SỬ DỤNG ĐỘ TRUNG GIAN CỦA CÁC CẠNH PHÁT HIỆN CỘNG ĐỒNG GỐI NHAU

Nguyễn Hiền Trinh[1*], Trần Hải Thanh[1], Cáp Thanh Tùng[2]
[1]*Trường Đại học Công nghệ thông tin & Truyền thông – ĐH Thái Nguyên*
[2]*Trường Đại học sư phạm - ĐH Thái Nguyên*

TÓM TẮT

Bài báo đề xuất một phương pháp cải tiến nhằm phát hiện các cộng đồng gối nhau trên đồ thị mạng xã hội dựa trên độ trung gian của cạnh. Nghiên cứu giới thiệu các khái niệm liên quan đến đồ thị mạng xã hội, độ đo trung tâm của đỉnh, của đồ thị, độ trung gian của cạnh, phần thực hiện thuật toán GN (Girvan-Newman) với phép duyệt BFS (Breadth-First Search) và cải tiến GN sử dụng độ đo trung gian cạnh để tỉa dần các đỉnh/cạnh không cần thiết và tối ưu hóa trình tự lựa chọn đỉnh để phát hiện các cộng đồng gối nhau trên đồ thị mạng hiệu quả.

Từ **khóa:** *Đồ thị, mạng xã hội, cấu trúc cộng đồng, độ trung gian cạnh, cộng đồng gối nhau.*

MỞ ĐẦU

Mạng xã hội thường được mô hình hóa như là đồ thị và được gọi là đồ thị mạng xã hội. Các thực thể là các đỉnh (nút) và cạnh (cung), giữa hai đỉnh là mối liên kết giữa hai thực thể trên mạng. Thông thường, đồ thị mạng xã hội là đồ thị vô hướng, ví dụ đồ thị mạng bạn bè trên Facebook. Nhưng, chúng có thể là đồ thị có hướng, như đồ thị mạng xã hội những người theo dõi (followers) trên Twitter hoặc Google+.

Trên mạng, một số đỉnh có liên kết chặt chẽ với nhau tạo thành từng cụm, và giữa các cụm đó được nối với nhau chỉ bằng một vài cạnh khác [1]. Cộng đồng 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. Ví dụ, trong mạng trích dẫn (Citation Network), cộng đồng có thể biểu diễn cho những bài báo về cùng một chủ đề, hay trên các trang web, cộng đồng có thể biểu diễn cho những trang web về những chủ đề có liên quan với nhau

Việc nghiên cứu, phát hiện và khai thác hiệu quả các thông tin trên các cấu trúc cộng đồng của mạng xã hội được nhiều nhà khoa học giới thiệu và tập trung nghiên cứu. Trong phạm vi bài báo, chúng tôi giới thiệu những khái niệm cơ bản về mô hình đồ thị mạng xã hội, các phương pháp tiếp cận nhằm phát hiện cấu trúc cộng đồng và đề xuất phương pháp cải tiến cách sử dụng độ trung gian của các cạnh nhằm phát hiện cộng đồng gối nhau trên mạng xã hội.

Phần còn lại của bài báo được tổ chức như sau: Phần 2 giới thiệu những khái niệm cơ sở liên quan và cần thiết của lý thuyết đồ thị trong phát hiện cấu trúc cộng đồng. Phần 3, trình bày Thuật toán phát hiện cộng đồng sử dụng độ trung gian GN và phần thực nghiệm bằng phép duyệt BFS, sau đó trình bày cải tiến GN sử dụng độ đo trung gian cạnh để tỉa dần các đỉnh/cạnh không cần thiết, tối ưu hóa trình tự lựa chọn đỉnh để phát hiện các cộng đồng gối nhau trên đồ thị mạng, kết quả thực hiện thuật toán GN với phép duyệt BFS và GN cải tiến, cuối cùng là kết luận.

MỘT SỐ VẤN ĐỀ CƠ SỞ

2.1. Đồ thị là cấu trúc rời rạc G = (V, E), trong đó V là tập các đỉnh và E là tập các cạnh. Đồ thị G = (V, E) là vô hướng nếu các cạnh $(v, v') \in E$ không phân biệt thứ tự, ngược lại là đồ thị có hướng. Đường đi có độ dài n đi từ nút v đến w là dãy các cạnh $(v_0, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n)$, trong đó $v_0 = v$, $v_n = w$ và $(v_i, v_{i+1}) \in E$.

2.2. Cho trước đồ thị mạng xã hội G = (V, E), trong đó V là tập các đỉnh và E là tập các cạnh. Cấu trúc cộng đồng (gọi tắt là cộng đồng) C là tập con các đỉnh của V, sao cho với mỗi đỉnh $v \in C$, có nhiều cạnh kết nối v

* *Tel. 0987 562055, Email: nhtrinh@ictu.edu.vn*

với những cạnh khác trong C và ít cạnh kết nối v với những đỉnh w khác thuộc V\C.

