KHÁM PHÁ CỘNG ĐỒNG MẠNG XÃ HỘI DỰA TRÊN CƠ SỞ LÝ THUYẾT ĐỒ THỊ

TS. Nguyễn Kim Quang

Email: quangnk@ptit.edu.vn

Tóm tắt: *Sự phát triển của mạng xã hội đã tác động làm thay đổi hành vi giao tiếp của con người và dần hình thành một cách khách quan những cộng đồng trực tuyến. Tới lượt mình, những cộng đồng này lại có những tác động ảnh hưởng, chi phối tới người tham gia cộng đồng cũng như tới xã hội thực nói chung. Vì vậy, việc khám phá cộng đồng trên mạng xã hội là một việc có nhiều ý nghĩa. Khám phá cộng đồng mạng xã hội dựa trên cơ sở lý thuyết đồ thị là một hướng nghiên cứu tiềm năng hiện đang nhận được nhiều quan tâm.*

1. MỞ ĐẦU

Các nền tảng truyền thông xã hội ngày càng đa dạng, các mạng xã hội với các kết nối tương tác không ngừng được phát triển đã tạo ra một xã hội ảo ngày càng sôi động bên cạnh xã hội hiện thực. Giống như một xã hội thông thường, trong mạng xã hội cũng hình thành những cộng đồng của mình. Cộng đồng mạng xã hội là một tập hợp các cá nhân, vượt qua những ranh giới địa lý để tương tác với nhau, thường quan tâm đến một chủ đề chung hoặc cùng theo đuổi những lợi ích hay mục tiêu chung. Mối quan hệ trong cộng đồng mạng xã hội được xem như một mạng lưới các liên kết của những thành viên và mối quan hệ thể hiện các mối liên kết đó.

Khám phá cộng đồng để tìm ra những chủ đề mà cộng đồng đó quan tâm, mức độ quan tâm và độ lan tỏa của cộng đồng, các thành viên nói chung và những thành viên có ảnh hưởng trong cộng đồng,… nhằm phục vụ cho những chiến lược quản lý xã hội, chiến lược marketing đào tạo ngành nghề, quảng bá du lịch,… là những nội dung đang rất được quan tâm.

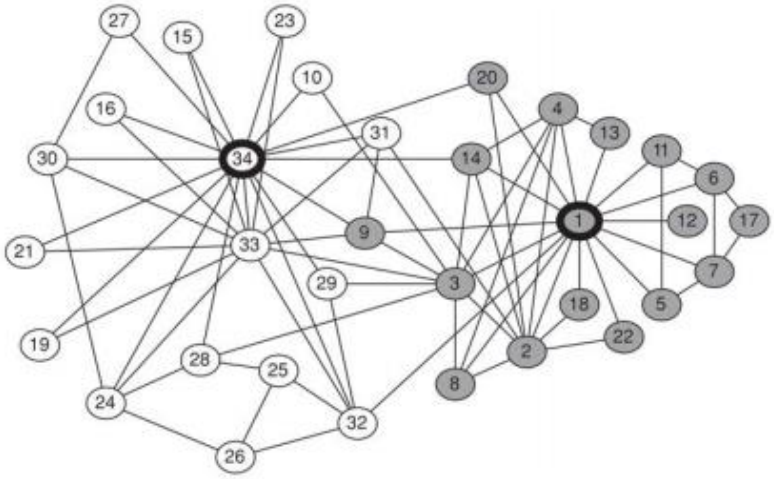
Có nhiều phương pháp khám phá cộng đồng đã và đang được nghiên cứu, đề xuất trong thời gian gần đây. Một trong những hướng nghiên cứu trong lĩnh vực này là sử dụng các cơ sở của lý thuyết đồ thị. Bài báo này nêu sơ lược về hướng nghiên cứu đó.

2. ÁP DỤNG LÝ THUYẾT ĐỒ THỊ TRONG BÀI TOÁN PHÁT HIỆN CỘNG ĐỒNG

Định nghĩa cộng đồng trên mạng xã hội:

Với một tập các đối tượng cùng loại cho trước S = {s1, . . . , sn}, một cộng đồng là một cặp C = (T, G), với T là chủ đề cộng đồng, còn $G \subseteq S$ là tập các đối tượng thuộc S chia sẻ chung chủ đề T. Khi đó $s_i \in G$ là thành viên của cộng đồng C.

