NGHIÊN CỨU ỨNG DỤNG LÝ THUYẾT ĐỒ THỊ TRONG PHÂN TÍCH MẠNG XÃ HỘI

Nguyễn Xuân Hùng[1], Lê Đức Hậu[2]
[1] Khoa CNTT, trường Đại học Thủy lợi. Email: hungnx185@gmail.com

1. GIỚI THIỆU

Kể từ khi xuất hiện, các mạng xã hội (social network) như là MySpace, Facebook, Cyword đã thu hút hàng triệu người sử dụng. Đây là nơi mà người dùng có thể kết bạn, chia sẻ các thông tin cá nhân, suy nghĩ, cảm xúc về một chủ đề cùng quan tâm. Có rất nhiều định nghĩa mạng xã hội được đưa ra như: Mạng xã hội là các dịch vụ dựa trên web cho phép cá nhân hóa và có thể kết hợp danh sách người sử dụng với nhau [1]; Mạng xã hội được xem như thế giới ảo cho phép kết nối các thành viên có cùng sở thích lại với nhau và có nhiều tính năng phục vụ người dùng [2]; Mạng xã hội là một cấu trúc xã hội được tạo lên từ các cá nhân hoặc tổ chức được gọi là các "nút" được lên kết hoặc kết nối bởi một hoặc nhiều kiểu phụ thuộc với nhau như: quan hệ bạn bè, họ hàng, có chung quyền lợi, trao đổi tài chính, cùng vấn đề quan tâm, quan hệ tình cảm, hoặc quan hệ dựa trên sự tin tưởng, sử hiểu biết lẫn nhau [3]; Còn theo nhà xã hội học Laura Garton – nhà nghiên cứu chiến lược trường đại học Toronto thì "khi một mạng máy tính kết nối mọi người hoặc các cá nhân tổ chức lại với nhau thì đó chính là mạng xã hội". Trong bài báo này, chúng tôi coi mạng xã hội như một đồ thị trong đó các nút có thể là một cá thể, tổ chức, còn các cạnh của đồ thị là quan hệ bạn bè, gia đình, công việc.

Cụ thể, chúng tôi khai phá và phân tích mạng xã hội Facebook – đây được coi là mạng xã hội lớn nhất trong số các mạng xã hội hiện có. Do hạn chế về chính sách bảo mật và riêng tư của mạng xã hội này, nên chúng tôi chỉ khai phá được một phần của mạng dùng cho việc phân tích. Trong đó, công cụ khai phá sẽ thu thập dữ liệu liên quan quan hệ bạn bè giữa những người dùng. Các mối quan hệ này được trực quan hóa và biểu diễn dưới dạng đồ thị với các nút là các người dùng và các cạnh là các quan hệ bạn bè. Tiếp đó, chúng tôi phân tích các đặc tính của đồ thị này với các thông số như: phân bố độ bậc (Degree distribution) của nút để xác định mô hình tiến hóa của mạng, các độ trung tâm của nút như: độ trung tâm dựa trên bậc của nút, độ trung tâm ở giữa, độ trung tâm gần nhất, độ trung tâm dựa trên trị riêng. Thêm vào đó, việc phân tích cộng đồng cũng được thực hiện với các thuật toán tìm kiếm nhóm trên đồ thị. Mặc dù các chính sách về bảo mật và riêng tư chỉ cho phép một phần của mạng xã hội Facebook được khai phá, tuy nhiên với những thông số cấu trúc đã tính toán được, chúng tôi thấy rằng sự phát triển của mạng xã hội tuân theo mô hình "quy mô-tự do" (scale-free). Đây chính là mô hình mà phần lớn các mạng phức tạp khác như mạng xã hội, mạng sinh học, mạng viễn thông, vv… tuân theo. Ngoài ra, các kết quả phân tích và đánh giá tầm quan trọng của người dùng cũng như việc xác định được các cộng đồng dựa trên độ trung tâm và thuật toán tìm kiếm nhóm cũng phù hợp với thực tế của nhóm người dùng trong mạng được phân tích này. Điều này mở ra triển vọng áp dụng các phân tích này cho mạng khai phá được có kích thước lớn hơn và cho các mạng xã hội khác.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP

