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

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

Hà Nội - 2021

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TẬP THỂ NGƯỜI HƯỚNG DẪN KHOA HỌC:

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

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng chấm luận cấp Học viện

Họp tại: Học viện Công nghệ Bưu chính Viễn thông

Vào hồi giờ ngày tháng năm 2021

Có thể tìm hiểu luận án tại:

- Thư viện Học viện Công nghệ Bưu chính Viễn thông
- Thư viện viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

Trong vài thập kỷ gần đây, các mạng xã hội 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 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 qui 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 trên mạng xã hội mang nhiều ý nghĩa quan trọng và có nhiều ứng dụng trong các lĩnh vực xã hội 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. Để giải quyết vấn đề này, nhiều thuật toán phát hiện cộng đồng trên mạng xã hội đã được đề xuất. Tuy nhiên, các thuật toán này phần lớn chưa đạt được hiệu quả trong việc phát hiện cộng đồng trên các mạng xã hội quy mô lớn. Phát hiện cộng đồng mạng xã hội còn được gọi là phân cụm đồ thị, là một trong những kỹ thuật phân tích mạng cơ bản và quan trọng được sử dụng để phát hiện các mối quan hệ giữa các thành viên trong mạng xã hội. Liên quan đến việc xác định số lượng cộng đồng trong mạng xã hội và số lượng thành viên của mỗi cộng đồng, với nhiều tương tác giữa các thành viên trong cùng một cộng đồng nhiều hơn giữa các thành viên trong cộng đồng của nó với phần còn lại của mạng. Với sự phát triển mạnh mẽ của công nghệ thông tin, việc sử dụng mạng xã hội trong xã hội của chúng ta đang phát triển theo cấp số nhân. Một hệ quả là sự thay đổi sâu sắc trong cách người dùng tương tác với nhau. Cộng đồng là một đặc tính quan trọng của mạng xã hội, cộng đồng thường đại diện cho các nhóm người dùng được tổ chức cụ thể với các thuộc tính, sở thích tương tự hoặc các mối quan hệ gần nhau hơn.

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 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 dựa vào 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 đối với 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 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 phương pháp 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 và 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: Girvan - Newman, Label Propagation Algorithm.
- 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.

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 gốc 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 điển hình Girvan-Newman (GN) sử dụng độ đo trung tâm trung gian.
- 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 các thuật toán gốc điển hình (Label Propagation Algorithm) LPA.

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

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

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 trên mạng xã hội và các phương pháp 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 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, …

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

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.

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.

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

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

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

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:

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

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.

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:

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

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:

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

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. Nhờ đó mà chi phí tính toán giảm, thời gian tính toán giảm nhưng không làm giảm chất lượng của giải pháp hay cấu trúc 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

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, $C = \{G_1, G_2, \ldots, G_k\}$, 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 $i \neq j = 1, 2, \ldots, k\}$.

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.
- 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.
- 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.
- 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.

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.

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: $v_1, v_2, \ldots, v_n$ 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 các giải pháp phát hiện cộng đồng trên mạng xã hội đã được nghiên cứu là rất nhiều và thường xuyên [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.

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.

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.

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

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

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 và hiệu quả các cộng đồng trên mạng xã hội.

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 nghiên cứu 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 và chưa thực sự hiệu quả. Một trong những cách tiếp cận để khắc phục nhược điểm 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 vẫn 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

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

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

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

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

Các cách tiếp cận rút gọn đồ thị phần lớn phụ thuộc vào các đặc tính cơ bản của lĩnh vực ứng dụng. Hầu hết 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 [71].

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.

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.

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].

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 rất

phổ biến, thông dụng, hữu hiệu được sử dụng thường xuyên để đánh giá hiệu quả, chất lượng phát hiện cộng đồng mạng xã hội [64], [88], [112], [113], [114].

