TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN TOÁN TỔ HỢP VÀ ĐỒ THỊ**

**CÁC ỨNG DỤNG CỦA LÝ THUYẾT ĐỒ THỊ TRONG MÔ HÌNH MẠNG XÃ HỘI**

*Người hướng dẫn*: **BÙI DANH HƯỜNG**

*Người thực hiện*: **HOÀNG KIẾN THIẾT – 51702187**

Lớp **: 17050202**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN TOÁN TỔ HỢP VÀ ĐỒ THỊ**

**CÁC ỨNG DỤNG CỦA LÝ THUYẾT ĐỒ THỊ TRONG MÔ HÌNH MẠNG XÃ HỘI**

Người hướng dẫn: **BÙI DANH HƯỜNG**

Người thực hiện: **HOÀNG KIẾN THIẾT**

Lớp **: 17050202**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Đầu tiên em xin gửi lời cảm ơn này đến gia đình và người thân, họ đã luôn là động lực giúp em không ngừng học tập và phát triển.

Trong quá trình nghiên cứu về đề tài này, các khó khăn em gặp phải như việc tìm và đọc tài liệu tiếng anh, tự bổ sung kiến thức về lý thuyết đồ thị hay hiểu và vận dụng các thuật toán trong mô hình hóa mạng xã hội đều đã được sự hỗ trợ từ thầy Bùi Danh Hường qua các buổi meeting online của thầy. Em xin chân thành cảm ơn thầy Bùi Danh Hường đã tận tâm giảng bài và giải đáp các thắc mắc, giúp em có thêm nguồn kiến thức để hoàn thiện bài tập cuối kì này. Cám ơn trường Đại học Tôn Đức Thắng vì đã tạo điều kiện cho sinh viên tiếp cận với môi trường học tập, nghiên cứu hiện đại, đặc biệt trong mùa dịch Covid-19 sinh viên phải học online thay cho hình thức học tập trung tại trường nhưng vẫn đảm bảo cung cấp nguồn tham khảo cần thiết cho đề tài ứng dụng của lý thuyết đồ thị trong mô hình mạng xã hội cũng như những tài liệu khác liên quan đến môn học Toán tổ hợp và đồ thị.

**BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm báo cáo của riêng tôi. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong báo cáo còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung tiểu luận của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện.

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Hoàng Kiến Thiết*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn ( nếu có )**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Bài báo cáo này sẽ trình bày quá trình nghiên cứu, tổng hợp kiến thức liên quan đến đề tài các ứng dụng của lý thuyết đồ thị trong mô hình mạng xã hội.

Để tiếp cận và hiểu hơn về đề tài, em đã tham gia các buổi meeting vào 20h tối thứ 2 hàng tuần của thầy Danh Hường. Từ đó, em có cái nhìn tổng quan hơn về đề tài này và có thể hệ thống kiến thức môn học Toán tổ hợp và Đồ thị để sử dụng cho việc nghiên cứu đề tài cuối kì. Song song với đó, em tìm đọc các bài báo, các bài nghiên cứu khoa học hay sách có liên quan đến đề tài các ứng dụng của lý thuyết đồ thị trong mô hình mạng xã hội. Từ đó em tổng hợp lại và vận dụng kiến thức của mình trong môn học Toán tổ hợp và Đồ thị để hoàn thành bài báo cáo cuối kì này.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc41249005)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc41249006)

[TÓM TẮT iv](#_Toc41249007)

[MỤC LỤC 1](#_Toc41249008)

[CHƯƠNG 1 – MẠNG XÃ HỘI VÀ CÁC VẤN ĐỀ PHÂN TÍCH MẠNG XÃ HỘI 3](#_Toc41249009)

[1. Mạng xã hội là gì? 3](#_Toc41249010)

[2. Tầm quan trọng của mạng xã hội và vấn đề phân tích mạng xã hội 3](#_Toc41249011)

[3. Mô hình hóa mạng xã hội bằng đồ thị 4](#_Toc41249012)

[CHƯƠNG 2 – CÁC PHƯƠNG PHÁP ĐO LƯỜNG TRONG MÔ HÌNH HÓA MẠNG XÃ HỘI 6](#_Toc41249013)

[2.1 Một vài khái niệm cơ bản. 6](#_Toc41249014)

[2.1.1 Đồ thị và các thành phần cơ bản của đồ thị 6](#_Toc41249015)

[2.1.2 Tỉ trọng, mật độ của đồ thị và các phương pháp tính 6](#_Toc41249016)

[2.1.2.1 Density là gì 7](#_Toc41249017)

[2.1.2.2 Tính trung tâm về mức độ (degree centrality) 8](#_Toc41249018)

[2.1.2.3 Tính trung tâm về độ giữa (betweenness centrality): 9](#_Toc41249019)

[2.1.2.4 Tính trung tâm về độ gần gũi (closeness centrality): 10](#_Toc41249020)

[2.1.2.5 Tính trung tâm phân cụm (clustering centrality): 10](#_Toc41249021)

[2.2 Key players và cách tìm key players 11](#_Toc41249022)

[2.2.1 Key players và các vấn đề xoay quanh key players 11](#_Toc41249023)

[2.2.2 Làm thế nào để phát hiện và duy trì key players 13](#_Toc41249024)

[2.3 Đồ thị signed và các vấn đề, ứng dụng của đồ thị signed trong mô hình hóa mạng xã hội 15](#_Toc41249025)

[2.3.1 Đồ thị signed và các vấn đề xoay quanh đồ thị signed 15](#_Toc41249026)

[2.3.1 Ứng dụng của đồ thị signed trong mô hình hóa mạng xã hội 16](#_Toc41249027)

[CHƯƠNG 3 – PHÁT HIỆN CỘNG ĐỒNG TRONG MẠNG XÃ HỘI 17](#_Toc41249028)

[3.1 Phát hiện cộng đồng là gì và ứng dụng của nó trong mạng xã hội 17](#_Toc41249029)

[3.2 Thuật toán module hóa và thuật toán cắt 19](#_Toc41249030)

[3.2.1 Thuật toán tối đa hóa module 19](#_Toc41249031)

[3.2.2 Thuật toán cắt 20](#_Toc41249032)

[3.3 Thuật toán Grivan-Newman 20](#_Toc41249033)

[3.4 Thuật toán Node similarity based 22](#_Toc41249034)

[3.5 Thuật toán Label Propagation Community Detection (LPA) 23](#_Toc41249035)

[CHƯƠNG 4 – MINH HỌA MỘT SỐ THUẬT TOÁN 24](#_Toc41249036)

[4.1 Import thư viện và lấy input 24](#_Toc41249037)

[4.2 Demo thuật toán 25](#_Toc41249038)

[4.2.1 Minh họa thuật toán Degree Centrality 25](#_Toc41249039)

[4.2.2 Minh họa thuật toán Closeness Centrality 26](#_Toc41249040)

[4.2.3 Minh họa thuật toán Betweeness Centrality 27](#_Toc41249041)

[4.2.4 Minh họa thuật toán Clusterring Centrality 28](#_Toc41249042)

[4.2.5 Minh họa thuật toán tìm key player 30](#_Toc41249043)

[4.3 Thực thi source code và kết quả của source code 31](#_Toc41249044)

[CHƯƠNG 5 – KẾT LUẬN 32](#_Toc41249045)

CHƯƠNG 1 – MẠNG XÃ HỘI VÀ CÁC VẤN ĐỀ PHÂN TÍCH MẠNG XÃ HỘI

1. Mạng xã hội là gì?

Mạng xã hội là một dịch vụ cho phép kết nối những người dùng với nhau qua Internet. Các người dùng có chung nhiều mục đích sử dụng mạng xã hội như chat, e-mail, chia sẻ phim ảnh, blog, các thông tin xã luận hay voice chat với nhau không phân biệt không gian, thời gian. Những người dùng này được gọi là cư dân mạng xã hội và được liên kết với nhau qua nhiều phương diện như bạn bè, thành viên nhóm, đối tác, thành viên trong gia đình và nhiều hơn thế nữa. Hiện nay trên thế giới có hàng trăm dịch vụ mạng xã hội khác nhau, nổi tiếng phải kể đến Facebook được dùng rộng rãi trong nước ta và thị trường Mỹ, Tây Âu.