Mạng xã hội bao gồm các nhân tố (actor) và mối liên hệ giữa chúng, do đó nó có thể được mô hình hóa thành một graph phức tạp, trong đó các nhân tố của mạng xã hội đó tương ứng với các nút của đồ thị còn các mối quan hệ tương ứng với các đường nối. Khi đó các cộng đồng được thể hiện bởi một graph với các node biểu thị cho các cá nhân và các cạnh biểu thị mối quan hệ giữa các cá nhân đó. Nói một cách tổng quát thì các cạnh cũng có thể biểu diễn nội dung hoặc các thuộc tính được chia sẻ giữa các cá nhân. Ví dụ, chúng ta có thể kết nối các cá nhân ở trong cùng một khu vực, cùng giới tính, cùng mua một loại sản phẩm,…Tương tự như vậy, các node cũng có thể biểu thị các sản phẩm, các trang web,…Trong một graph thì mật độ các đường nối giữa các nút tập trung cao hơn ở trong các nhóm và thấp hơn giữa các nhóm với nhau. Đó là cơ sở để nhận diện cấu trúc cộng đồng trong một network, trong đó các cộng đồng là các nhóm node trong hệ thống, được liên hệ chặt chẽ với nhau.

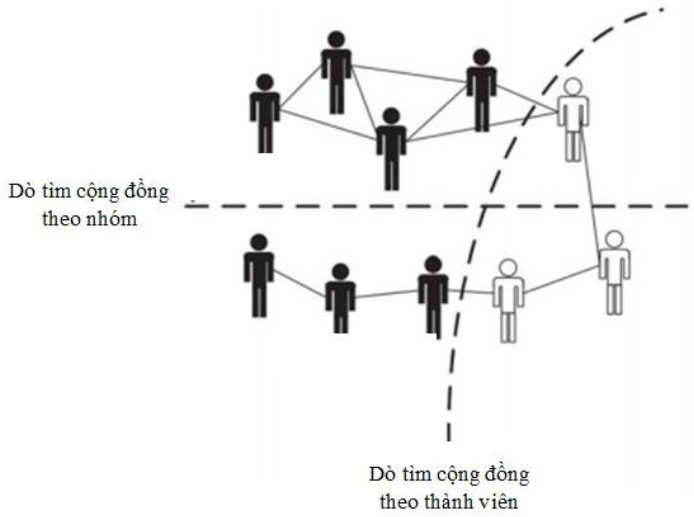


Hình 1: Biểu diễn cộng đồng bằng một graph

Khi khám phá cộng đồng, chúng ta thường quan tâm đến việc khám phá cộng đồng:

- Chứa một số thành viên cụ thể (1), hoặc
- Có dạng cộng đồng cụ thể nào đó (2).

