìm các cấu trúc cộng đồng chồng chéo trên đồ thị mạng xã hội sử dụng độ đo trung tâm trung gian cục bộ. Những cải tiến, đề xuất về thuật toán tính nhanh độ đo trung tâm trung gian cục bộ đã được nghiên cứu sinh trình bày trong các công trình [CT5].

3. Mạng xã hội đang phát triển rất nhanh, với số lượng người dùng và mối quan hệ trong mạng với nhau rất lớn. Từ đó, yêu cầu khách quan đặt ra là phải có những phương pháp nghiên cứu và các kỹ thuật phân tích mạng xã hội phù hợp. Vì vậy, việc phát triển các thuật toán song song để thực hiện đồng thời công việc phát hiện các cấu trúc cộng đồng trên mạng xã hội nhằm giảm thiểu thời gian tính toán trên dữ liệu mạng xã hội có quy mô lớn là quan trọng và cần thiết hơn bao giờ hết.

DANH MỤC CÁC CÔNG TRÌNH CÓ LIÊN QUAN ĐẾN LUẬN ÁN

CT1	**Nguyễn Xuân Dũng**, Đoàn Văn Ban, Đỗ Thị Bích Ngọc, "A Method to improve the time of computing Betweenness centrality in socical network graph", *Tạp chí Khoa học và công nghệ*, Viện hàn lâm khoa học và công nghệ Việt Nam, Số 3, 2019, Tr 344-355.
CT2	**Nguyễn Xuân Dũng**, Đoàn Văn Ban, "Một phương pháp cải tiến thời gian phát hiện cấu trúc cộng đồng trên đồ thị mạng xã hội", *Tạp chí Khoa học, Trường đại học sư phạm Hà Nội,* Tập 63, số 11A, 2018, Tr 145-158.
CT3	**Nguyen Xuan Dung**, Doan Van Ban, Truong Tien Tung, "A method to improve the time of computing for detecting community structure in social network graph", *International journal of engineering and advanced technology, Blue eyes intelligence engineering & sciences*, Volume 8, Issue 6, 2019, Tr 933-937, Scopus Indexed Journal.
CT4	**Nguyễn Xuân Dũng**, Đoàn Văn Ban, "Một phương pháp tính nhanh độ đo trung gian để phát hiện cộng đồng trên mạng xã hội", *Kỷ yếu Hội thảo Quốc gia lần thứ XXI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Thanh Hoá, 2018, Tr 198-204.
CT5	**Nguyễn Xuân Dũng**, Đoàn Văn Ban, Đỗ Thị Bích Ngọc, "Tiền xử lý dữ liệu đồ thị cải tiến thời gian tính độ đo trung gian cục bộ trên đồ thị mạng xã hội", *Kỷ yếu Hội thảo Quốc gia lần thứ XXII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Thái Bình, 2019, Tr 169-174.

TÀI LIỆU THAM KHẢO

Tài liệu Tiếng Việt

[1]. Hoàng Thị Thanh Giang, Nguyễn Thị Thúy Hạnh, Nguyễn Hoàng Huy, So sánh một số thuật toán phân cụm phổ cho dữ liệu biểu diễn gene, *Tạp chí Khoa học và Phát triển*, tập 13, số 6: 1008-1015, (2015).

Tài liệu tiếng Anh

[2]. A., Ng., Jordan, M. and Weiss, Y.: On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, Dietterich T., S. Becker, and Z. Ghahramani (Eds.), MIT Press, 14: 849 - 856 (2002).

[3]. Ahuja, M. S., Singh, J.: Future Prospects in Community Detection, *International Journal of Computer Science Engineering and Information Technology Research*, vol. 4, no. 5, pp 37-48, (2014).

[4]. Amelio, A. and Pizzuti, C.: Overlapping Community Discovery Methods: *A Survey 2014*, (2014).

[5]. Arab, M., Hasheminezhad M.: Efficient community detection algorithm with label propagation using node importance and link weight, *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 510-518 (2018).