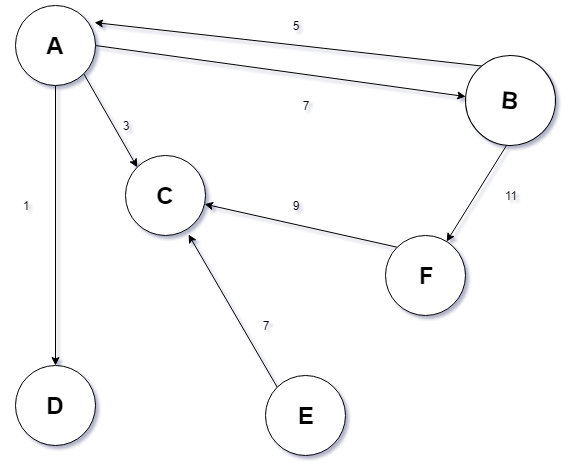
2. Tầm quan trọng của mạng xã hội và vấn đề phân tích mạng xã hội

Mạng xã hội đang rất phổ biến và có tầm quan trọng to lớn trong xã hội ngày nay. Các mạng xã hội trực tuyến như Facebook, Twitter, YouTube đã trở thành một trong những kênh truyền thông phổ biến để truyền tải tin tức, tài liệu, quảng cáo và tiếp thị đến người dùng. Do đó phân tích mạng xã hội đang là một lĩnh vực nghiên cứu cực kỳ quan trọng.

Trong những năm gần đây, các nghiên cứu trong lĩnh vực mạng xã hội lớn thu hút được nhiều sự chú ý từ cộng đồng khoa học. Một trong những lĩnh vực nghiên cứu quan trọng nhất trong lĩnh vực này là mô hình hóa mạng xã hội. Mô hình hóa mạng xã hội giúp ta hiểu được các đặc điểm cấu trúc và các chức năng của một mạng xã hội từ quy mô nhỏ đến lớn và phức tạp. Ngoài ra, mô hình hóa mạng xã hội sẽ giúp ta dễ quản lý, kiểm soát quá trình phát triển của mạng xã hội và các mảng lớn, nhỏ bên trong nó. Ta có thể mô hình hóa mạng xã hội bằng các đồ thị, trong đó một mạng xã hội là một đồ thị với nhóm rất nhiều các đỉnh được liên kết chặt chẽ với các đỉnh khác bên trong các đồ thị con và các đỉnh còn lại bên ngoài.

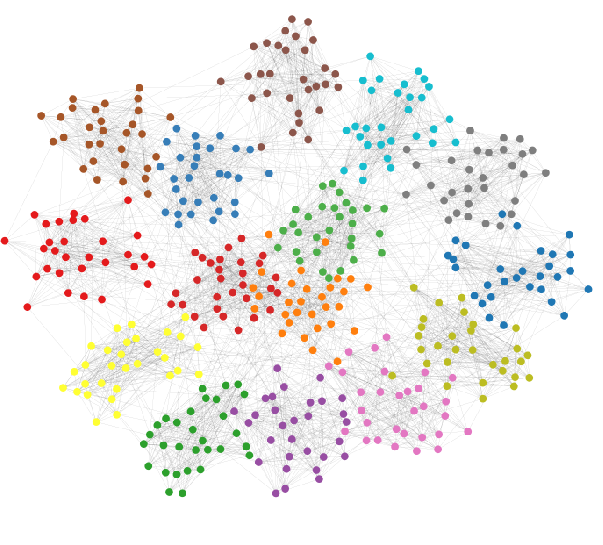
3. Mô hình hóa mạng xã hội bằng đồ thị

Một mạng xã hội được biểu diễn một như một đồ thị 𝐺 (𝑉, 𝐸), trong đó 𝑉 là các đỉnh trong đồ thị G biểu diễn các đối tượng như người dùng, nhóm người dùng và 𝐸 là các cạnh trong đồ thị G biểu diễn các quan hệ hoặc sự liên kết giữa các người dùng. Cùng với đó là các thuật toán phát hiện cộng đồng như thuật toán Odularity và cắt, thuật toán Grivan Newman, thuật toán dựa trên tương tự nút hay thuật toán LPA để xây dựng nên một đồ thị biểu diễn cho một mạng xã hội to phức tạp. Các đồ thị vô hướng cũng được dùng để biểu diễn sự tương tác giữa các đối tượng là các đỉnh. Để lấy ví dụ minh họa, ta xét đồ thị sau :



Hình 1.1 Đồ thị minh họa một phần nhỏ của mạng xã hội

Trong đồ thị trên, các đỉnh A, B, C, D, E, F minh họa cho các đối tượng trong một mạng xã hội như người dùng, nhóm người dùng. Các cạch có hướng biểu diễn cho một sự liên kết, tương tác nào đó giữa các đối tượng như nhắn tin, kết bạn hay cùng thả tim cho một đối tượng khác. Các trọng số của các cạnh này thể hiện cho cường độ của sự tương tác hay các mối quan hệ trên cạnh tương ứng. Từ đây ta có thể dễ dàng kiểm soát, quản lý cũng như phát triển các đối tượng trong một mạng xã hội hơn so với khi chưa có mô hình này. Hơn nữa ta có thể kiểm soát các luồng thông tin từ mạng xã hội, thúc đẩy giao tiếp giữa các đối tượng với nhau và chia sẻ rộng rãi hơn ra cộng đồng người dùng mạng xã hội. Rõ ràng quá trình mô hình hóa mạng xã hội bằng đồ thị này giúp ích rất nhiều và hầu như không có nhược điểm. Trên thực tế, mạng xã hội rất to lớn và phức tạp, để mô hình hóa được một mạng xã hội hay một phần của nó thôi cũng tốn nhiều công sức, nhưng hiệu quả đạt được thì rất nhiều như em đã nói ở trên.



Hình 1.2 Mô hình hóa mạng xã hội bằng đồ thị

Nguồn: [paperswithcode.com]

Trong quá trình nghiên cứu về đề tài mô hình hóa mạng xã hội, em nhận thấy đây là một lĩnh vực rất rộng lớn và phức tạp. Trọng tâm của bài báo cáo chỉ đề cập đến hai mảng lớn là các phương pháp đo lường và mảng phát hiện các cộng đồng trong lĩnh vực mô hình hóa mạng xã hội.