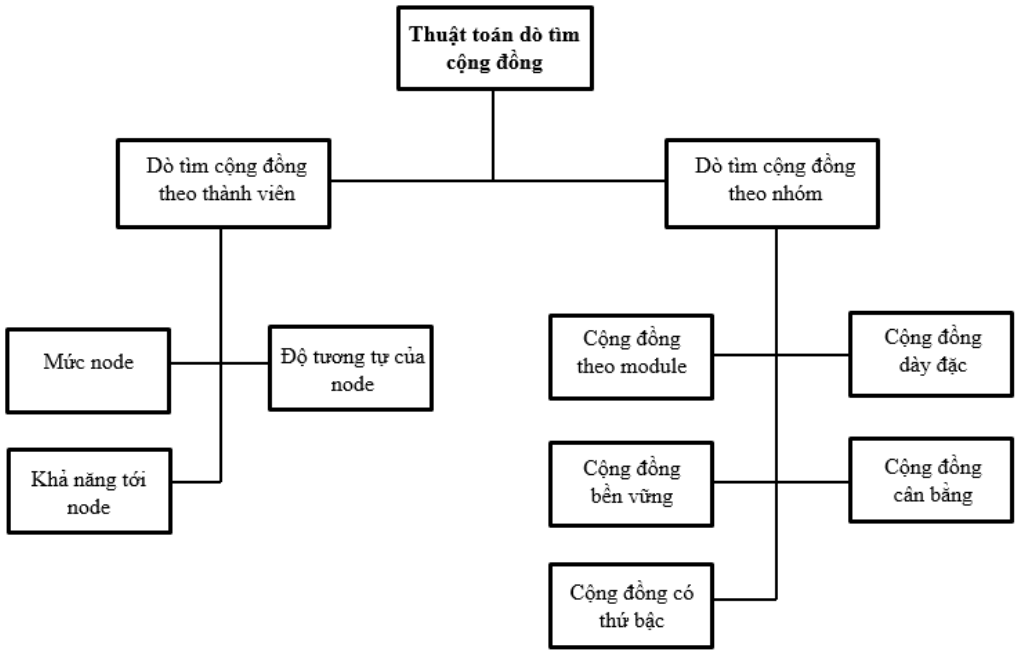
Khám phá cộng đồng theo (1) là phương thức *khám phá cộng đồng theo thành viên* còn theo (2) là *khám phá cộng đồng theo nhóm*. Hình 2 mô tả về hai cách khám phá này.



Hình 2: Phương thức khám phá cộng đồng

Theo hình 2, có thể tìm được hai cộng đồng dựa theo đặc tính thành viên là cộng đồng mặc áo trắng (3 người) và cộng đồng mặc áo đen (7 người); đồng thời cũng có thể phân thành 2 cộng đồng theo nhóm dựa trên mật độ liên hệ giữa các thành viên (cộng đồng phía trên và cộng đồng phía dưới)

Các thuật toán khám phá cộng đồng được phân loại như hình sau:



Hình 3: Các phương thức dò tìm cộng đồng

Sau đây, bài báo trình bày một số thuật toán khám phá cộng đồng dựa trên lý thuyết đồ thị hay được sử dụng hiện nay

2.1. Khám phá cộng đồng theo thành viên

Thuật toán khám phá cộng đồng theo thành viên thường được xây dựng dựa trên các thuộc tính của node như độ tương tự, mức độ, hoặc khả năng tiếp cận.

a) Thuật toán dựa trên mức node (Node Degree)

Với phương thức dò tìm cộng đồng này, chúng ta tìm những graph con, trong đó các node (hoặc nhóm các node) có mức node nhất định. Mức node ở đây là số lượng các cạnh vào hoặc ra một node. Trong thuật toán này, người ta sử dụng khái niệm Clique.

Clique là một đồ thị con đầy đủ trong đó tất cả các cặp node đều nối với nhau (mỗi node trong clique kết nối tới tất cả các node còn lại). Một clique có kích thước k là một đồ thị con trong đó tất cả các node có mức node là k-1. Để tìm các Clique trong một mạng nhỏ, người ta thường sử dụng thuật toán cưỡng chế (*brute-force*)

Algorithm 6.1 Brute-Force Clique Identification

Require: Adjacency Matrix A, Vertex v_x
1: **return** Maximal Clique C containing v_x
2: CliqueStack = $\{\{v_x\}\}$, Processed = {};
3: **while** CliqueStack not empty **do**
4: C=pop(CliqueStack); push(Processed,C);
5: v_{last} = Last node added to C;
6: $N(v_{last}) = \{v_i | A_{v_{last},v_i} = 1\}$.
7: **for all** $v_{temp} \in N(v_{last})$ **do**
8: **if** $C \bigcup \{v_{temp}\}$ is a clique **then**
9: push(CliqueStack, $C \bigcup \{v_{temp}\}$);
10: **end if**
11: **end for**
12: **end while**
13: Return the largest clique from Processed

Hình 4: Thuật toán xác định clique bằng brute-force

Phương pháp này có thể tìm thấy tất cả các clique tối đa trong một graph. Với mỗi đỉnh v_x, chúng ta sẽ tìm clique tối đa có chứa node v_x đó.