[6]. Arasteh M., Alizadeh. S.: A fast divisive community detection algorithm based on edge degree betweenness centrality, *Springer Science Business Media*, LLC, part of Springer Nature, 49 (2): 689 - 702, (2018).

[7]. Aref, M., Moawad, I. F., Mahmoud, M.: A Survey on Graph Reduction Methods and Applications, *Egyptian Journal of Language Engineering*, Vol. 1, No. 2, (2014).

[8]. Arif, T.: The Mathematics of Social Network Analysis: Metrics for Academic Social Networks, *International Journal of Computer Applications Technology and Research*, Volume 4 - Issue 12, 889 - 893, ISSN: 2319-8656, (2015).

[9]. Baagyere, E. Y., Qin, Z., Xiong, H., and Zhiguang, Q.: The Structural Properties of Online Social Networks and their Application Areas, *IAENG International Journal of Computer Science*, 43:2, IJCS_43_2_03, (2016).

[10]. Bader, D. A., Kintali, S., Madduri, K., Mihail, M.: Approximating Betweenness centrality. *In WAW* (2007).

[11]. Bader, D. A., Madduri, K..: Parrallel algorithm for evaluating centrality indices in real-world networks. *In ICPP* (2006).

[12]. Bai, L., Liang, J., Du, H. and Guo, Y.: A novel community detection algorithm based on simplification of complex networks, *Knowledge-Based Systems*, 143:58-64, (2018).

[13]. Barber, M. J., Clark, J. W.: Detecting network communities by propagating labels under constraints. *Physical Review E-Statistical*, Nonlinear, and Soft Matter Physics. 80(2)026129, (2009).

[14]. Bezdek, J. C.: Pattern recognition with fuzzy objective function algorithms: *Springer Science & Business Media* (2013).

[15]. Blondel, V. D, Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J*ournal of Statistical Mechanics: Theory and Experiment*. 2008. doi:10.1088/1742-5468/2008/10/P10008, (2008).

[16]. Boettcher, S., and Percus, A. G.: Optimization with extremal dynamics. *Physical Review Letters*, 86(23), 5211 (2001).

[17]. Bortner, D. and Han, J.: Progressive clustering of networks using structure connected order of traversal. *ICDE*, pages 653-656, (2010).

[18]. Brandes, U., Pich, C.: Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*. (2007).

[19]. Brandes, U.: A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163-177. (2001).

[20]. Carlos, A. and Coello, C.: An Introduction to Evolutionary Algorithms and Their Applications, *ISSADS 2005*, LNCS 3563, pp.425-442, (2005).

[21]. Chen, J. and Saad, Y.: Dense subgraph extraction with application to community detection. *TKDE*, 24(7):1216-1230, (2012).

[22]. Clauset, A., Newman, M. E., and Moore, C.: Finding community structure in very large networks. *Physical Review E*, 70(6):066111, (2004).

[23]. Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C.: Introduction to Algorithms. MIT Lincoln Laboratory Series, *The MIT Press*, 3 ed. 17 (2009).

[24]. Creusefond, J.: Characterising and detecting communities in social networks. *Ph.D. thesis*. Normandie Universite (2017).

[25]. De Meo, P., Nocera, A., Terracina, G., and Ursino, D.: Recommendation of similar users, resources and social networks in a social internetworking scenario, *Information Sciences*, vol. 181, no. 7, pp. 1285-1305, (2011).

[26]. Dhumal, A., and Kamde, P.: Survey on Community Detection in Online Social Networks. *International Journal of Computer Applications*, 121(9) (2015).

[27]. Duch, J., and Arenas, A.: Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2), 027104 (2005).

[28]. Dutta, K.: Graph Theoretic Approach to Social Network Analysis, *International Journal of Scientific Research in Science and Technology*, (4) 2: 1550-1557, (2018).

[29]. Edmonds, N., Hoefler, T., Lumsdaine, A.: A space efficient parallel algorithm for computing betweeness centrality in distributed memory. *In HiPC*. (2010).