CHƯƠNG 2 – CÁC PHƯƠNG PHÁP ĐO LƯỜNG TRONG MÔ HÌNH HÓA MẠNG XÃ HỘI

Trong chương này em sẽ tìm hiểu cách mô hình hóa mạng xã hội bằng đồ thị thông qua các khái niệm cơ bản của một đồ thị. Cùng với đó là các khái niệm về density, bậc và các loại của mật độ. Khái niệm về key players và cách phát hiện key players. Và cuối cùng là định nghĩa được một đồ thị mô hình hóa cho mạng xã hội cùng với các vấn đề và ứng dụng của nó.

2.1 Một vài khái niệm cơ bản.

2.1.1 Đồ thị và các thành phần cơ bản của đồ thị

Một mạng xã hội được biểu diễn một như một đồ thị 𝐺 (𝑉, 𝐸), trong đó 𝑉 là các đỉnh trong đồ thị G biểu diễn các đối tượng như người dùng, nhóm người dùng và 𝐸 là các cạnh trong đồ thị G biểu diễn các quan hệ hoặc sự liên kết giữa các người dùng. Những sự liên kết này được biểu diễn thông qua các ma trận kề , trong đó = 0 biểu thị (𝑣𝑖, 𝑣𝑗) ∈ 𝐸, và 0 biểu thị cho (𝑣𝑖, 𝑣𝑗) 𝐸. Bậc của các nút 𝑣𝑖 là d𝑖. Nếu các cạnh giữa các nút là các cạnh có hướng các bậc tương ứng được ký hiệu lần lượt là 𝑑− và 𝑑 +. Số lượng đỉnh và cạnh của một mạng xã hội biểu diễn theo đồ thị lần lượt là ∣V∣ = 𝑛 và ∣E∣ = m. Đường đi ngắn nhất giữa một cặp nút và được gọi là trắc địa và khoảng cách trắc địa giữa hai nút được gọi là 𝑑(𝑖,j). Đồ thị G bao gồm rất nhiều đồ thị con trong đó, các đồ thị con này kí hiệu là Gs(𝑉𝑠, 𝐸𝑠).

2.1.2 Tỉ trọng, mật độ của đồ thị và các phương pháp tính

Trong lý thuyết đồ thị, đồ thị dày là đồ thị có số cạnh gần bằng với số cạnh tối đa của đồ thị đó. Ngược lại, một đồ thị chỉ có một vài cạnh là một đồ thị thưa. Sự phân biệt giữa các biểu đồ thưa và dày là khá mơ hồ, và phụ thuộc vào bối cảnh mà ta định nghĩa chúng. Do đó, các khái niệm về tỉ trọng và mật độ của đồ thị được định nghĩa giúp cho việc sử dụng đồ thị trong các ứng được thuận tiện hơn.

2.1.2.1 Density là gì

Đặt 𝐶 là biểu đồ con của đồ thị G, mật độ của biểu đồ con 𝐶 được tính theo công thức:

density(C) = với

Từ công thức trên, ta định nghĩa mật độ của đồ thị của đồ thị là tỷ số giữa số cạnh | E | đối với số cạnh tối đa có thể có của đồ thị đó.

Đối với các đồ thị vô hướng, mật độ của đồ thị là: = =

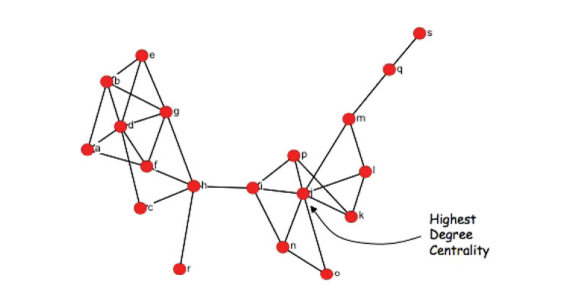
Còn đối với các đồ thị có hướng, số cạnh tối đa có thể nhiều gấp đôi so với các đồ thị vô hướng, do đó mật độ được tính bởi : = = . Trong đó là số cạnh tối đa của một đồ thị vô hướng. Khi đó tỷ trọng tối đa của đồ thị là 1 và tối thiểu là 0.

Để nói về mật độ của đồ thị, ta có các khái niệm về tính trung tâm trong một mạng xã hội được Freeman xây dựng trong những năm 1978, 1979 và được phát triển bởi De Nooy, Mrvar, và Batagelj trong cuốn Hanneman & Riddle xuất bản năm 2005. Cụ thể:

* Tính trung tâm về mức độ (degree centrality): Các thông tin trong và ngoài mỗi nút được xem như một trung tâm.
* Tính trung tâm về sự gần gũi (closeness centrality): Là khoảng cách của một đối tượng trong mạng xã hội so với tất cả các đối tượng khác trong mạng xã hội đó.
* Tính trung tâm về mặt giữa (betweenness centrality): Là mức độ đối tượng được định vị trên con đường ngắn nhất (trắc địa) giữa các cặp đối tượng khác trong mạng xã hội.
* Tính trung tâm phân cụm (clustering centrality): Là tỷ lệ của các mối quan hệ, ràng buộc các một đỉnh trong mạng xã hội so với nhau.

2.1.2.2 Tính trung tâm về mức độ (degree centrality)

Tính trung tâm về mức độ là số lượng các mối quan hệ, liên kết hay sự ràng buộc mà một nút hay một đối tượng trong mạng xã hội đang sỡ hữu. Bậc của các nút này được xác định thông qua các mối quan hệ, liên kết đi (indegree) và đến (outdegree) giữa các nút với nhau trong một mạng xã hội. Em lấy ví dụ trong trường hợp của một đồ thị các mối quan hệ giữa một status trên mạng xã hội với dòng status đó là một nút, những người dùng tương tác với status đó (ví dụ như thả tim, bình luận hay chia sẻ) được gọi là outdegree của nút status, và người dùng đăng dòng status đó được gọi là indegree của nó. Các mức độ trung tâm này thường được chuẩn hóa phương pháp tính theo phần trăm độ trong một mạng xã hội mà ta đang xét.



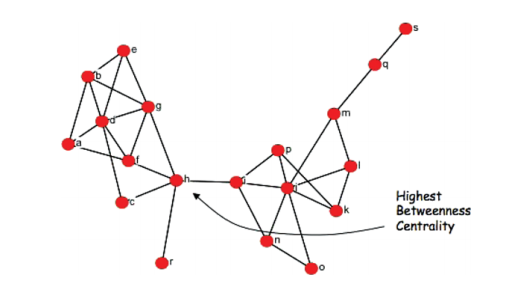
Hình 2.1 Minh họa Degree Centrality

Nguồn: (Borgatti, ARTS lecture 2009)

2.1.2.3 Tính trung tâm về độ giữa (betweenness centrality):

Mức độ giữa là thước đo mức độ thường xuyên xuất hiện của một đỉnh trong đồ thị nằm trên đường đi ngắn nhất (trắc địa) qua các đỉnh khác trong đồ thị. Ứng dụng của việc tính mức độ giữa này là ta có thể tính toán được khả năng của một nút có thể hoạt động như một đỉnh mấu chốt, đỉnh liên kết nhiều đỉnh khác hay không, từ đó giúp ta kiểm soát sự giao tiếp giữa các nút trong đồ thị. Nếu một nút có mức độ giữa cao mà bị xóa khỏi một mạng, mạng sẽ rơi vào trạng thái phân cụm (clustering). Không giống như trung tâm về mức độ (degree centrality), mức độ giữa là một phép tính và được chuẩn hóa theo tỷ lệ của tất cả các trắc địa bao gồm các đỉnh đang xét. Nếu được định nghĩa là số lượng đường trắc địa giữa i và j, và là số lượng các trắc địa đi qua k, mức độ tính trung tâm giữa được Farrall định nghĩa vào năm 2005 như sau:

với i j k.