Thuật toán khởi đầu với một stack trống dùng để lưu clique. Stack này được khởi đầu với node v_x đang được phân tích (tức clique kích thước bằng 1), sau đó lấy ra một clique C chứa v_x. Node cuối cùng được add vào clique C là v_{last}. Các điểm lân cận của v_{last} sẽ được lần lượt thêm vào clique C, và nếu tạo thành một clique lớn hơn thì clique đó sẽ được đẩy vào stack. Quy trình này được thực hiện cho tới khi không thêm được node vào nữa.

Thuật toán brute-force không áp dụng được cho các mạng lớn. Đối với các mạng lớn, người ta điều chỉnh thuật toán bằng cách tỉa bớt các node và các cạnh. Chẳng hạn nếu clique được tìm kiếm có kích cỡ bằng k hoặc lớn hơn, thì chúng ta có thể giả sử rằng tất cả các node trong clique có mức node lớn hơn hoặc bằng k-1. Do vậy chúng ta có thể loại bỏ tất cả các node có mức node nhỏ hơn k-1 và các cạnh nối tới nó. Cách này giảm bớt tương đối khối lượng tính toán

Giả sử cộng đồng được hình thành từ một clique lõi, người ta sử dụng thuật toán clique lan truyền (*clique percolation metod- CPM*) để xác định cộng đồng.

Algorithm 6.2 Clique Percolation Method (CPM)

Require: parameter k

1: **return** Overlapping Communities
2: $Cliques_k$ = find all cliques of size k
3: Construct clique graph $G(V, E)$, where $|V| = |Cliques_k|$
4: $E = \{e_{ij} \mid$ clique i and clique j share $k-1$ nodes$\}$
5: Return all connected components of G

Hình 5: Thuật toán CPM

Thuật toán được trình bày trên hình Hình . Với số k cho trước, đầu tiên thuật toán sẽ tìm tất cả các clique có kích cỡ là k. Rồi một graph sẽ được tạo ra bằng cách biểu diễn các clique đó dưới dạng các node, từ đó hình thành nên cộng đồng.

b) Thuật toán dựa trên độ tương tự của node (Node Similarity)

Thuật toán độ tương tự node sẽ xác định độ tương tự giữa hai node v_i và v_j. Các node tương tự nhau (hoặc các node tương tự nhất) được cho là ở cùng trong một cộng đồng. Sau khi xác định được độ tương tự, thuật toán phân nhóm (clustering algorithm) sẽ được sử dụng để tìm ra các cộng đồng.

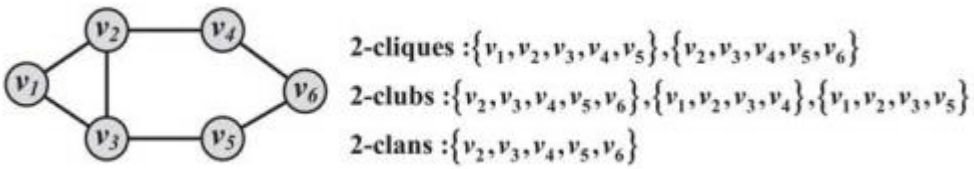
Nếu $N(v_i)$ và $N(v_j)$ tương ứng là lân cận của các đỉnh v_i và v_j thì độ tương tự giữa hai node có thể được xác định như sau:

$$\sigma(v_i, v_j) = |N(v_i) \cap N(v_j)|$$

c) Thuật toán dựa trên khả năng kết nối tới node (Node Reachability)

Khi tìm theo node reachability, chúng ta tìm những graph con (subgraph) trong đó các node có thể được kết nối tới các node khác thông qua một đường dẫn (path) nào đó.

- k-Clique là graph con đầy đủ trong đó đường dẫn ngắn nhất giữa hai node bất kỳ luôn luôn nhỏ hơn hoặc bằng k. Lưu ý rằng trong các k-clique, các node trên đường dẫn ngắn nhất đó không nhất thiết phải thuộc vào graph con đó.
- k-Club tương tự như k-clique nhưng định nghĩa chặt hơn. Các node trên đường dẫn ngắn nhất phải thuộc về subgraph đó.
- k-Clan là một k-clique, mà đối với tất cả các đường dẫn ngắn nhất trong subgraph, khoảng cách nhỏ hơn hoặc bằng k, $k - Clan = k - Clique \cap k - Club$