[30]. Erdos, D., Ishakian, V., Bestavros, A., Terzi, E.: A divide and Conquer Algorithm for Betweeness Centrality. *Proceedings of the 2015 SIAM International Conference on Data Mining*, SDM, pp. 433-441, (2015).

[31]. Fortunato, S.: Community detection in graphs, *Technical Report,* Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Torino, ITALY, arXiv:0906.0612v2 (2010).

[32]. Freeman, L. C., and Linton, C.: A set of measures of centrality based on Betweenness centrality. *Sociometry*, 40(1), Pages 35-41 (1977).

[33]. Freeman, L. C., Linton, C.: Centrality in social networks conceptual clarification. *Social Networks* 1.3:215-239 (1978).

[34]. Freeman, L.: The Development of Social Network Analysis. *Vancouver: Empirical Press*, (2006).

[35]. Furht, B.: Handbook of social network technologies and applications: *Springer Science & Business Media* (2010).

[36]. Geisberger, R., Sanders, P., Schultes, D.: Better approximation of betweeness centrality. *In ALENEX*. (2008).

[37]. Girvan, M., Newman, M. E. J.: Community structure in social and biological networks, *Proceedings of the National Academy of Sciences of the United States of America*, Vol.99, No.12, pp. 7821-7826 (2002).

[38]. Gregory, S.: A Fast Algorithm to Find Overlaping Communities in Network. In: Joint European conference on machine learning and knowledge discovery in databases. *Springer*, Berlin, Heidelberg, pp. 408 - 423 (2008).

[39]. Gregory, S.: Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12 (10): 103018. (2008).

[40]. Gregory, S.: Finding overlapping communities using disjoint community detection algorithms. *Complex networks*,: 47-61 (2009).

[41]. Halkidi, M., Batistakis, Y.,: "Cluster validity methods: part I." *ACM SIGMOD Record* 31(2): 40-45, (2002).

[42]. Hanneman, R. A., Riddle, M.: Introduction to social network analysis. *Riverside, CA: University of California*, (2005).

[43]. Hu, Y., Yang, B.: Characterizing the structure of large real networks to improve community detection. *Neural Comput*. Appl. 28, 2321-2333, (2016).

[44]. Huang, J., Sun, H., Song, Q., Deng, H., and Han, J.: Revealing density-based clustering structure from the core-connected tree of a network. *TKDE*, 25(8):1876-1889, (2013).

[45]. Hübler C, Kriegel HP, Borgwardt K, Ghahramani Z.: Metropolis algorithms for representative subgraph sampling. *In: Proceedings of the 2008 eighth IEEE international conference on data mining*, ICDM ’08, pp 283-292, (2008).

[46]. Huq, S. T., Ravi, V. and Deb, K.: Evolutionary Multi Objective Optimization Algorithm for Community Detection in Complex Social Networks, *SN Computer Science*, 1 (2021).

[47] Institute of Web Science and Technologies - University of Koblenz - Landau: KONECT: The Koblenz Network Collection, Avaiable: https://konect.uni-koblenz.de.

[48]. Jamour, F., Skiadopoulos, S., and Kalnis, P.: Parallel Algorithm for Incremental Betweenness Centrality on Large Graph, *IEEE Transactions on Parallel and Distributed Systems*, Volume 29, Issue 3, Pages 659-672, (2018).

[49]. Jia, H. C., and Ratnavelu, K.: A semi-synchronous label propagation algorithm with constraints for community detection in complex networks. *Scientific Reports*, 7: 45836 (2017).

[50]. Kernighan, B. W., and Lin, S.: An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2), 291-307 (1970).

[51]. Khan, K., Nawaz, W. and Lee, Y.: Set-based unified approach for summarization of a multi-attributed graph, *World Wide Web*, 20(3): 543-570, (2017).