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

Luận án thực hiện tính tỷ lệ rút gọn đồ thị Compression (VN) của thuật toán đề xuất, từ việc phân tích hiệu suất rút gọn đồ thị cho thấy hiệu quả của thuật toán rút gọn đồ thị mạng xã hội đề xuất.

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 qui 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, 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 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 đơn thể hóa, các thuật toán dựa vào độ đo trung tâm trung gian và các thuật toán 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 không thực sự hiệu quả. Một trong những cách tiếp cận để khắc phục nhược điểm trên là 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 cấu trúc cộng đồng của đồ thị mạng xã hội gốc, nên không thể áp dụng cho bài toán phát hiện cộng đồng. Các chương sau sẽ 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, và áp dụng để phát triển các thuật toán nhanh phát hiện cộng đồng 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 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 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.

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 cạnh của đồ thị.

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)

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. [83]

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. 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 $w \in V - \{s \in V \mid \deg(s) = 1\}$, $t \in V - \{u, v, w\}$ (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)

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)

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 $G = (V, E)$, $u \in V$ 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).

Nhận xét 2.1. Nếu u là đỉnh sườn và G không phải là clique thì chắc chắn có ít nhất một đỉnh $v \in N(u)$ có bậc khác với bậc của đỉnh sườn u ($\deg(v) > \deg(u)$) trên đồ thị G. Ký hiệu $\Gamma(u) = \{u\} \cup N(u)$ là tập các đỉnh liền kề với u và kể cả u.

Nhận xét 2.2. Đồ thị con sinh bởi $\Gamma(u)$ 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).

$\Gamma1(u) = \{ v \in \Gamma(u) \mid \deg(v) = \deg(u)\}$ - Tập những đỉnh có cùng bậc với đỉnh sườn u trong clique sinh bởi $\Gamma(u)$.

Nhận xét 2.3. Nếu G không phải là clique thì $\Gamma2(u) = \Gamma(u) - \Gamma1(u) \neq \phi$, nghĩa là chắc chắn có ít nhất một đỉnh $v \in \Gamma2(u)$ trên clique sinh bởi N(u) (hay $\Gamma(u)$) 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).

Tính chất 2.5. Giả sử S là tập các đỉnh sườn tương đương, S = {v_1, v_2, …, 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 $j \neq i$, $v \in N(S)$.

Tính chất 2.6. Giả sử S là tập các đỉnh sườn tương đương, S = {v_1, v_2, …, 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 $v_i, v_j \in S$, $v \in N(S)$.

Tính chất 2.7. Giả sử S là tập các đỉnh sườn tương đương, S = {v_1, v_2, …, 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 $V_S = V - S$.

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

(i) $\delta_{st}(v) = 0$, với mọi $v \in \Gamma1(u)$, $s \neq u \neq t \in V$ (2.13)

(ii) $C_B(v) = 0$, với mọi $v \in \Gamma1(u)$ (2.14)

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.

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

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 $\Gamma2(S'_i) = \Gamma2(u)$, với $u \in S_i$. Ta có tính chất sau:

(i) $\tau_{S'_i t}(v) = |S_i| * \delta_{ut}$, $u \in \Gamma2(S'_i)$, i =1..h, u = v, $t \notin \{u, S'_1, S'_2, \ldots, S'_h\}$ (2.15)

(ii) $\tau_{S'_i t}(v) = |S_i| * \tau_{ut}(v)$, $u \in \Gamma2(S'_i)$, i = 1..h, $u \neq v$, $t \notin \{u, v, S'_1, S'_2, \ldots, S'_h\}$ (2.16)

(iii) $\tau_{s S'_i}(v) = |S_i| * \delta_{sw}$, $w \in \Gamma2(S'_i)$, i = 1..h, w = v, $s \notin \{v, S'_1, S'_2, \ldots, S'_h\}$ (2.17)

(iv) $\tau_{s S'_i}(v) = |S_i| * \tau_{sw}(v)$, $w \in \Gamma2(S'_i)$, i =1..h, $v \neq w$, $s \notin \{w, S'_1, S'_2, \ldots, S'_h\}$ (2.18)