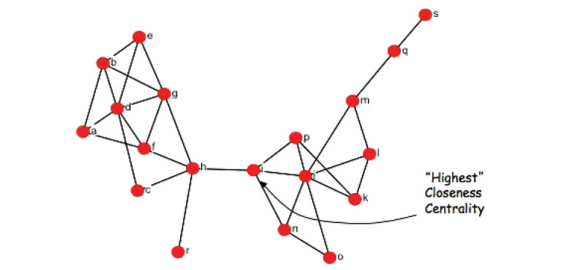


Hình 2.2 Minh họa Betweenness Centrality

Nguồn: (Borgatti, ARTS lecture 2009)

2.1.2.4 Tính trung tâm về độ gần gũi (closeness centrality):

Tính trung tâm gần cũng được định nghĩa là một tỷ lệ. Đầu tiên ta tính tổng khoảng cách của một đỉnh đến tất cả các đỉnh khác trong mạng. Sau đó ta xác định tỷ lệ trung tâm gần bằng tỷ lệ giữa số lượng các đỉnh khác chia cho tổng ở bước đầu tiên. Tính trung tâm gần cung cấp một thước đo toàn cầu về vị trí của một đỉnh trong mạng. Trong khi đó tính trung tâm giữa được xác định với tham chiếu đến vị trí cục bộ của một đỉnh.

****

Hình 2.3 Minh họa Closeness Centrality

Nguồn: (Borgatti, ARTS lecture 2009)

2.1.2.5 Tính trung tâm phân cụm (clustering centrality):

Phân cụm là một quá trình phân vùng một tập hợp dữ liệu thành các tập hợp con có ý nghĩa sao cho tất cả dữ liệu trong cùng một nhóm tương tự và dữ liệu trong các nhóm khác nhau không giống nhau ở một số giác quan. Đó là một phương pháp khám phá dữ liệu và cách nhìn cho các mẫu hoặc cấu trúc trong dữ liệu được quan tâm. Phân cụm có ứng dụng rộng rãi trong khoa học xã hội, sinh học, hóa học, và khoa học thông tin.

Tính trung tâm phân cụm tức là tính tỷ lệ của các mối quan hệ, ràng buộc các một đỉnh trong mạng xã hội so với nhau. Về mặt này, phân cụm toàn cầu của một mạng là trung bình của các hệ số phân cụm cục bộ cho tất cả các nút trong mạng. Ta có thể tính trung tâm phân cụm theo hệ phương trình sau:

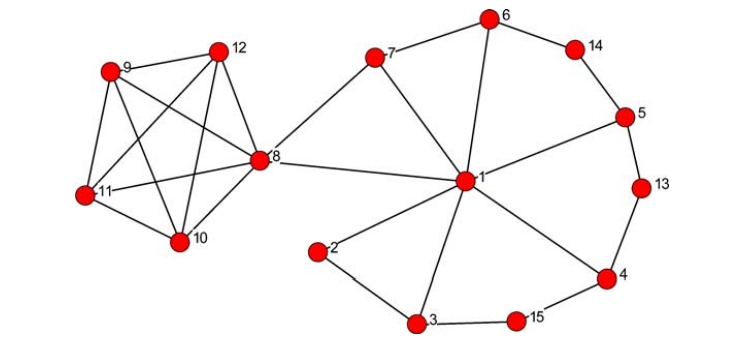
w( u, v) =

Trong đó đồ thị G = (V, E, w) với V là tập hợp đỉnh, E là tập hợp cạnh của đồ thị G và w là hàm trọng số cạnh. Hàm w(u, V)) và w(e) biểu thị trọng lượng của cạnh 𝑒 giữa hai đỉnh và V. Nếu không có cạnh giữa hai đỉnh và V, thì w(u,V)= 0. Ngược lại w(u, V) = 1.

2.2 Key players và cách tìm key players

2.2.1 Key players và các vấn đề xoay quanh key players

Key players trong mạng xã hội là những người dùng có sức ảnh hưởng nhất định đến mạng xã hội đó. Ví dụ như những người nổi tiếng, ca sĩ, diễn viên hay chính trị gia có tầm ảnh hưởng lớn trong xã hội. Hoặc cũng có thể là một admin của một fan page hay một group nào đó. Những người dùng này có nhiều liên kết, ràng buộc với các người dùng hay các đối tượng khác và có sức ảnh hưởng đến mạng xã hội nên được gọi chung là key players. Trong mô hình hóa đồ thị của mạng xã hội, các key players được hiểu là một đỉnh của đồ thị mà đồ thị đó bao gồm nhiều đỉnh khác có nhiều liên kết đến key players. Hay nói cách khác, các đỉnh có bậc cao nhất trong đồ thị là key players. Ta xét đồ thị sau để minh họa cho định nghĩa một key players trong đồ thị hóa mô hình mạng xã hội:



Hình 2.4 Đồ thị minh họa cho định nghĩa key palyers

Nguồn: (Comput Math Organiz Theor (2006) pp. 23)

Trong đồ thị trên, các đỉnh 1 có bậc 7 và đỉnh 8 có bậc 6 có thể hiểu là một key players bởi các đỉnh này có nhiều liên kết đối với các đỉnh còn lại trong đồ thị. Điều này mô phỏng các mối liên hệ, sự liên kết, ràng buộc hay ảnh hưởng của các key players trong mạng xã hội đối với các đối tượng trong và ngoài phạm vi mảng của mạng xã hội ta đang xét. Đồ thị trên gồm 2 đồ thị con liên kết với nhau, trong đó đỉnh 8 là key players của đồ thị con bên trái, đỉnh 1 là kye players của đồ thị con bên phải. Ta thấy các đồ thị này đều có cạch nối với nhau, điều này mô phỏng cho việc các key players trong một mạng xã hội có thể tưởng tác và ảnh hưởng với nhau và với toàn mạng xã hội.

Trong đồ thị trên, nếu ta xóa đi đỉnh 1 và đỉnh 8, 2 đồ thị con sẽ tách biệt nhau. Trong mô hình mạng xã hội cũng vậy, nếu một key players bị xóa đi, mạng sẽ hội sẽ bị phân cụm, mất đi sự liên kết giữa các người dùng và giữa các đối tượng với nhau. Trong quá trình xây dựng và phát triển mạng xã hội, sự liên kết chặt chẽ giữa các người dùng luôn là mục tiêu hướng đến. Nếu mạng xã hội bị phân cụm sẽ dẫn đến sự rời rạc, mất liên kết giữa những người dùng với nhau. Do đó vấn đề xây dựng và duy trì các key players này trong mạng xã hội là rất quan trọng.