[52]. Khorasgani, R. R., Chen, J., and Za¨iane, O. R.: Top leaders community detection approach in information networks. *SNA-KDD*, (2010).

[53]. Kirianovskii I., Granichin O., Proskurnikov,A.: A new Randomized Algorithm for Community Detection in Large Networks, *12 th IFAC International Workshop on Adaptation and Learning in Control and Signal Processing*, Eindhoven, The Netherlands, June 29 - July 1, (2016).

[54]. Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P.: Optimization by simmulated annealing. *Science*, 220(4598), 671-680 (1983).

[55]. Knuth, D. E.: The Art of Computer Programming Vol 1. 3rd ed., *Boston: Addison-Wesley*, ISBN 978-0-201-89683-1 (1997).

[56]. Krackhardt, D.: Assessing the Political Landscape: Structure, Cognition, and Power in Organizations, *Administrative Science Quarterly*, 35, pp. 342-369, (1990).

[57]. Kumpula, J. M., Kivela¨, M., Kaski, K., and Sarama¨ki, J.: Sequential algorithm for fast clique percolation. *Physical Review E*, 78(2):026109, (2008).

[58]. Lee, B., Plaisant C., Parr, C. S., Fekete, J., and Henry, N.: Task taxonomy for graph visualization. In Proc. *Of the 2006 AVI Workshop on Beyond time and errors: novel evaluation methods for in formation visualization*, BELIV'06, pages 1-5, (2006).

[59]. Lee, C., Reid, F., Mcdaid, A.: Detecting highly overlapping community structure by greedy clique expansion. (2010).

[60]. Leskovec J., Krevl A.: SNAP Datasets: Stanford large network dataset collection, Avaiable: https://snap.stanford.edu (2014).

[61]. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In Proc. *of the 20th ACM SIGKĐ Intl. Conf. On Knowledge Discovery and Data Mining*, pages 631-636, (2006).

[62]. Li, W., Kang, Q., Kong, H., Liu, C. and Kang, Y.: A novel iterated greedy algorithm for detecting communities in complex network, *Social Network Analysis and Mining*, pp.10-29, (2020).

[63] Lin, H., Zhap, Z., Li, H. and Chen, Z.: A novel graph reduction algorithm to identify structural conflicts, in Proc. *35th Hawaii Int. Conf. Syst. Sci*., vol. 9, p. 289, (2002).

[64]. Linhares, Claudio D. G., Bruno A. N. Travençolo, Jose Gustavo S. Paiva, and Luis E C Rocha.: Visual analysis for evaluation of community detection algorithms. *Multimedia Tools and Applications,* volume 79, pages 17645-17667 (2020).

[65]. Liu, W., Jiang, X., Pellegrini, M.: Discovering communities in complex networks by edge label propagation. *Sci Rep*, 6: 22470. (2016).

[66]. Liu, X, and Murata, T.: Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A*. 389(7):1493-1500, (2010).

[67]. Luo, W., Lu, L., Ni, L., Zhu, W. and Ding, W.: Local community detection by the nearest nodes with greater centrality, *Information Sciences*, 517:377-392, (2020).

[68]. Lusseau, D., Newman, M. E.: Identifying the role that animals play in their social networks, *Proceedings of the Royal Society of London B: Biological Sciences* 271 (Suppl 6) S477-S481 (2004).

[69]. MacQueen, J.: Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, (Vol. 1, pp. 281-297, Vol. 14): Oakland, CA, USA (1967).

[70]. Maiya AS, Berger-Wolf TY.: Sampling community structure. In: *Proceedings of the 19th international conference on World wide web*, WWW'10, pp 701-710, (2010).

[71]. Mislove, A. E.: Online social networks: measurement, analysis, andapplication stodistributed information systems. *ProQuest*, Rice University, Ann Arbor, United States (2009).

[72] Moosavi, S. A., Jalali, M., Misaghian, N., Shamshirband, S., & Anisi, M. H.: Community detection in social networks using user frequent pattern mining. *Knowledge and Information Systems*, 51(1), 159-186, (2017).