2.3. Xét mạng xã hội G = (V, E). Ta qui ước, những đỉnh (nút) gần nhau (closed) nếu chúng có cạnh nối trực tiếp giữa chúng, ngược lại là xa nhau (distant). Khoảng cách giữa đỉnh x và y ∈ V, ký hiệu là d(x, y) và được định nghĩa:

1. d(x, y) = 0 nếu (x, y) ∈ E, ngược lại là d(x, y) = 1.
2. Hoặc d(x, y) = 1 nếu có cạnh nối giữa chúng, và bằng ∞ khi chúng xa nhau.

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.

Xét một số độ đo (Measures) trung tâm của đồ thị dựa trên khái niệm khoảng giữa (Betweenness) để phát hiện các cộng đồng mạng được đề xuất bởi Freeman [2].

2.4. Với mạng G = (V, E), xét một đỉnh $v_k \in V$ và cặp các đỉnh $\{v_i, v_j\}$ bất kỳ, không phân biệt thứ tự với $i \neq j \neq k$. Độ trung gian bộ phận của đỉnh v_k đối với $\{v_i, v_j\}$, ký hiệu là $B_{ij}(v_k)$ được định nghĩa:

Nếu giữa v_i và v_j không có đường đi thì $B_{ij}(v_k) = 0$

Ngược lại, nếu giữa chúng có đường đi, nghĩa là chúng có thể truyền thông với nhau qua một số đường đi. Khi đó xác suất trao đổi, quan hệ giữa chúng là $\frac{1}{g_{ij}}$, với g_{ij} là số đường đi ngắn nhất giữa v_i và v_j. Như vậy, tiềm năng mà v_k điều khiển (control) thông tin trao đổi giữa v_i với v_j được xác định bằng chính xác suất mà v_k nằm trên những đường đi ngắn nhất giữa chúng. Ký hiệu $G_{ij}(v_k)$ là số đường đi ngắn nhất có đi qua v_k, ta có

$$B_{ij}(v_k) = \frac{G_{ij}(v_k)}{g_{ij}} \quad (1)$$

2.5. Độ trung tâm của đỉnh v_k trong đồ thị G = (V, E), ký hiệu là $C(v_k)$ được xác định như sau:

$$C(v_k) = \sum_{i,j,i<j}^{n} B_{ij}(v_k), \text{ với } |V| = n \quad (2)$$

Độ trung tâm của đỉnh v_k, $C(v_k)$ chính là hệ số tiềm năng để điều khiển sự liên kết giữa các đỉnh trên đồ thị. Độ trung tâm của đỉnh v_k đạt được giá trị cực đại khi mọi đỉnh khác trong G đều có cạnh nối với v_k và v_k nằm trên tất cả các đường đi ngắn nhất có độ dài lớn hơn 1. Trong đồ thị, khi mọi đỉnh đều đến được tới tất cả các đỉnh khác (trực tiếp hoặc gián tiếp đi qua v_k) thì số đường đi giữa chúng là $\frac{n(n-1)}{2}$, trong đó có n-1 đường được nối với v_k. Vậy, độ trung tâm của đỉnh v_k cực đại sẽ là $\max C(v_k) = \frac{n*(n-1)}{2} - (n-1) = \frac{n^2-3n+2}{2}$ (3)

2.6. Độ trung tâm tương đối của đỉnh v_k trong đồ thị G : $C'(v_k) = \frac{2C(v_k)}{n^2-3n+2}$ (4)

Các giá trị $C(v_k)$ và $C'(v_k)$ đều có thể sử dụng để so sánh giữa các đồ thị đối xứng liên thông hoặc không liên thông. Cả hai giá trị này đều đạt được cực đại ở những tâm điểm của những đồ thị hình sao hoặc hình bánh xe.

2.7. Đỉnh trung tâm của đồ thị G là đỉnh v_k^* có $C'(v_k^*)$ đạt giá trị cực đại.

Khi đó, độ trội (ưu thế) của đỉnh trung tâm nhất trong đồ thị sẽ là $C_G = \frac{\sum_{i=1}^{n}(C(v_k^*) - C'(v_i))}{n-1}$ (5)

Đây chính là độ lệch trung bình của đỉnh trung tâm nhất so với các đỉnh khác trên đồ thị. Giá trị C_G luôn nằm trong khoảng 0 và 1. Dễ nhận thấy, $C_G = 0$ với mọi đồ thị G mà độ trung tâm của tất cả các đỉnh đều bằng nhau, và $C_G = 1$ cho những đồ thị hình sao hoặc hình bánh xe.