Tuy nhiên khi đã hình thành một mạng xã hội lớn và phức tạp, việc mạng xã hội đó thay đổi và phát triển dẫn đến các key players mới hình thành, và nhiều đến mức cần các key players lớn hơn hay cả một đồ thị các key players để quản lý. Trong đó còn có các key players cũ dần mất đi sự liên kết giữa các người dùng với nhau. Từ vấn đề này, ta cần đưa ra giải pháp cho việc tìm và nhận dạng một key player để duy trì và phát triển các key players trong mạng xã hội, gọi là detect key players.

2.2.2 Làm thế nào để phát hiện và duy trì key players

Xét một đồ thị mô phỏng mô hình mạng xã hội, ta có thể tìm một key players theo các tiêu chí sau:

* Tìm đỉnh có nhiều cạnh nối với các đỉnh khác nhất, sử dụng phương pháp tính trung tâm về mức độ (degree centrality).
* Tìm đỉnh có số lần đi qua trắc địa nhiều nhất, tức đỉnh đó nằm trên đường đi ngắn nhất của đồ thị ta đang xét nhiều nhất có thể. Sử dụng phương pháp tính trung tâm về mặt giữa (betweenness centrality).
* Tìm đỉnh có sự gần gũi nhất với tất cả các đỉnh khác, sử dụng phương pháp tính trung tâm về sự gần gũi (closeness centrality).

Từ các tiêu chí trên ta có thể xây dựng thuật toán xác định key players trong một mạng xã hội được mô hình hóa bằng đồ thị vô hướng G(V, E) với n đỉnh bao gồm các đồ thị con sau :

* (): ): Đồ thị biểu diễn cho những người dùng như các nút và quan hệ giữa chúng. Số lượng đỉnh được ký hiệu là .
* (): Đồ thị biểu diễn các tổ chức, nhóm người dùng dưới dạng các nút và các mối quan hệ giữa chúng. Số lượng đỉnh ký hiệu là .
* (): Đồ thị biểu diễn các vị trí, địa điểm dưới dạng các nút và các mối quan hệ giữa chúng. Số lượng đỉnh ký hiệu là .

Tổng của tất cả các nút trong các đồ thị con trên bằng tổng số nút trong toàn bộ biểu đồ, tức là, n = + + . Ngoài các cạnh đã cho trong () , (), có ba tập con bổ sung của các cạnh là một phần của tập hợp E trong (V, E). Những bộ cạnh này là:

* : Các cạnh thể hiện quan hệ giữa người dùng và nhóm người dùng, tổ chức.
* : Các cạnh thể hiện quan hệ giữa người dùng và vị trí, địa điểm
* : Các cạnh thể hiện quan hệ giữa các nhóm người dùng, tổ chức và vị trí, địa điểm.

Sau khi có các định nghĩa cần thiết, ta bắt đầu thuật toán tìm key players với 5 bước sau :

* Bước 1: Tính mật độ trung tâm (degree centrality) cho mỗi nút thuộc tập hợp trong đồ thị ) theo công thức :

() =

* Bước 2: Tính mật độ trung tâm (degree centrality) cho mỗi nút thuộc tập hợp trong đồ thị ) theo công thức :

() =

* Bước 3: Tính độ gần trung tâm (closeness centrality) cho mỗi nút thuộc tập hợp V trong đồ thị (V, E) theo công thức:

() =

* Bước 4: Tính độ gần trung tâm (closeness centrality) cho mỗi nút thuộc tập hợp theo công thức :

() =

* Bước 5: Tính độ giữa trung tâm (betweenness centrality) cho mỗi nút thuộc tập hợp và cả trong tập hợp trong đồ thị (theo công thức:

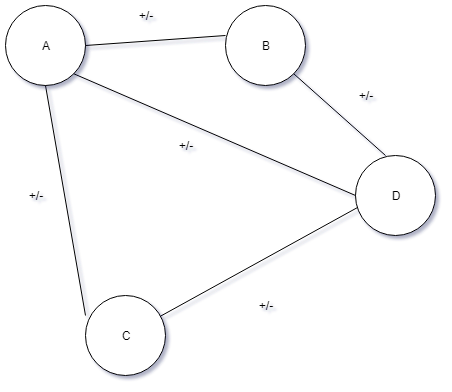
()=

Sau khi tất cả các mật độ trung tâm đã được tính toán theo các bước của thuật toán trên, các key players cần tìm trong mạng xã hội được xác định chính là các nút có tổng số cao nhất so với hầu hết các nút còn lại. Khi đó nút ta vừa tìm được bao gồm các thuộc tính gần nhất với tất cả các nút trong đồ thị, nằm trên đường đi ngắn nhất qua các đỉnh của đồ thị và có nhiều cạnh nối với các đỉnh còn lại nhất. Từ đó thỏa mãn các đặc tính của một key player.

2.3 Đồ thị signed và các vấn đề, ứng dụng của đồ thị signed trong mô hình hóa mạng xã hội

2.3.1 Đồ thị signed và các vấn đề xoay quanh đồ thị signed

Một đồ thị signed là một đồ thị có các cung mang giá trị là các dấu dương hoặc âm. Đồ thị signed Z bao gồm một đồ thị không dấu, ký hiệu là |Z| và một phần ánh xạ = E(|Z|) {}, được xác định trên tất cả các cung trừ một nửa vòng cung và được yêu cầu phải tích cực trên các vòng cung đó. Ta ký hiệu N (E) cho tập hợp đỉnh, E (Z) cho tập hợp cung của Z. Ta có thể cũng viết Z = (r, a) nếu muốn khi dùng với đồ thị cơ bản là r, hoặc Z = (N, E, a) để chỉ ra rằng tập hợp đỉnh là N và tập hợp cạnh là E.



Hình 2.5. Đồ thị Signed

Từ khái niệm trên, ta có ba vấn đề cơ bản về một đồ thị signed là: Đồ thị signed có được cân bằng không? Kích thước lớn nhất của một cạnh cân bằng được đặt trong đồ thị signed là bao nhiêu? Và số đỉnh nhỏ nhất phải được xóa để làm cho nó cân bằng là bao nhiêu? Câu hỏi đầu tiên là dễ dàng để giải quyết nhanh chóng. Đồ thị signed được cân bằng khi có số cung dương bằng số cung âm. Hai câu hỏi sau có thể giải quyết bằng các thuật toán.

2.3.1 Ứng dụng của đồ thị signed trong mô hình hóa mạng xã hội

Các biểu đồ thị signed được sử dụng để mô hình hóa các tình huống trong một mạng xã hội. Các cung mang dấu dương đại diện cho tình bạn hay các mặt tích cực trong một mối quan hệ giữa các đối tượng người dùng. Và các cung mang dấu âm cạnh tranh tiêu cực giữa các đối tượng người dùng. Em lấy ví dụ 3 chu kỳ tích cực là ba người bạn chung hoặc hai người bạn với một kẻ thù chung; trong khi 3 chu kỳ âm là ba kẻ thù chung hoặc hai kẻ thù có chung bạn bè.