(v) $\tau_{S'_i S'_j}(v) = |S_i| * |S_j| * \tau_{u w}(v)$, $u \in \Gamma2(S'_i)$, $w \in \Gamma2(S'_j)$, i, j =1..h, $v \notin \{u, w, S'_i, S'_j\}$ (2.19)

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). 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)

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)

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)

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

Dựa trên các tính chất của các đỉnh tương đương theo độ đo trung tâm trung gian được trình bày ở Mục 2.1, 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 (**)).

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

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.

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, nhưng 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 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 đạt 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).

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).

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.

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).

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

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].

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

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 Bảng 2.3 số lượng đỉnh và số lượng cạnh được giảm sau khi thực hiện 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.

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 Bảng 2.4 ta 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 được khẳng định là hiệu suất rút gọn tăng khi quy mô mạng xã hội tăng lên và số lượng đỉnh và cạnh rút gọn được này có ý nghĩa quan trọng đối với bài toán phát hiện cộng đồng trên mạng xã hội.

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 Bảng 2.5 số lượng đỉnh và số lượng 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 Bảng 2.6 ta thấy tỉ lệ rút gọn đồ thị mạng xã hội là khá lớn lần lượt là 0.239, 0.262 và 0.292 đối với các mạng Com-DBLP, Com-Amazon và Com-Youtube. Như vậy, một nhận xét quan trọng được khẳng định là hiệu suất rút gọn tăng khi quy mô mạng xã hội tăng lên và số lượng đỉnh và cạnh rút gọn được này có ý nghĩa quan trọng đối với bài toán phát hiện cộng đồng trên mạng xã hội.

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 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 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 đạt được hiệu quả.

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 rất lớn không ngừng phát triển theo thời gian, vì vậy các thuật toán phát hiện cộng đồng mất rất nhiều thời gian và chưa thực sự hiệu quả. Một trong những cách tiếp cận để khắc phục được nhược điểm 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 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 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 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 và thời gian thực hiện

của thuật toán đề xuất giảm thiểu đáng kể so với các thuật toán đã công bố. 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. 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 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 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.

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

3.2.1. Duyệt đồ thị theo chiều rộng

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. Tư 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 bảo toàn được 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 = ∅, queue Q = ∅, 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 đồ thị 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.

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 được biết tới 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 3.1.

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 3.1.

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.

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).

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 cấu trúc 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 cấu trúc 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, tác giả thực hiện tiến hành thực nghiệm trên cùng các bộ dữ liệu được giới thiệu trong Mục 2.4.1 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 với các kết quả MAA đã công bố.

3.5.1. Cài đặt thực nghiệm

3.5.1.1. Độ đo

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.

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 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].

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. Sau đó, thực nghiệm hai 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] và 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]. Nội dung phần này cũng trình bày những kết quả tính toán về thời gian và độ chính xác thông qua chất lượng cộng đồng được phát hiện của các thuật toán đề xuất so với những thuật toán phổ biến gần đây như các thuật toán MAA [6], và OLP [82].

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

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 nhiều. 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.

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].

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**

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.

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. Qua số liệu Bảng 3.9 kết quả 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.

- 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.

- 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 cộng đồng trên mạng xã hội lớn, phức tạp.

- Kết quả thực nghiệm cho thấy, so với thuật toán đã công bố, các thuật toán đề xuất nhanh hơn và hiệu quả hơn do giảm thiểu đáng kể số đỉnh, số cạnh của đồ thị mạng xã hội. Các kết quả nghiên cứu được công bố trong công trình [CT2], [CT3].

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

I. Kết quả đạt được của luận án

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. 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. 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. 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 các 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 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, "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 Hóa, 2018, Tr 198-204.
CT5	**Nguyễn Xuân Dũng**, Đoàn Văn Ban, Đỗ 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.