2.8. Độ trung gian của cạnh (u, v) là số các cặp đỉnh x và y mà cạnh (u, v) nằm trên đường đi ngắn nhất nối giữa x và y.

Như vậy, cạnh (u, v) giữa hai cộng đồng thì u và v không nằm trong cùng một cộng đồng. Một cạnh nằm giữa hai cộng đồng (được xem như là cầu nối giữa hai cộng đồng đó), do vậy số các đường đi ngắn nhất đi qua cạnh đó thường là khá lớn. Ba độ đo $C(v_k)$, $C'(v_k)$, C_G được dùng để xác định những tâm điểm của đồ thị và chúng được ứng dụng trong nhiều mục đích khác nhau. Tuy nhiên, việc sử dụng những độ đo này chỉ phù hợp cho những mạng trong đó khái niệm độ trung gian

(Betweenness) được xem là quan trọng trong tiềm năng ảnh hưởng tới quá trình xử lý sự liên kết giữa các đỉnh. Như trong nghiên cứu các mạng truyền thông (Communication Network), vấn đề quan trọng là cần xác định những điểm có tiềm năng điều khiển truyền thông để đảm bảo mạng truyền thông hiệu quả và bền vững.

THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG SỬ DỤNG ĐỘ TRUNG GIAN

Thuật toán Girvan-Newman (GN)

Nếu một mạng lưới bao gồm các cộng đồng hoặc nhóm các cộng đồng chỉ được liên kết yếu bằng một số ít các cạnh nối chúng với nhau, thì tất cả các đường đi ngắn nhất giữa các cộng đồng khác nhau sẽ phải đi dọc theo một trong số ít những cạnh đó. Vì vậy, những cạnh kết nối các cộng đồng sẽ là cạnh có độ trung gian cao. Thuật toán Girvan-Newman (GN) [3] duyệt qua mỗi đỉnh (nút) v một lần và tính số đường đi ngắn nhất từ v tới những đỉnh khác có đi qua từng cạnh đó.

Thuật toán GN_BFS được mô phỏng:

Input: mạng xã hội được biểu diễn bởi một đồ thị G = (V, E) không có trọng số; ngưỡng trung gian α

Output: Thành phần(cụm/cộng đồng) được phân chia

Begin

1. For v∈V Do

If visited(v)=False Then BFS(v);

2. For u∈BFS(v) do <gán nhãn c[u] theo Qui tắc 1>

3. For e∈E do <gán nhãn c[e] theo Qui tắc 2>

4. While (e∈E and c[e] ≤ α) or (không còn cạnh e trung gian) Do

4.1 Chọn cạnh e={ (u,v)∈E and c(u,v)→max}

4.2 E=E-{e}

4.3 Xác định các cụm sau khi loại cạnh e

4.4 For e∈E Do < Tính lại c[e] theo Qui tắc 2>

5. Return <các cụm>;

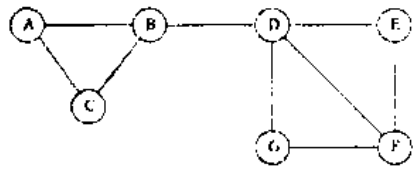
End

Để duyệt đồ thị, ta có thể sử dụng phương pháp duyệt đồ thị theo chiều rộng BFS (Breadth-First Search) hoặc theo chiều sâu DFS(Depth-First Search), bắt đầu từ đỉnh v nào đó. Để thực hiện hiệu quả, nhóm tác giả sử dụng thuật toán duyệt đồ thị theo chiều rộng BFS. Mức của mỗi đỉnh trong biểu diễn duyệt theo BFS chính là độ dài của đường đi ngắn nhất đi từ v tới đỉnh đó. BFS(v): Thực hiện duyệt đồ thị G bắt đầu từ đỉnh v và trả về thứ tự duyệt. Hàm visited(v)=False nếu v chưa được thăm và bằng True nếu v đã được thăm.

Các cạnh giữa các mức sẽ tạo thành đồ thị định hướng, phi chu trình, được gọi DAG (Directed Acyclic Graph). Mỗi cạnh của DAG sẽ là một phần của một đường đi ngắn nhất đi từ đỉnh gốc X. Nếu có cạnh (Y, Z) trên DAG, trên đó Y ở mức trên của Z (gần gốc X hơn), thì Y là cha của Z và Z là con của Y, mặc dù các đỉnh cha không cần thiết phải là duy nhất ở trong DAG giống như trên cây (tree).