2.1. Dữ liệu

Có rất nhiều công cụ để khai phá mạng xã hội như: IPython Notebook, Natural Language Toolkit, NetworkX và nhiều công cụ tính toán khác [15]. Trong bài báo này, chúng tôi sử dụng IPython Notebook để khai phá dữ liệu trên mạng xã hội Facebook. IPython Notebook là một công cụ mạnh mẽ dựa trên ngôn ngữ lập trình Python. Sau khi dữ liệu khai phá với quyền tối đa được cấp, những dữ liệu này được biểu diễn dưới dạng đồ thị gồm 136 nút và 1489 cạnh, trong đó nút là tài khoản người dùng facebook và cạnh là các quan hệ bạn bè. Dữ liệu này sau đó sẽ được sử dụng trong việc phân tích các đặc tính cấu trúc dựa trên lý thuyết đồ thị.

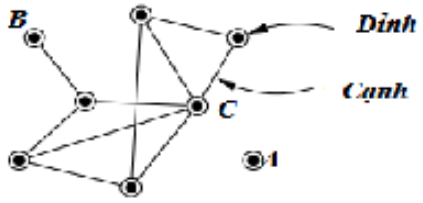
2.2. Phương pháp

Mạng xã hội được mô hình hóa như một đồ thị vô hướng G = (V, E) với V là được ký hiệu là các tập đỉnh (nút), E là tập cạnh (liên kết, quan hệ) (Hình 1). Việc phân tích mạng này được thực hiện

thông qua việc tính toán và đánh giá các thông số cấu trúc sau:

Phân bố độ bậc của nút (node degree distribution) P(k) là phân bố xác suất của bậc của các nút trong toàn mạng. Trong đó, bậc của nút (k) là số kết nối hay số cạnh của nút kết nối với nút khác. Nếu một mạng có tổng số n nút và n_k là số nút có độ bậc k, thì xác suất của một nút có bậc k là P(k) = n_k/n. Phân bố bậc của nút quyết định mô hình phát triển của mạng.

Theo [2] [4] là mạng ngẫu nhiên hay theo quy luật gắn kết ưu tiên (preferential attachment). Khi phát triển theo quy luật gắn kết ưu tiên, mạng đó sẽ có đặc tính "quy mô-tự do" (scale-free). Trong mạng này có một số ít các nút có bậc lớn hơn rất nhiều độ bậc trung bình của mạng. Các nút có bậc cao này thường được gọi là trung tâm (hub). Do đó, với một mạng scale-free, phân bố bậc của nút sẽ tuân theo hoặc tiệm cận hàm mũ: $P(k) \sim k^{-\gamma}$ với γ là tham số nằm trong khoảng $2<\gamma<3$.



Hình 1. *Một đồ thị với 8 đỉnh và 10 cạnh*

Độ trung tâm (centrality) của một nút trong mạng được sử dụng đánh giá tầm quan trọng của một người dùng trên mạng xã hội [5] [6]. Có nhiều độ trung tâm của nút đã được định nghĩa, bao gồm:

Độ trung tâm dựa trên bậc của nút (degree centrality). Đây là độ đo chỉ dựa vào bậc của nút và được xác định bằng số cạnh hay số kết nối mà một nút có. Công thức tính như sau: $C_D(V)=deg(v)$. Với deg(v) là số cạnh mà nút đó có.

Độ trung tâm ở giữa (betweenness centrality) chỉ ra vị trí cả nút trong mạng và khả năng kết nối các thành phần hoặc nhóm trong mạng. Nó được tính bằng $C_B(v) = \Sigma_{s\neq v\neq t\in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$ với σ_{st} là tổng số đường đi ngắn nhất từ s tới t. $\sigma_{st}(v)$ là tổng số đường đi ngắn nhất từ s tới t mà đi qua nút v.

Độ trung tâm dựa trên sự gần gũi (closeness centrality) chỉ ra một nút trong mạng có thể truy cập nhanh tới nhiều nút khác trong mạng. Nó được tính bằng $C_c(v) = \frac{\Sigma_{tc/v} d_c(vt)}{n-t}$ với $S_{t\in V\setminus v} d_c(v,t)$ là tổng số đường đi ngắn nhất từ nút v tới các nút còn lại, n là số nút trong mạng.