Theo lý thuyết của một đồ thị signed cân bằng em đã đề cập ở trên, các chu kỳ tích cực khi được cân bằng được gọi là tình huống xã hội ổn định, trong khi các chu kỳ tiêu cực, tức không cân bằng được gọi là tình huống xã hội không ổn định. Trong ví dụ trên, trường hợp ba kẻ thù chung xảy ra do việc chia sẻ một kẻ thù chung có khả năng khiến hai trong số những kẻ thù trở thành bạn bè. Trong trường hợp hai kẻ thù chia sẻ một người bạn, người bạn được chia sẻ có khả năng chọn một người khác và biến một trong những tình bạn của mình thành kẻ thù.

Để cân bằng một đồ thị signed thì chỉ cần thay đổi cho số các cung dương bằng với số các cung âm là được. Nhưng trên thực tế khi ta áp dụng vào mô hình hóa mạng xã hội, khi đó đồ thị signed to lớn và phức tạp, bất kì sự thay đổi nào cũng có thể ảnh hưởng ít nhiều đến cả một hệ thống thậm chí cả một mạng xã hội. Do đó vấn đề cân bằng đồ thị signed trên các hệ thống lớn là một vấn đề khó. Các mối quan hệ thân thiện gắn kết các người dùng trong một mạng xã hội lại với nhau, trong khi một mạng xã hội bị chia thành nhiều phe sẽ rất bất ổn. Do đó ứng dụng quan trọng của một đồ thì signed là mô phỏng các mối quan hệ tích cực, tiêu cực giữa các người dùng chung trong mạng xã hội và giải quyết bài toán cân bằng đồ thị signed đó chính là giải pháp để duy trì sự ổn định và cân bằng mạng xã hội.

CHƯƠNG 3 – PHÁT HIỆN CỘNG ĐỒNG TRONG MẠNG XÃ HỘI

3.1 Phát hiện cộng đồng là gì và ứng dụng của nó trong mạng xã hội

Trong một mạng xã hội mọi người thường có xu hướng thành lập các nhóm trong môi trường làm việc, sở thích, gia đình, bạn bè của họ. Từ đó hình thành nên một cộng đồng. Các cộng đồng trong mạng xã hội này cho phép chúng ta tạo ra một bản đồ tỷ lệ lớn của mạng xã hội thông qua các cộng đồng riêng lẻ hoạt động như các nút trong mạng xã hội, giúp cho việc nghiên cứu và phát triển mạng xã hội thuận tiện hơn.

Các cộng đồng riêng lẻ này cũng làm rõ chức năng của hệ thống được đại diện bởi mạng xã hội vì các cộng đồng thường tương ứng với các đơn vị chức năng của hệ thống. Ví dụ trong một cộng đồng người dùng thích xem phim ảnh, các nhóm người thích xem phim thể loại hành động có chức năng như một đồ thị con và khi người dùng mạng trong nhóm này tương tác với nhau tương ứng với các chu trình hoặc đường đi trong đồ thị. Cộng đồng này tương ứng với các nút trong đồ thị lớn hơn mô phỏng cộng đồng những người thích xem phim và đều có chức năng tương tự nhau. Tương tự, các nhóm người dùng thích bóng đá, bơi lội hình thành các cộng đồng theo chủ đề thể thao. Việc xác định các cấu trúc phụ này trong một mạng xã hội có thể cung cấp cái nhìn sâu sắc về cách chức năng mạng và cấu trúc liên kết ảnh hưởng lẫn nhau. Cái nhìn cụ thể, sâu sắc như vậy có thể hữu ích trong việc cải thiện một số thuật toán trên đồ thị như thuật toán phân cụm.

Một lý do khác khiến các cộng đồng có tầm ảnh hưởng quan trọng là chúng thường có các thuộc tính rất khác so với các thuộc tính trung bình của các mạng. Do đó, chỉ tập trung vào các thuộc tính trung bình thường bỏ lỡ nhiều tính năng quan trọng và thú vị bên trong các mạng. Ví dụ, trong một mạng xã hội nhất định, cả hai nhóm tập thể và kín đáo có thể tồn tại đồng thời với nhau.

Sự tồn tại của các cộng đồng nói chung cũng ảnh hưởng đến các quá trình khác nhau như lan truyền tin đồn chưa được kiểm chứng hoặc các thông tin sai sự thật, thông tin phản động, chống phá Nhà Nước lan truyền trên mạng. Do đó để hiểu đúng bản chất của các cộng đồng, điều quan trọng đầu tiên là phát hiện các cộng đồng, từ đó nghiên cứu cách chúng ảnh hưởng đến các người dùng trong cộng đồng và các cộng đồng khác trong mạng xã hội.

Một ứng dụng quan trọng nữa mà phát hiện cộng đồng mang lại trong khoa học mạng là dự đoán các liên kết bị thiếu và xác định các liên kết sai trong mạng. Trong quá trình đo lường, một số liên kết có thể không được quan sát vì một số lý do nào đó. Tương tự, một số liên kết có thể nhập sai dữ liệu do các lỗi, sai số trong các thuật toán đo lường, tính toán. Cả hai trường hợp này đều được xử lý tốt bởi thuật toán phát hiện cộng đồng vì nó cho phép người ta gán xác suất tồn tại của một cạnh giữa một cặp nút cho trước.

Từ vai trò to lớn của phát hiện cộng đồng, ta sẽ đi đến các thuật toán phát hiện cộng đồng dưới đây.

3.2 Thuật toán module hóa và thuật toán cắt

3.2.1 Thuật toán tối đa hóa module

Mặc dù có những hạn chế nhất định, một trong những phương pháp được sử dụng rộng rãi nhất để phát hiện cộng đồng là tối đa hóa module. Tính module là một chức năng có lợi cho việc đo lường chất lượng của một bộ phận cụ thể trong mạng xã hội, từ đó hình thành các cộng đồng.

Phương pháp tối đa hóa module để phát hiện các cộng đồng tìm một hoặc nhiều module tối đa đặc biệt bằng cách tìm kiếm trên các bộ phận có thể có của một mạng xã hội để. Do tìm kiếm toàn diện trên tất cả các bộ phận có thể thường là khó hiểu và phức tạp nên các thuật toán thực tế dựa trên các phương pháp tối ưu hóa gần đúng như thuật toán tham lam, thuật toán mô phỏng hoặc tối ưu hóa quang phổ nhưng với các cách tiếp cận khác nhau mang lại sự cân bằng khác nhau giữa tốc độ hội tụ và độ chính xác, độ sai số của thuật toán.

Một cách tiếp cận với thuật toán tối đa hóa module phổ biến là phương pháp Louvain, phương pháp lặp lại tối ưu hóa các cộng đồng địa phương cho đến khi module toàn cầu không còn có thể được cải thiện do nhiễu loạn đối với trạng thái cộng đồng hiện tại. Một phương pháp khác nữa là sử dụng sơ đồ RenEEL, một ví dụ về mô hình học tập cực đoan (EEL). Hiện tại đây là thuật toán tối đa hóa mô đun hóa tốt nhất.