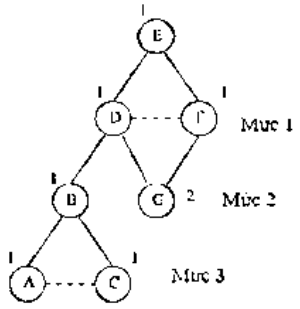
Để gắn nhãn cho từng đỉnh (nhãn của đỉnh v là số các đường đi ngắn nhất đi từ gốc tới v), ta thực hiện theo Qui tắc 1: Thực hiện từ trên/xuống: Đỉnh gốc bắt đầu được gắn nhãn bằng 1; nhãn của các đỉnh Y ở mức tiếp theo bằng tổng số nhãn của các đỉnh cha của chúng.

Ví dụ 1: Cho đồ thị G được biểu diễn bởi Hình 1

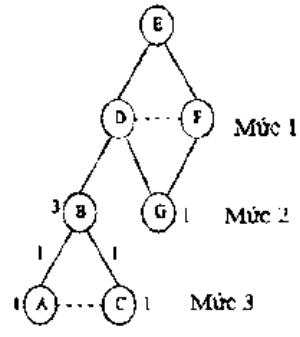


Hình 1.

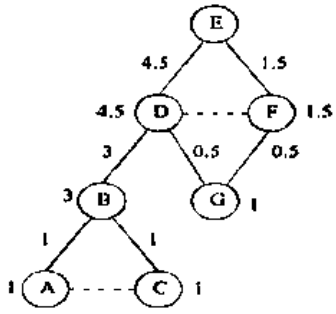
Xét dạng biểu diễn duyệt theo BFS đồ thị G, bắt đầu từ đỉnh E:



Hình 2. *Bước 1 của GN*



Hình 3. *Bước cuối của GN, mức 3*



Hình 4. *Bước cuối của GN, mức 2 và 3*

Các cạnh nét liền là các cạnh của đồ thị DAG, còn những cạnh nét đứt là nối giữa các đỉnh trong cùng một mức. Mỗi đỉnh trên DAG được gắn nhãn là số đỉnh cha của chúng. Trước tiên đỉnh gốc được gắn nhãn là 1, sau đó D, F đều có một đỉnh cha là E, nên cũng được gắn nhãn là 1. Đỉnh G có hai cha, mỗi cha có nhãn là 1 nên được gắn nhãn là 2 = 1 + 1. Với mỗi cạnh e ∈ E, ta tính tổng trên tất cả các tỉ số các đường đi ngắn nhất đi từ gốc tới đỉnh v mà có đi qua e. Qui tắc 2 được thực hiện từ dưới/lên như sau:

1. Mỗi đỉnh lá trên DAG, đỉnh không có cạnh tới đỉnh khác ở mức tiếp theo, được nhận giá trị 1.

2. Những đỉnh ở trên nó (cha), không phải là lá, nhận nhãn có giá trị là 1 cộng với tổng của các cạnh trong DAG đi từ đỉnh đó tới những đỉnh ở mức tiếp theo.

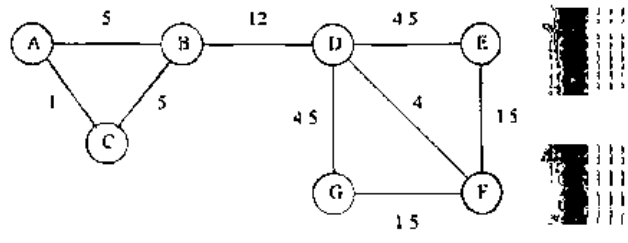
3. Một cạnh e của DAG nối đỉnh X với những đỉnh ở mức trên trực tiếp (cha của X) được chia sẻ bởi giá trị nhãn của X là tỉ lệ thuận với tỉ số các đường đi ngắn nhất đi từ gốc tới X mà có đi qua e. Một cách hình thức, giả sử cha của X là $Y_1, Y_2, \ldots, Y_k$ (ở mức trên) và p_i là số đường đi ngắn nhất đi từ gốc tới Y_i được xác định ở Bước 2 cũng chính là nhãn của X. Khi đó nhãn của cạnh (Y_i, X) sẽ là $p_i / \sum p_j$.

Sau khi tính xong các giá trị nhãn của các đỉnh, thì tính nhãn cho các cạnh. Kết quả sau khi thực hiện theo thuật toán GN và các qui tắc 1, 2 ta thu được kết quả ở hình 2, hình 3, hình 4 với đồ thị G ở hình 1 và phép duyệt BFS bắt đầu từ đỉnh E.