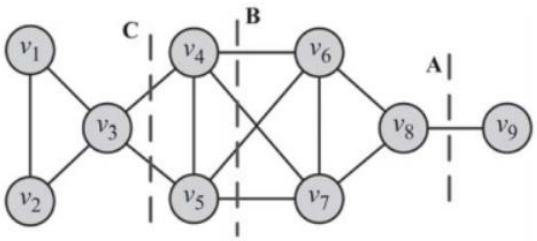


Hình 6: Minh họa k-Clique, k-Club và k-Clan

2.2. Khám phá cộng đồng theo nhóm

a) Khám phá cộng đồng cân bằng (balanced communities)

Trong việc phân vùng tìm cộng đồng trên graph, chúng ta cắt graph thành các phần nhỏ hơn (set/cutset) và dựa vào các đặc tính để xác định cộng đồng. Kích cỡ của một phép cắt là số lượng các cạnh bị cắt hoặc tổng trọng số của các cạnh đối với các graph có trọng số. *Phép cắt tối thiểu* (minimum cut/min-cut) là phép cắt mà kích cỡ của phép cắt là nhỏ nhất. Ví dụ ở Hình phép cắt B có kích cỡ là 4, A có kích cỡ 1 và là phép cắt tối thiểu.



Hình 7: Phép cắt tối thiểu và phép cắt cân bằng trong graph

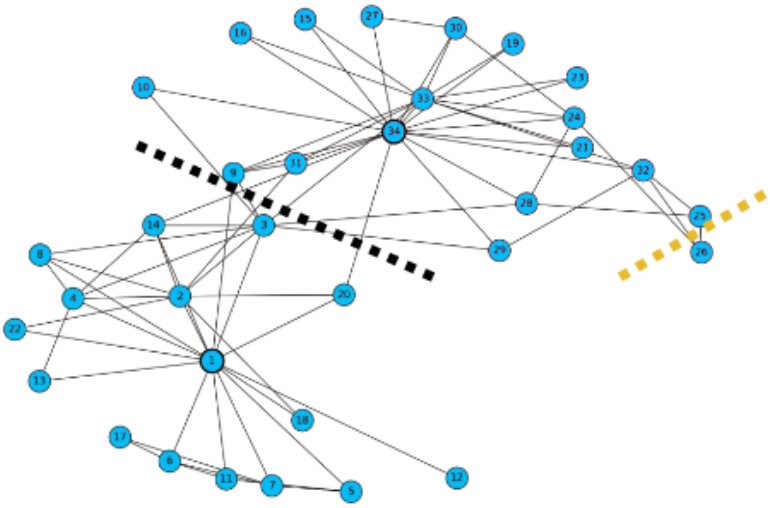
Mặc dù phép cắt tối thiểu có thể được thực hiện dễ dàng hơn nhưng phép cắt này không phải là phép cắt tối ưu để tìm cộng đồng trên graph vì thông thường nó cắt một node tách riêng ra khỏi phần còn lại của graph. Phép cắt như vậy tạo ra những phần không cân bằng của graph và không phản ánh cộng đồng thực tế. Chúng ta tìm kiếm những phép cắt cân bằng hơn.

Xem xét một graph $G(V, E)$, với việc chia nhỏ G thành k phần tạo thành một bộ $P = (P_1, P_2, \dots P_k)$, sao cho $P_i \subseteq V, P_i \cap P_j = \emptyset$ và $\bigcup_{i=1}^{k} P_i = V$. Khi đó hàm mục tiêu cho các phép cắt ratio cut và normalized cut được định nghĩa như sau:

$$Ratio\ Cut(P) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(P_i, \bar{P}_i)}{|P_i|},$$