Độ trung tâm dựa trên giá trị riêng (eigenvalue centrality) [8] chỉ ra nút đó có độ trung tâm nhất trong toàn mạng. Nó được tính bằng $x_v = \frac{1}{\lambda}\Sigma_{t\in M(v)} x_t = \frac{1}{\lambda}\Sigma_{t=1}^{n} A_{vt}\ x_t$ với $A_{v,t}$ là ma trần kề của mạng, M(v) là tập nút mà được kết nối tới nút thứ v, n là tổng số nút và λ là một hằng số. Với nút có trị số đặc trưng cao chỉ ra nút đó có độ trung tâm cao hơn các thực thể khác và đây cũng là thước đo lợi thế về vị trí.

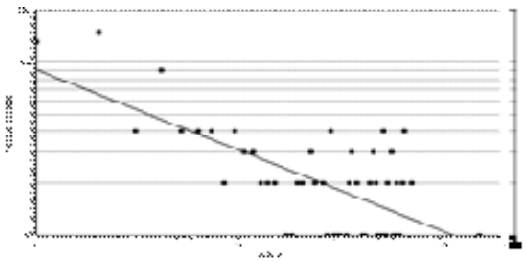
Trong phân tích đồ thị, ngoài việc đánh giá các thông số liên quan tới nút, chúng tôi cũng quan tâm tới đặc tính nhóm/cộng đồng (cluster/community) của các nút trên đồ thị. Các nút trong cùng một nhóm sẽ có một số đặc trưng giống nhau. Đây cũng là một đặc tính khá phổ biến trong mạng thực tế. Ví dụ, với mạng xã hội, các nhóm/cộng đồng được hình thành dựa trên yếu tố vùng miền, các vấn đề quan tâm, yêu thích, công việc… [9]. Có rất nhiều thuật toán để có thể tìm kiếm các cộng đồng trên mạng (community detection). Thuật toán được sử dụng phổ biến nhất được đề xuất bởi Girvan-Newman. Ngoài ra còn có phương pháp Module cực đại, tìm kiếm dựa trên nhóm (clique) lớn nhất. Một thuật toán được biết tới như một phương thức tự động tìm kiếm những đồ thị con có tính liên kết cao (coi như các cộng đồng) là thuật toán MCODE. Thuật toán này có hiệu năng tính toán cao đặc biệt cho các mạng có kích thước lớn như mạng xã hội. Trong nghiên cứu này, chúng tôi sẽ sử dụng MCODE để tìm các cộng đồng trên mạng đã khai phá được. Cụ thể, thuật toán này thực thi theo ba bước chính như sau: B1: Gán điểm số cho các nút, nút nào có nhiều kết nối với các nút khác có điểm số cao hơn; B2: Tìm kiếm từ nút hạt giống (nút có điểm số cao nhất) và lặp lại quá trình này đi ra các nút là hàng xóm của các nút hạt giống, loạt bỏ các nút dưới ngưỡng cho phép. Quá trình này được lặp lại và cuối cùng cộng đồng được lọc ra chứa ít nhất k phần tử. B3: Cắt tỉa và làm mịn, loại bỏ đi những nút có kết nối đơn.

3. KẾT QUẢ

Việc tính toán các thông số cấu trúc của mạng được thực hiện bởi phần mềm mã nguồn mở Cytoscape. Đây là một công cụ rất mạnh dùng cho việc trực quan hóa và phân tích các mạng phức tạp nói chung.

3.1. Phân bố bậc của nút

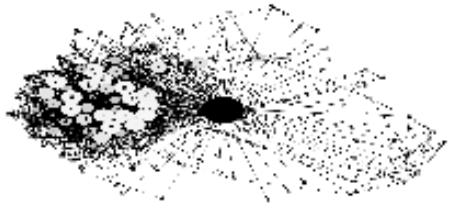
Hình 2 cho thấy rằng sự phân bố độ bậc của nút, sự phân bố này tuân theo hàm mũ (đường kẻ thẳng trong biểu đồ). Điều này chứng tỏ mạng tuân theo quy luật gắn kết ưu tiên và mạng sẽ chứa các nút có bậc cao hơn mức trung bình và các nút này là những nút trung tâm.