Sử dụng độ trung gian để tìm cộng đồng

Độ trung gian của các cạnh trong đồ thị đôi khi được xem như là độ đo độ trung gian các đỉnh trên đồ thị. Nó không hoàn toàn là độ đo khoảng cách (Metric) truyền thống, bởi theo định nghĩa, nó không thỏa bất đẳng thức tam giác. Song, chúng ta có thể phân cụm dựa vào độ trung gian của các cạnh bằng cách sắp xếp các cạnh của đồ thị theo thứ tự tăng dần của độ trung gian và tạo ra một đồ thị tương ứng. Mỗi bước, những thành phần liên thông của đồ thị sẽ tạo thành các cụm. Như vậy, độ trung gian lớn hơn, thì nhận được nhiều cạnh hơn và các cụm sẽ lớn hơn.

Phương pháp này được thể hiện qua quá trình loại bỏ dần các cạnh có độ trung gian lớn hơn một ngưỡng nào đó. Bắt đầu từ đồ thị có tất cả các cạnh cho trước, loại bỏ đi những cạnh có độ trung gian lớn nhất, tiếp tục cho đến khi đồ thị được chia thành các thành phần liên thông theo yêu cầu.



Hình 5. *Độ đo trung gian của đồ thị ở hình 1*

Ví dụ 2. Xét đồ thị cho trước G ở Hình 1.

Sử dụng thuật toán GN để tính độ đo độ trung gian của G như trong Hình 5. Dựa vào GN để loại bỏ đi những cạnh có độ trung gian lớn nhất để phân cụm đồ thị. Đầu tiên, loại bỏ cạnh (B, D) vì nó có độ trung gian lớn nhất là 12.

Khi đó ta có 2 cụm là {A, B, C} và {D, E, F, G}.

Ta có thể tiếp tục loại những cạnh có độ trung gian lớn nhất trong số còn lại là (A, B) và (B,

C) độ trung gian là 5, đồ thị sẽ chia thành 3 cụm: {A, C}, {B}, và {D, E, F, G}, ...

Thuật toán phát hiện các cộng đồng gối nhau

Các thuật toán phát hiện cộng đồng (gối nhau) trên mạng có thể chia thành hai nhóm chính: các thuật toán dựa vào đỉnh (node-based) và thuật toán dựa vào liên kết (link-based). Thuật toán phát hiện cộng đồng gối nhau dựa vào đỉnh [4] chia trực tiếp các đỉnh của mạng thành những cộng đồng khác nhau. Các liên kết trong mạng biểu diễn cho những quan hệ duy nhất, nên những thuật toán phát hiện cộng đồng gối nhau dựa vào liên kết thực hiện, trước tiên phân cụm các cạnh, sau đó ánh xạ các cộng đồng liên kết vào các cộng đồng đỉnh bằng cách gộp những đỉnh liên thuộc với tất cả các cạnh bên trong mỗi cộng đồng [5].

Có nhiều thuật toán phân cụm đồ thị được sử dụng để phát hiện các cộng đồng rời nhau, nghĩa là mỗi đỉnh chỉ thuộc một cộng đồng. Tuy nhiên, trong thực tế trên mạng xã hội nói riêng, mạng truyền thông nói chung thì phần lớn các cộng đồng không rời nhau hoàn toàn mà chúng có thể gối lên nhau (overlap), chồng lấp hay giao nhau trong một phạm vi nào đó, nghĩa là một số đỉnh có thể thuộc nhiều hơn một cộng đồng. Ví dụ mạng cộng tác trong nghiên cứu khoa học (collaboration networks), một tác giả có thể tham gia nghiên cứu cùng với một số nhà khoa học khác trong nhiều nhóm khác nhau, hay trong mạng logic-sinh học (bio-logical networks), một protein có thể tương tác với nhiều nhóm của các protein khác.

Trong bài viết này, chúng tôi mở rộng thuật toán GN để phát triển thuật toán phát hiện các cộng đồng gối nhau trên mạng dựa vào độ đo độ trung gian cạnh và kết hợp cả yếu tố đỉnh và cạnh trong phân cụm đồ thị.

Như đã trình bày, thuật toán GN được thực hiện theo kỹ thuật phân cụm phân cấp. Bắt đầu cả mạng n đỉnh được xem như một cụm. Sau đó việc loại bỏ một hoặc một số cạnh có độ trung gian lớn sẽ chia mạng đó thành hai thành phần (hai cụm). Tiếp theo, trong mỗi thành phần tiếp tục loại đi một hoặc một số cạnh để chia chúng thành hai cụm nhỏ hơn. Thuật toán lặp lại cho đến khi loại bỏ hết các cạnh của từng mạng thành phần thì mạng ban đầu sẽ chia thành n thành phần [6]. Trong phân cụm mạng, giả thiết rằng mỗi đỉnh phải nằm trong một cụm cùng với ít nhất một phần tử láng giềng của nó, trừ khi nó là một đỉnh độc lập hoặc hoàn toàn không có cụm. Vì vậy, đỉnh v có thể phân thành nhiều nhất deg(v) bản sao, trong đó deg(v) là bậc của v. Chúng ta phải quyết định chia nó bao nhiêu lần và khi nào cần phân chia. Cơ sở của phương pháp phân chia đỉnh được thực hiện như sau:

i) Khi chia một đỉnh v bất kỳ thành v_1 và v_2, ta bổ sung một cạnh ảo nối v_1 với v_2. Nếu u là lân cận của v_1 (có cạnh nối với nhau) và w là lân cận của v_2, thì tất cả các đường đi ngắn nhất đi qua v dọc theo các cạnh {u, v}, {v, w} bây giờ sẽ dọc theo {u, v_1}, {v_1, v_2}, {v_2, w}.

ii) Cạnh ảo có trọng số là 0: độ dài đường đi qua nó không thay đổi, và không có đường đi ngắn nhất nào mới được tạo ra: những đường đi bắt đầu từ v sẽ không đi qua cạnh này.

iii) Tính độ trung gian C({v_1, v_2}) của cạnh ảo được bổ sung.

Dễ thấy, có $2^{deg(v)-1}-1$ cách để chia đỉnh v bất kỳ thành hai phần. Do vậy, để thực hiện phân chia đỉnh hiệu quả thì cần chọn những đỉnh v có độ độ trung gian C({v_1, v_2}) của cạnh ảo là cực đại [4].

Thuật toán GN được cải tiến sao cho trong mỗi bước thực hiện, ta đều xem xét độ trung tâm của các đỉnh cũng như độ trung gian của các cạnh. Nếu độ trung tâm cực đại của đỉnh lớn hơn độ trung gian cực đại của cạnh thì thực hiện phân chia đỉnh, ngược lại phân chia theo cạnh (loại bỏ cạnh).

Định nghĩa 1. Độ trung gian của đỉnh v là số đường đi ngắn nhất đi qua v và nối một phần tử bất kỳ của V_1 với một phần tử bất kỳ của V_2, trong đó V_1 và V_2 là hai tập chứa tất cả đỉnh lân cận của v.

Theo định nghĩa, số đường đi ngắn nhất này không lớn hơn số đường đi ngắn nhất đi qua v. Do vậy, có thể tính độ trung gian của đỉnh

v thông qua độ trung gian của cạnh e [4].

$$C(v) = \frac{1}{2}\sum_{e\in\Gamma(v)} C(e) - (n-1) \quad (6)$$

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

Một cách tối ưu là có thể sử dụng độ trung gian của đỉnh như là cận trên để thực hiện phân chia: Nếu độ trung gian của đỉnh không lớn hơn độ trung gian cực đại của các cạnh thì không cần tính độ trung gian để chia tách đỉnh.

Định nghĩa 2. Độ trung gian cặp đỉnh (pair betweennesses) của đỉnh v đối với {u, w}, trong đó u và w là lân cận của v với u ≠ w, là số đường đi ngắn nhất đi qua cả hai cạnh {u, v} và {v, w}.

Như vậy, độ trung gian của đỉnh v chính bằng tổng tất cả các độ trung gian cặp của đỉnh đó. Ta có thể biểu diễn độ trung gian cặp của đỉnh v có bậc deg(v) = k, bằng một k-cliques (tập các đỉnh của đồ thị con trong G mà giữa 2 đỉnh bất kỳ đều có đường đi đến nhau với độ dài nhỏ hơn hay bằng k), trong đó mỗi đỉnh được ghi nhãn bằng một trong số các lân cận của v và mỗi cạnh {u, w} được ghi nhãn bằng độ trung gian cặp ("số điểm") của v đối {u, w}. Thủ tục xác định độ trung gian cặp và cách chia đỉnh v:

Procedure Mer(T:k-cliques)

Begin

1. For e∈T Do <Tính c(e) theo Qui tắc 2>
2. For i:=1 to k-2 Do

2.1. Chọn cạnh e={ (u,v)∈T and c(u,v)→min}

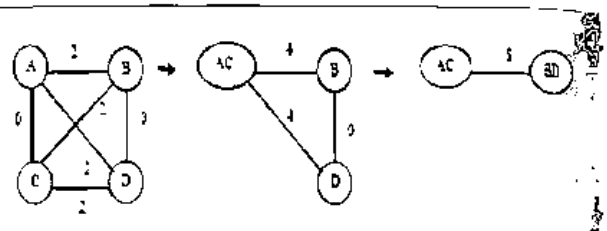
2.2. Kết hợp u, v tạo ra đỉnh đơn uv

2.3. For x∈ T Do

Thay cạnh (u,x) có nhãn t_1 và cạnh (v,x) có nhãn t_2 bằng cạnh (uv,x) có nhãn t_1+t_2

End

Ví dụ 3. Hình 6. chỉ ra cách tìm đỉnh phân chia tốt nhất. Thuật toán trên lặp lại k-2 = 2 lần; cạnh được chọn trong bước 2 của mỗi lần lặp được tô đậm.