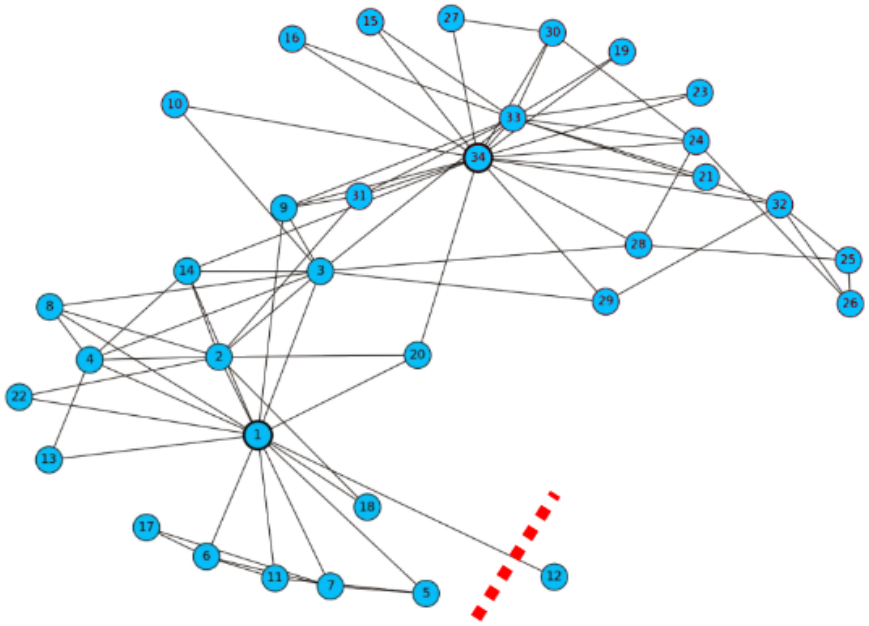
$$Normalized\ Cut(P) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(P_i, \bar{P}_i)}{vol(P_i)},$$

trong đó $\bar{P}_i = V - P_i$ là phần bù của cutset, $cut(P_i, \bar{P}_i)$ là kích cỡ của vết cắt, và $vol(P_i) = \sum_{v \in P_i} d_v$ là tổng mức của tất cả các node thuộc P_i. Cả hai hàm mục tiêu trên đều cho ra những cộng đồng kích cỡ cân bằng hơn bằng cách chuẩn hóa kích cỡ vết cắt bởi số lượng các đỉnh hoặc số lượng đỉnh có tính trọng số. Một ví dụ về normalized cut:



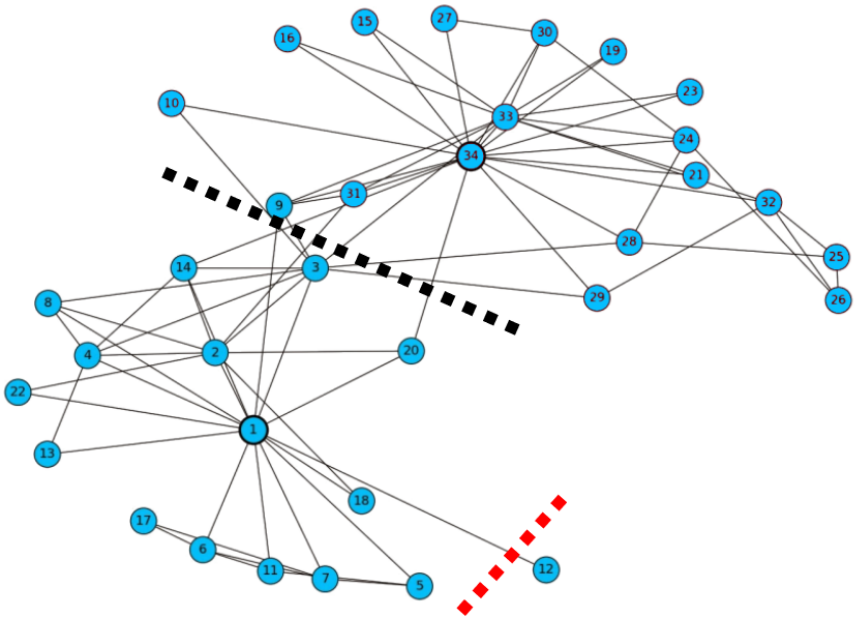
$$\text{Ratio Cut}(\cdots) = \frac{1}{2}(\frac{3}{33} + \frac{3}{1}) = 1.54545$$
$$\text{Ratio Cut}(\cdots) = \frac{1}{2}(\frac{9}{16} + \frac{9}{18}) = 0.53125$$