Hình 2. *Sự phân bố độ bậc của nút (node degree distribution)*

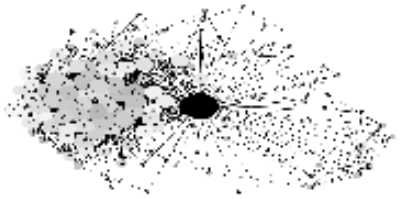
3.2. Độ trung tâm

Độ trung tâm của từng nút được tính toán và trực quan hóa trên Hình 3 và Hình 4 trong đó, các nút có kích thước lớn và màu đỏ biểu thị rằng nút đó có độ trung tâm lớn, các nút có màu vàng, màu xanh với độ lớn nhỏ hơn tương ứng với các nút có độ trung tâm thấp hơn. Hình 3 cho biết người dùng 85 có độ trung tâm ở giữa lớn nhất (màu đỏ) và chứng tỏ người dùng này có kết nối trong mạng là lớn nhất. Trên mạng facebook thực tế mà chúng tôi khai phá được, người dùng 85 này là bạn bè của hầu hết người dùng còn lại trong mạng.



Hình 3. *Trực quan hóa độ trung tâm của các nút theo độ trung tâm ở giữa (betweeness centrality).*

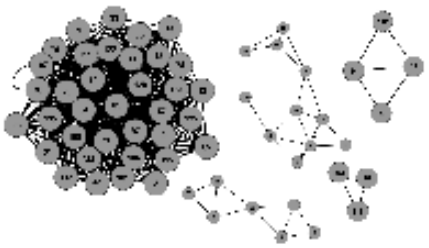
Hình 4 cho chúng ta thấy rằng người dùng 85 có chỉ số độ trung tâm dựa trên sự gần gũi cao nhất (màu đỏ) chúng tỏ khả năng truyền thông tin đến toàn bộ thành viên trong mạng là nhanh nhất và có ảnh hưởng lớn đến mạng khi có sự việc xảy ra. Trong mạng facebook, khi người dùng 85 đưa một thông tin lên mạng, hầu hết các thành viên trong mạng đều có thể tiếp cận, bình luận trao đổi thông tin với các thành viên còn lại trong mạng rất dễ dàng.



Hình 4: *Trực quan hóa độ trung tâm của các nút theo độ trung tâm dựa trên sự gần gũi (closeness centrality).*

3.3. Cộng đồng

Hình 5 là kết quả sau khi thực thi thuật toán MCODE để tìm kiếm cộng đồng trong mạng, với mạng này thuật toán đưa ra được 5 cộng đồng, trong các cộng đồng các thành viên kết nối với nhau chặt chẽ. Trên mạng thực tế, các thành viên trong các cộng đồng này thường xuyên trao đổi thông tin, chia sẻ hình ảnh với nhau hơn hẳn các thành viên khác.



Hình 5. *Kết quả tìm kiếm cộng đồng với MCODE. Có năm cộng đồ được xác định trong mạng.*

4. KẾT LUẬN

Việc phân tích các mạng phức tạp nói chung và các mạng xã hội nói riêng là một lĩnh vực quan trọng. Gần đây lý thuyết đồ thị được ứng dụng nhiều để phân tích các mạng này. Trong bài báo này, chúng tôi đã sử dụng một số thông số cấu trúc như: phân bố bậc của nút, độ trung tâm, tính cộng đồng để phân tích một mạng được khai phá từ mạng xã hội Facebook. Dựa vào các chỉ số này, chúng ta đánh giá được mô hình phát triển của mạng, tầm trọng của từng thành viên và sự kết nối theo nhóm/cộng đồng của thành viên trong mạng. Mặc dù còn nhiều hạn chế về kích thước mạng khai phá được do các chính sách về bảo mật và riêng tư của nhà quản lý mạng xã hội, tuy nhiên, các kết quả tính toán và phân tích được đã chỉ ra tiềm năng của việc ứng dụng lý thuyết đồ thị trong phân tích mạng xã hội khi có nhiều thuật toán đã và đang được phát triển cho việc phân tích đồ thị.

TÀI LIỆU THAM KHẢO

[1]. N. b. Danah M.Boyd, "Social Network Sites: Definition, History, and Scholarship," Journal of Computer - Mediated Communication 13 , pp. 210-230, 2008.

[2]. "Bách khoa toàn thư mở Wikipedia," 10 2013. [Online]. Available: http://vi.wikipedia.org. [Accessed Octorber 2013].

[3]. S. Wasserman and K. Faust, "Social Network Analysis: Methods and Applications," in Social Network Analysis: Methods and Applications, ISBN-13: 978-0521387071 ISBN-10: 0521387078, 1994.

[4]. M. S. Oliver Hein and W. Konig, "Scale-Free Network - The impact of fat tailed degree distribution of diffusion and communication processes," in Wirtscharftsinformatik 48 (2006), p.267-275, 2006.