Thuật toán GN cải tiến làm tăng độ trung gian của cạnh {v, w} đối với tất cả các đường đi ngắn nhất bắt đầu từ mỗi đỉnh s Trong thuật toán, ta làm tăng độ trung gian cặp của v đối với tất cả những cặp {u, w} sao cho w là lân cận của v trên đường đi giữa u và w. Thuật toán GN cải tiến (GN_New):

Input: mạng xã hội được biểu diễn bởi một đồ thị G = (V, E) không có trọng số

Output: Thành phần (cụm/cộng đồng) được phân chia

Begin

1. For e∈E Do <Tính c(e) theo Qui tắc 2>;
2. While < còn cạnh trung gian chưa bị loại> Do

2.1 For v∈V Do <Tính c(v) theo công thức 6>;

2.2 S={ u∈V và c(u)≥max{c(v) với v∈V }};

2.3 If S≠ϕ Then

2.3.1 Mer(S);

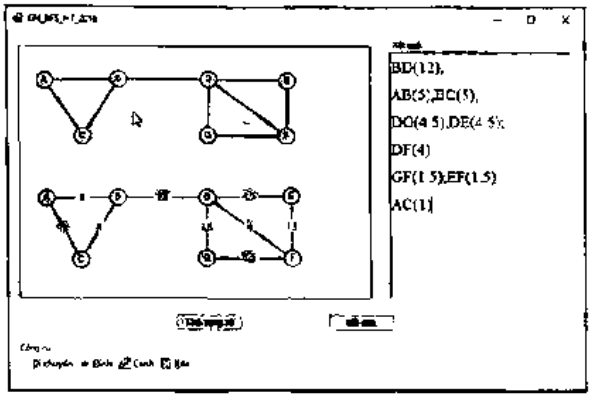
2.3.2 For v∈S Do <Tính c(v) theo công thức 6>;

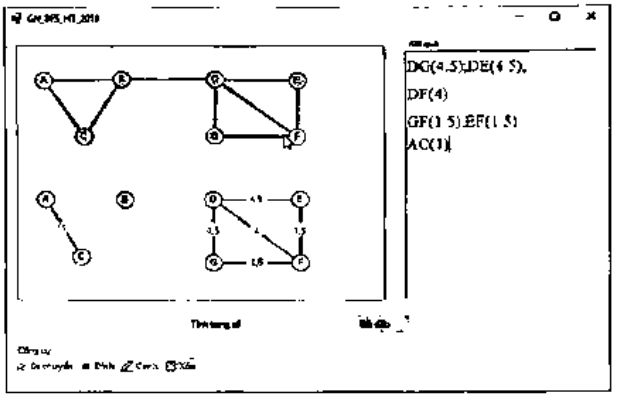
3. Chọn cạnh e={ (u,v)∈E and c(u,v)→max and u,v∈S};
4. Loại e hoặc chia đỉnh v nếu có c(e) →max hoặc c(v) →max;
5. Xác định cụm được phân chia
6. Return <các cụm> ;

End

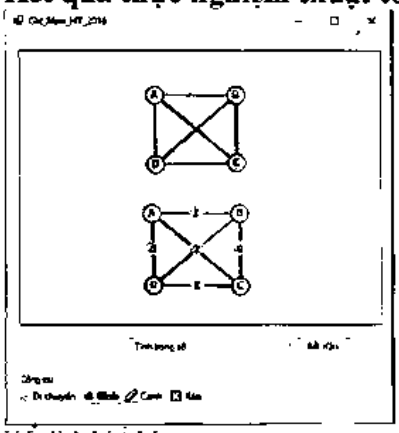
Thuật toán GN có độ phức tạp tính toán là $O(m^2n)$, trong đó m là số cạnh và n là số đỉnh của đồ thị biểu diễn mạng. Trong thuật toán cải tiến, mỗi bước thực hiện phân chia đỉnh trung bình thành 2m/n đỉnh, nghĩa là số bước phân chia đỉnh có độ phức tạp O(m); số lần lặp là O(m) và số cạnh không thay đổi. Do vậy, độ phức tạp tính toán của thuật toán là $O(m^2)$.