Tuy nhiên, có những trường hợp kết quả không chính xác:



$$\text{Ratio Cut}(\cdots) = \frac{1}{2}(\frac{1}{33} + \frac{1}{1}) = 0.51515$$

Bởi vậy, thay vì "cào bằng", coi các node là tương đương nhau trong cộng đồng, người ta đưa trọng số cho các node có ảnh hưởng lớn hơn.



$$\text{Norm. Cut}(\diagdown\!\diagup) = \frac{1}{2}\left(\frac{1}{155} + \frac{1}{1}\right) = 0.50322$$
$$\text{Norm. Cut}(\diagdown\!\diagup) = \frac{1}{2}\left(\frac{9}{76} + \frac{9}{80}\right) = 0.11546$$

Với phương pháp này, một phép chia tốt là phép phân chia mạng thành các cộng đồng sao cho có ít cạnh giữa các cộng đồng hơn là dự tính. Trên thực tế nếu số cạnh giữa các cộng đồng ít hơn rất nhiều so với dự tính thì kết quả thường không chính xác. Tính mô đun được tính bằng số lượng các cạnh trong các nhóm trừ đi số lượng tương ứng trong mạng khi các cạnh được đặt một cách ngẫu nhiên.

Biểu diễn toán học của tính mô đun:

$$Q = \sum_{k=1}^{K} (e_{kk} - a_k^2)$$

trong đó:

$$e_{kk} = \frac{\textit{số lượng cạnh có cả hai đầu nằm trong cộng đồng } k}{\textit{tổng số cạnh}},$$

$$a_k = \frac{\textit{số lượng các đầu mút nằm trong cộng đồng } k}{\textit{tổng số mút}}$$

chỉ số e_{kk} thể hiện thành phần các cạnh nằm trong cộng đồng k; a_k thể hiện thành phần nếu các cạnh được đặt ngẫu nhiên. Mục tiêu của phương pháp này là tối đa hóa Q để cho cấu trúc cộng đồng khác xa với việc sắp đặt ngẫu nhiên nhất.

b) Khám phá cộng đồng có thứ bậc

Các phương pháp trước xem xét các cộng đồng ở mức đơn lẻ. Trên thực tế chúng ta thường có các cộng đồng với kiến trúc có thứ bậc, tức là các cộng đồng trong đó có chứa các cộng đồng nhỏ và trong các cộng đồng nhỏ có thể chứa các cộng đồng nhỏ hơn nữa.

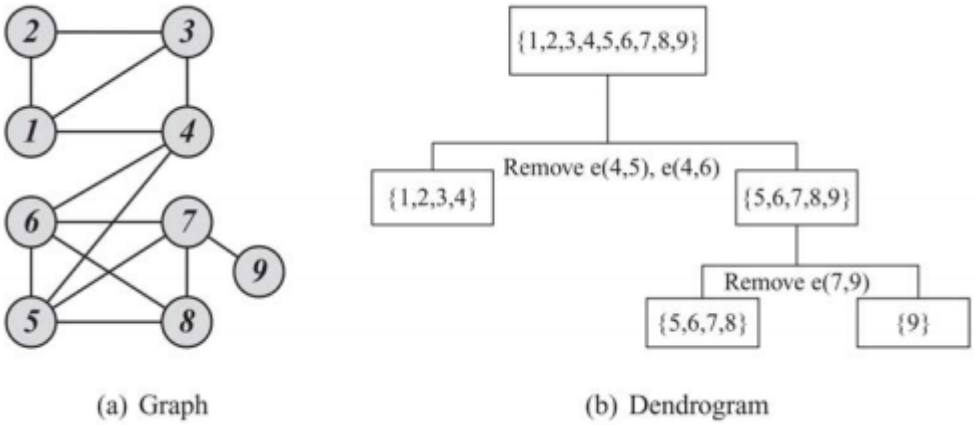
Phương pháp này thường khởi đầu với việc xem xét toàn bộ *n* node là 1 cộng đồng (chứa toàn bộ node) hoặc *n* cộng đồng (mỗi cộng đồng 1 node). Các cộng đồng

sau đó sẽ được sát nhập hoặc phân tách ra, tùy thuộc vào thuật toán được sử dụng. Thuật toán Girvan-Newman thường được sử dụng để phân nhóm cộng đồng có cấu trúc.