[73]. Mudduri, K., Ediger, D., Jiang, K., Bader, D. A., Chavarria-Miranda, D. G.: A faster paralel algorithm and effcient multithreaded implementations for evaluating Betweenness centrality on massive datasets. *In IPDPS*. (2009).

[74] Nakajima, K. and Shudo. K.: Estimating Properties of Social Networks via Random Walk considering Private Nodes. KDD 2020: 720-730 (2020).

[75] Nam, P. Nguyen., Thang, N. Dinh., Ying, Xuan., and My T. Thai.: Adaptive algorithms for detecting community structure in dynamic social networks, *in INFOCOM*, 2011 Proceedings IEEE, pp. 2282 - 2290, (2011).

[76]. Newman, M. E. J., Girvan, M.: Finding and evaluating community structure in networks. *Physical Review E*: Statistical, Nonlinear, and Soft Matter Physics. 69(2)026113 (2004).

[77]. Newman, M. E. J.: Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* Vol.74, Article ID 036104. (2006).

[78]. Newman, M. E. J.: Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*. 103(23):8577-8582, (2006).

[79]. Newman, M. E. J.: Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* 64, 016132 (2001).

[80]. Newman, M. E.: Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6), 066133 (2004).

[81]. Newman, M.: A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39-54, (2005).

[82]. Pirouz M., Zhan. J.: Optimized Label Propagation Community Detection on Big Data Networks, *Association for Computing Machinery*, ACM ISBN 978-1-4503-6358-7/18/03, (2018).

[83]. Potterat, J., Phillips-Plummer, L., Muth, S., Rothenberg, R., Woodhouse, D., Maldonado-Long, T., Zimmerman, H., and Muth, J.: Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs, *Sexually Transmitted Infections*, 78, pp. 159-163, (2002).

[84]. Puzis, R., Zilberman, P., Elovici, Y., Dolev, S, Brandes, U.: Heuristics for Speeding up Betweenness centrality Computation, *ASE/IEEE International Conference on Privacy*, Security, Risk and Trust. (2012).

[85]. Raghavan, U. N., Albert, R., and Kumara S.: Nearlinear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, (2007).

[86]. Rand, W. M.: Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* 66, 846 (1971).

[87]. Riondato, M., Kornaropoulos, E. M.: Fast approximation of Betweenness centrality through sampling *WSDM'14*, pages 413-422. (2014).

[88]. Rossetti G., Pappalardo L., Rinzivillo S.: A novel approach to evaluate community detection algorithms on ground truth. *Complex Networks VII*, 133-144, (2016).

[89] Sadiq, W. and Orlowska, M. E.: Analyzing Process Models Using Graph Reduction Techniques, *Information Systems*, 25(2): 117-134, (2000).

[90] Sadiq, W. and Orlowska, M. E.: Applying Graph Reduction Techniques for Identifying Structural Conflicts in Process Models, *Springer-Verlag Berlin Heidelberg* (1999).

[91]. Sariyuce, A. E., Saule, E., Kaya, K., Catalyurek, U. V.: Shattering and compressing networks for Betweenness centrality. *In SDM*, (2013).

[92]. Satuluri, V. and Parthasarathy, S.: Scalable graph clustering using stochastic flows: applications to community discovery. *SIGKDD*, pages 737-746, (2009).

[93] Scheuermann, B. and Rosenhahn, B.: SlimCuts: GraphCuts for High Resolution Images Using Graph Reduction, *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, July (2011).

[94]. Schuetz, P., and Caflisch, A.: Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Physical Review E*. 77(4)046112, (2008).

[95]. Scott, J.: Social network analysis: a Handbook. *London: SAGE publications*, (1991).

[96]. Shamma, D. A., Kennedy, L., Churchill, E. F.: Tweet the Debates: Understanding Community Annotation of Uncollected Sources, *In Proceedings of the first SIGMM workshop on Social media*, ACM, USA, (2009).