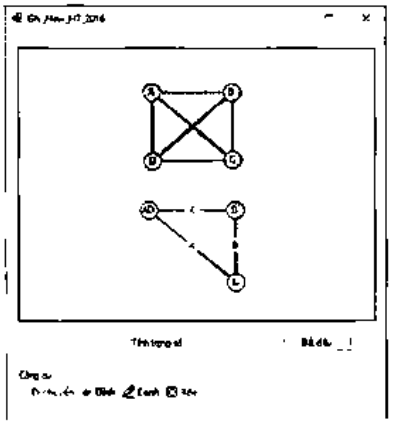
Kết quả thực nghiệm thuật toán GN_BFS

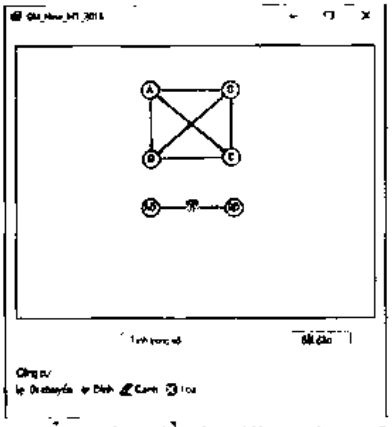




Kết quả thực nghiệm thuật toán GN_New







KẾT LUẬN

Bài báo đã nghiên cứu khái niệm liên quan đến đồ thị mạng xã hội, độ đo trung tâm của đỉnh, của đồ thị, độ trung gian của cạnh, thực nghiệm thuật toán GN (Girvan-Newman) với phép duyệt BFS(Breadth-First Search) phân cụm đồ thị và cải tiến thuật toán GN sử dụng độ đo trung gian cạnh để tỉa dần các đỉnh/cạnh không cần thiết và tối ưu hóa trình tự lựa chọn đỉnh để phát hiện các cộng đồng gối nhau trên đồ thị mạng, thuật toán GN cải tiến làm tăng độ trung gian cặp của v đối với tất cả những cặp {u, w} sao cho w là lân cận của v trên đường đi giữa u và w. Thuật toán cải tiến đã giảm thiểu phép toán phân chia đỉnh giúp cho độ phức tạp của thuật toán được cải thiện và nhanh chóng phát hiện các cộng đồng gối nhau trong phân tích và khai phá các cấu trúc cộng đồng trên mạng xã hội.

Tiếp theo chúng tôi sẽ nghiên cứu ứng dụng của thuật toán GN cải tiến trong mạng xã hội sức khỏe nhằm phát hiện và khai thác các cộng đồng bệnh nhân có tính chất gối nhau.

Bài báo là sản phẩm của đề tài Khoa học & công nghệ cấp cơ sở mã số T2016-07-10, được tài trợ kinh phí bởi Trường Đại học Công nghệ thông tin và Truyền thông - ĐHTN.

TÀI LIỆU THAM KHẢO

1. Santo Fortunato, (2010), *Community detection in graphs*, Technical Report.

2. Freeman, L.C. (2007), *A set of measures of centrality based on betweenness*. Sociometry 40, 35-41.

3. M. Girvan, M. E. J. Newman. *Community structure in social and biological networks*, Proc. Natl. Acad. Sci., 99(12), 7821, 2002

4. Gregory, S.: *An Algorithm to Find Overlapping Community Structure in Networks*. In: Kok, J.N., Koronacki, J., López de Mántaras, R., Matwin, S., Mladeni$, D., Skowron, A. (eds.) (2007), PKDD 2007. LNCS (LNAI), vol. 4702, pp. 91–102. Springer, Heidelberg.

5. Chuan Shi, Yanan Cai, Di Fu, Yuxiao Dong, Bin Wu, (2013) A link clustering based overlapping community detection algorithm, *Data & Knowledge Engineering 87* , 394–404.

6. M.E.J. Newman and M. Girvan. (2013), Finding and evaluating community structure in networks. *Preprint cond-mat/0308217*.

SUMMARY

USE OF EDGES BETWEENNESS FOR MINING ADJACENT COMMUNITIES

Nguyen Hien Trinh[1*], Tran Hai Thanh[1], Cap Thanh Tung[2]
[1]*College of Information and Communication Technology - TNU*
[2]*College of Education - TNU*

This article proposes an improved approach to mine adjacent communities in social network based on edge betweeness. The research introduces concepts concerning social network graph, measure of vertex centrality, of edge betweeness, implementation of GN(Girvan-Newman) algorithm with BFS (Breadth-First search application and an improved GN using edge betweeness measurement to trim unnecessary edges vertex and optimize vertex choice order to mine adjacent communities in social networks effectively.

Keywords: *Graphs, social network, community structure, edges betweenness, adjacent communities*

* *Tel: 0987 562055, Email: nhtrinh@ictu.edu.vn*