Thuật toán này giả sử rằng nếu một mạng có một tập các cộng đồng và những cộng đồng này được kết nối với nhau qua một số lượng không lớn các cạnh, thì các đường nối ngắn nhất giữa các thành viên của các cộng đồng khác nhau phải đi qua những cạnh này. Bằng cách loại bỏ những cạnh này, chúng ta có thể tìm ra các cộng đồng trong mạng. Để tìm ra những cạnh như thế, thuật toán Girvan-Newman sử dụng một chỉ số gọi là *edge betweenness* (độ xen giữa cạnh) và loại bỏ các cạnh với mức *edge betweenness* cao hơn.

Sơ lược các bước trong thuật toán Girvan-Newman biểu diễn như sau:

1. Tính toán *edge betweenness* cho tất cả các cạnh trong graph.
2. Loại bỏ cạnh có *edge betweenness* cao nhất.
3. Tính toán lại *edge betweenness* cho tất cả các cạnh bị ảnh hưởng bởi sự loại bỏ cạnh ở 2.
4. Lặp lại cho tới khi các cạnh bị loại bỏ.



Hình 8: Minh họa thực hiện Girvan-Newman

Trong cộng đồng có thứ bậc, chúng ta có thể có các độ đo mức độ trung tâm của một node như sau:

- *Mức độ trung tâm theo mức node (Degree centrality)*: Đây là cách tính mức độ trung tâm đơn giản nhất, được tính bằng mức của node, tức là số lượng các cạnh khởi đầu từ hoặc kết thúc ở node đó.
- *Mức trung tâm gần (Closeness centrality)*: được đo bằng khoảng cách trung bình ngắn nhất tới tất cả các node khác trong mạng.
- *Mức trung tâm xen giữa (Betweenness centrality)* của một node x được tính bằng số lượng các path ngắn nhất giữa hai node a và b bất kỳ trong graph đi qua node x.

- *Mức trung tâm Katz (Katz centrality)* được tính bằng tổng số các node có thể tới được từ một node cho trước, có trừ đi độ suy giảm cho các node ở xa.

3. KẾT LUẬN

Cộng đồng mạng xã hội đang được hình thành và phát triển và đầy biến động. Cộng đồng có ảnh hưởng và chi phối tới hành vi, thói quen của từng người dùng tham gia vào cộng đồng. Chính vì vậy, việc khám phá cộng đồng mạng xã hội từ nhiều nguồn dữ liệu khác nhau để từ đó nắm được, hiểu được những gì đang diễn ra trong cộng đồng là một việc có ý nghĩa. Đây cũng là một vấn đề đang thu hút nhiều nhà nghiên cứu hiện nay. Trong bài báo này, chúng tôi trình bày những nét phác thảo của một trong các hướng nghiên cứu tỏ ra hiệu quả để khám phá cộng đồng mạng xã hội là hướng nghiên cứu dựa trên cơ sở lý thuyết đồ thị. Việc ứng dụng các kết quả nghiên cứu theo hướng này đã cho những kết quả khả quan đồng thời cũng tiếp tục đặt ra các bài toán mới còn cần tiếp tục giải quyết .

TÀI LIỆU THAM KHẢO

1. *R. Zafarani, M. A. Abbasi và H. Liu, Social Media Mining - An Introduction, Cambridge University Press, 2014.*
2. *"CSE 190 - Data Mining and Predictive Analytics - Community Detection," [Trực tuyến]. Available: https://cseweb.ucsd.edu/~jmcauley/cse190/slides/week3/lecture6.pdf.*
3. *M. E. J. Newman, "Modularity and community structure in networks," 2006. [Trực tuyến]. Available: http://www.pnas.org/content/103/23/8577.full.pdf.*
4. *D. Hoffman và M. Fodor, "ResearchGate," 2010. [Trực tuyến]. Available: https://www.researchgate.net/publication/228237594_Can_You_Measure_the_ROI_of_Your_Social_Media_Marketing. [Đã truy cập 9 September 2016].*