Theo nhận định của em, lợi ích của thuật toán tối ưu hóa module này vẫn còn phải cân nhắc bởi tối ưu hóa module thường không phát hiện được các cụm nhỏ hơn theo một số tỷ lệ nhất định và còn phải tùy thuộc vào kích thước của mạng. Hơn nữa, kết quả của các giá trị module này được đặc trưng bởi sự suy biến rất lớn của các cộng đồng trong mạng xã hội và có tính module cao, gần với mức tối đa tuyệt đối nên chúng có thể rất khác nhau, từ đó khó phân biệt các cộng đồng có đặc tính tương đương nhau.

3.2.2 Thuật toán cắt

Một trong những thuật toán lâu đời nhất để phân chia mạng thành các cộng đồng là thuật toán cắt tối thiểu. Các biến thể như thuật toán cắt tỷ lệ và thuật toán cắt chuẩn hóa cũng tương tự nên em chỉ xét thuật toán cắt tối thiểu.

Trong phương pháp cắt tối thiểu, mạng được chia thành một số thành phần được xác định trước, thường có cùng kích thước, được chọn sao cho số lượng cạnh giữa các nhóm đỉnh được giảm thiểu. Phương thức này hoạt động tốt trong nhiều ứng dụng nhưng không lý tưởng cho việc phát hiện cộng đồng trong các mạng xã hội phức tạp vì nó sẽ tìm thấy các cộng đồng bất kể chúng có ẩn trong cấu trúc khác hay không và nó sẽ chỉ tìm thấy một số cộng đồng nhất định.

3.3 Thuật toán Grivan-Newman

Thuật toán Girvan-Newman phát hiện các cộng đồng bằng cách loại bỏ dần các cạnh khỏi mạng ban đầu. Các thành phần được kết nối của mạng còn lại là các cộng đồng. Thay vì cố gắng xây dựng một biện pháp cho chúng ta biết các cạnh nào là trung tâm nhất của cộng đồng, thuật toán Girvan-Newman tập trung vào các cạnh có khả năng là "giữa" của các cộng đồng.

Vertex betweenness là một chỉ số của các nút trung tâm cao trong các mạng. Đối với bất kỳ nút nào, độ giữa của đỉnh được định nghĩa là số đường đi ngắn nhất giữa các cặp nút chạy qua nó. Nó có liên quan đến các mô hình trong đó mạng điều chỉnh chuyển hóa giữa điểm bắt đầu và điểm kết thúc cho trước, giả sử việc chuyển giao này tìm kiếm đường đi ngắn nhất có sẵn.

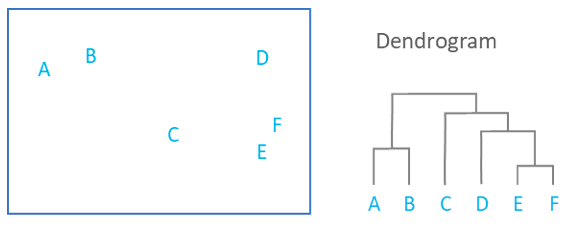
Thuật toán Girvan-Newman mở rộng định nghĩa này cho trường hợp các cạnh, xác định "độ giữa" của cạnh là số đường đi ngắn nhất giữa các cặp nút chạy dọc theo nó. Nếu có nhiều hơn một đường dẫn ngắn nhất giữa một cặp nút, thì mỗi đường dẫn được gán trọng số bằng nhau sao cho tổng trọng số của tất cả các đường dẫn bằng với sự thống nhất. Nếu một mạng chứa các cộng đồng hoặc nhóm chỉ được kết nối lỏng lẻo bởi một vài cạnh giữa các nhóm, thì tất cả các đường dẫn ngắn nhất giữa các cộng đồng khác nhau phải đi dọc theo một trong các cạnh này. Do đó, các cạnh kết nối cộng đồng sẽ có độ giữa cạnh cao. Bằng cách loại bỏ các cạnh này, các nhóm được tách biệt với nhau và do đó cấu trúc cộng đồng cơ bản của mạng được phát hiện.

Các bước của thuật toán để phát hiện cộng đồng được tóm tắt dưới đây:

* Bước 1: Độ giữa của tất cả các cạnh hiện có trong mạng được tính toán trước. Các cạnh có độ giữa cao nhất được loại bỏ. Khoảng giữa của tất cả các cạnh bị ảnh hưởng bởi việc loại bỏ được tính toán lại.
* Bước 2 và bước 3 được lặp lại cho đến khi không còn cạnh nào và kết quả là cấu trúc một cộng đồng.

Trên thực tế, các độ giữa duy nhất được tính toán lại chỉ là những thứ bị ảnh hưởng bởi việc loại bỏ, có thể làm giảm thời gian chạy mô phỏng của quá trình chạy thuật toán trên máy tính. Tuy nhiên, tính trung tâm giữa phải được tính toán lại với từng bước hoặc khi xảy ra lỗi nghiêm trọng nào đó. Lý do là mạng thích nghi với các điều kiện mới được đặt sau khi loại bỏ cạnh. Chẳng hạn, nếu hai cộng đồng được kết nối bởi nhiều hơn một cạnh, thì không có gì đảm bảo rằng tất cả các cạnh này sẽ có độ giữa cao. Bằng cách tính toán lại giữa các khoảng cách sau khi loại bỏ từng cạnh, ta có thể đảm bảo rằng ít nhất một trong các cạnh còn lại giữa hai cộng đồng sẽ luôn có giá trị cao.

Kết quả cuối cùng của thuật toán Girvan-Newman là một sơ đồ đại diện cho một cây (gọi là dendrogram).



Hình 3.1 Minh họa một dendrogram

Nguồn: [displayr.com]

Khi thuật toán Girvan-Newman chạy, dendrogram được tạo từ trên xuống (tức là mạng chia thành các cộng đồng khác nhau với việc loại bỏ liên kết liên tiếp). Các nút lá của dendrogram là các nút riêng lẻ. Độ phức tạp của thuật toán Girvan-Newman là .

3.4 Thuật toán Node similarity based

Thuật toán phát hiện cộng đồng dựa trên sự tương tự giữa các nút (Node similarity based algorithm) so sánh một tập hợp các nút dựa trên các nút mà chúng được kết nối. Hai nút được coi là tương tự nhau nếu chúng chia sẻ nhiều hàng xóm giống nhau. Thuật toán Node similarity based tính toán các điểm tương đồng theo cặp dựa trên số liệu Jaccard.

Đầu vào của thuật toán này là một biểu đồ lưỡng cực được kết nối với nhau và có chứa hai tập hợp nút khác nhau. Mỗi mối quan hệ bắt đầu từ một nút trong tập nút đầu tiên và kết thúc tại một nút trong tập nút thứ hai. Thuật toán Node similarity based so sánh tất cả các nút từ nút đầu tiên được đặt với nhau dựa trên mối quan hệ của chúng với các nút trong tập thứ hai. Độ phức tạp của phép so sánh này tăng theo phương trình bậc hai với số lượng nút cần so sánh. Ta có thể giảm độ phức tạp của thuật toán bằng cách bỏ qua các nút bị ngắt kết nối.

Ngoài độ phức tạp tính toán, yêu cầu bộ nhớ để tạo kết quả cũng có tỷ lệ xấp xỉ bậc hai. Để sử dụng bộ nhớ bị ràng buộc, thuật toán yêu cầu giới hạn rõ ràng về số lượng kết quả để tính toán trên mỗi nút. Đây là thông số “topK”. Nó có thể được đặt thành bất kỳ giá trị nào, ngoại trừ giá trị 0.