[97]. Shi, J., and J. Malik: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888 - 905 (2000).

[98]. Staudt, C. L., Sazonovs, A., Meyerhenke, H.: NetworKit: A tool suite for large-scale complex network analysis. *Network Science*, 4(4), 508-530, (2016).

[99]. Steinhaeuser, K., Chawla, N. V.: Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5): pp. 413-421, (2010).

[100]. Stumpf, M. P., Wiuf C., and May, R. M.: Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc.Natl.Acad. Sci.U.S.A*, 102(12): 4221-4224, (2005).

[101]. Tan, G., Tu, D., Sum, N.: A parallel algorithm for computing Betweenness centrality. *In ICPP*. (2009).

[102]. Tang, L. and Liu, H.: Community detection and mining in social media. *Synthesis Lecftures on Data Mining and Knowledge Discovery*, 2(1), pages 1-137, (2010).

[103] Tripathy, A., Yelick, K., Buluc, A.: Reducing Communication in Graph Neural Network Training, SC20, *IEEE* (2020).

[104]. Wang, T., Qian, X., Wang, X.: HLPA: A hybrid label propagation algorithm to find communities in large-scale networks// IEEE, *International Conference on Awareness Science and Technology*. IEEE, :135-140. (2015).

[105]. Wasserman, S., Faust, K.: Social network analysis: methods and applications, volume 8 of structural analysis in the social sciences. *Cambridge University Press*, Cambridge (1994).

[106]. Wellman, Barry and Berkowitz S. D.: Social Structures: A Network Approach. Cambridge: *Cambridge University Press* (1988).

[107]. Whang JJ, Sui X, Dhillon IS.: Scalable and memory-efcient clustering of large-scale social networks. In: *2012 IEEE 12th international conference on data mining, ICDM'12*, pp 705-714, (2012).

[108]. Wilson, R. J.: Introduction to Graph Theory. *Pearson Publisher*, 5 ed (2010).

[109]. Wu, Z. H., Lin, Y. F., Gregory, S.: Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology*, 27 (3): 468 - 479. (2012).

[110]. Xie, J., Szymanski, B. K.: Towards linear time overlapping community detection in social network, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg,: 25 - 36. (2012).

[111] Xu, Y, Salapaka, S. M., and Beck, C. L.: On reduction of graphs and Markov chain models, *in Proc. CDC-ECE*, pp.2317-2322, (2011).

[112]. Yang Z, Algesheimer R, Tessone CJ.: A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6, (2016).

[113]. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approarch. *In proceedings of the sixth ACM international conference on Web search and data mining*. Pages 587-596. ACM, (2013).

[114]. Yin C, Zhu S, Chen H, Zhang B, David B,: A method for community detection of complex networks based on hierarchical clustering. *IJDSN 2015*, 849140:1-849140:9, (2015).

[115]. Zachary W.: "An information flow model for conflict and fission in small groups", *Journal of Anthropological Research, vol.33,* pages 452-473. (1977).

[116]. Zhang, A. P. , Ren, G., Jia, B. Z., Cao, H., Zhang, S.B.: Generalization of label propagation algorithm in complex networks. *Proceedings of the 25th IEEE Chinese Control and Decision Conference*; Guiyang, China. pp. 1306-1309, (2013).

[117]. Zhang, A., Ren, G., Lin, Y., Jia, B., Cao, H., Zhang, J., and Zhang. S.: Detecting Community Structures in Networks by Label Propagation with PREGiction of Percolation Transition, Hindawi Publishing Corporation, *the Scientific World Journal* Volume 2014, Article ID 148686, 14 pages, (2014).

[118]. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases, *Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD'96*. pp. 103-114. doi:10.1145/233269.233324, (1996).

[119]. Zhao, F. and Tung, A. K.: Large scale cohesive subgraphs discovery for social network visual analysis. *VLDB*, pages 85-96, (2012).

[120]. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. *CMU CALD tech report CMU-CALD-02-107*, (2002).