Đầu ra của thuật toán là các mối quan hệ mới giữa các cặp của tập nút đầu tiên. Điểm tương đồng được thể hiện thông qua các thuộc tính mối quan hệ.

3.5 Thuật toán Label Propagation Community Detection (LPA)

Thuật toán Label Propagation Community Detection (LPA) là một thuật toán nhanh để phát hiện các cộng đồng trong mạng xã hội. Nó phát hiện các cộng đồng này bằng cách sử dụng cấu trúc mạng độc lập làm hướng dẫn và không cần phải có chức năng mục tiêu được xác định trước hoặc thông tin trước về cộng đồng.

Một tính năng của LPA là các nút có thể được gán nhãn sơ bộ để thu hẹp phạm vi giải pháp được tạo. Điều này có nghĩa là nó có thể được sử dụng như một cách tìm kiếm cộng đồng bán giám sát nơi chúng tôi chọn một số cộng đồng ban đầu.

Ở điều kiện ban đầu, các nút mang nhãn biểu thị cộng đồng quản lý các nút này. Tư cách thành viên trong một cộng đồng sẽ được thay đổi dựa trên các nhãn mà các nút lân cận sở hữu. Sự thay đổi này tùy thuộc vào số lượng nhãn tối đa trong một mức độ của các nút. Mỗi nút được khởi tạo với một nhãn duy nhất, sau đó các nhãn khuếch tán qua mạng. Do đó, các nhóm kết nối dày đặc nhanh chóng đạt được một nhãn chung. Khi nhiều nhóm dày đặc như vậy được tạo ra trên toàn mạng, ta tiếp tục mở rộng ra bên ngoài cho đến khi không thể làm như vậy được nữa thì thuật toán kết thúc và các cộng đồng trong mạng xã hội được tìm thấy. Em mô phỏng thuật toán này qua 5 bước như sau:

* Bước 1. Khởi tạo nhãn tại tất cả các nút trong mạng. Đối với một nút đã cho x, ta có (0) = x.
* Bước 2. Đặt t = 1.
* Bước 3. Sắp xếp các nút trong mạng theo thứ tự ngẫu nhiên và đặt nó thành X.
* Bước 4. Với mỗi x ∈ X được chọn theo thứ tự cụ thể đó, hãy để (t) = f ( (t), ..., (t), (t - 1), ..., (t - 1) ). Hàm f ở đây trả về nhãn xảy ra với tần suất cao nhất trong số các hàng xóm. Chọn nhãn ngẫu nhiên nếu có nhiều nhãn tần số cao nhất.
* Bước 5. Kiểm tra nếu mỗi nút có một nhãn mà số hàng xóm tối đa của chúng có thì dừng thuật toán. Nếu không, đặt t = t + 1 và quay lại bước 3.

CHƯƠNG 4 – MINH HỌA MỘT SỐ THUẬT TOÁN

Trong phần minh họa các thuật toán này em sẽ lần lượt minh họa 5 thuật toán đã trình bày trong chương 2, cụ thể là các thuật toán: degree centrality, closeness centrality, betweenness centrality, clustering centrality và thuật toán tìm key players. Trong phần viết code để demo thuật toán, em sử dụng ngôn ngữ lập trình Python3 bởi ngôn ngữ này có sẵn các thư viện rất phù hợp cho việc demo các thuật toán ở trên.

Thay vì đọc input từ một file để có resource cho source code thì em code luôn trong file main.py để đọc input. Do đó folder Demo sẽ không có file resource mà chỉ bao gồm 2 file là file source code tên main.py và file release tên output.txt chỉ tạo ra khi thực thi file source code main.py.

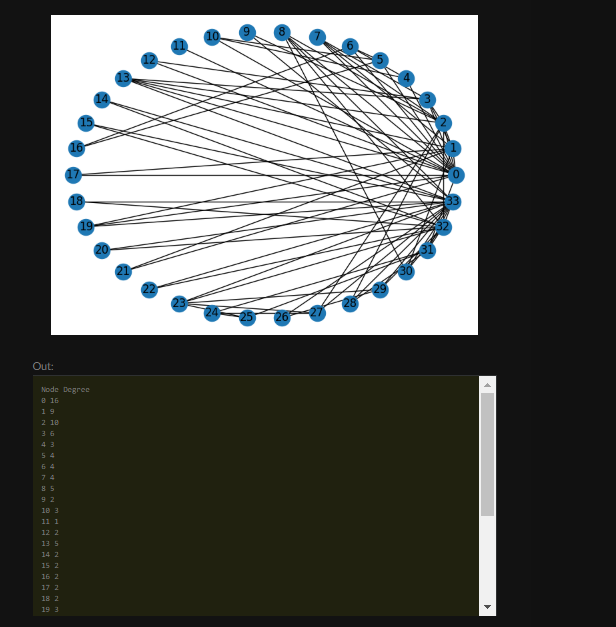
4.1 Import thư viện và lấy input

Thư viện em sử dụng là thư viện networkx trong Python – một thư viện của Python cho phép thao tác, nghiên cứu cấu trúc, động lực và chức năng của các thành phần trong mạng xã hội phức tạp.



Hình 4.1 Đoạn code import thư viện networkx và input

Trong đoạn code trên, em gán input G là một đồ thị tên karate club có sẵn trong thư viện networkx. Đồ thị G khi mô phỏng qua hình vẽ bên dưới:

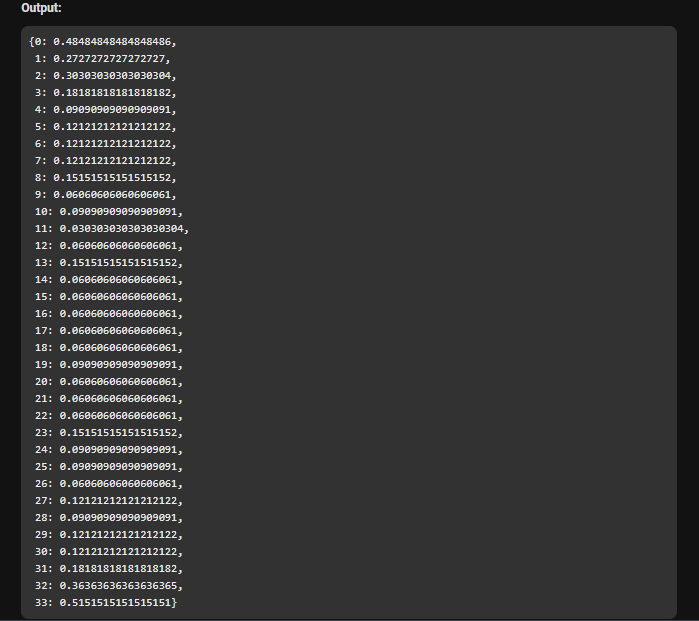


Hình 4.1 Mô phỏng đồ thị input G

4.2 Demo thuật toán

4.2.1 Minh họa thuật toán Degree Centrality

Em truyền input G làm argument cho hàm degree\_centrality() có sẵn trong module nx và gán kết quả trả về của hàm này cho biến deg. Sau khi in biến deg ra màn hình console, kết quả là degree của mỗi nút trong đồ thị G.

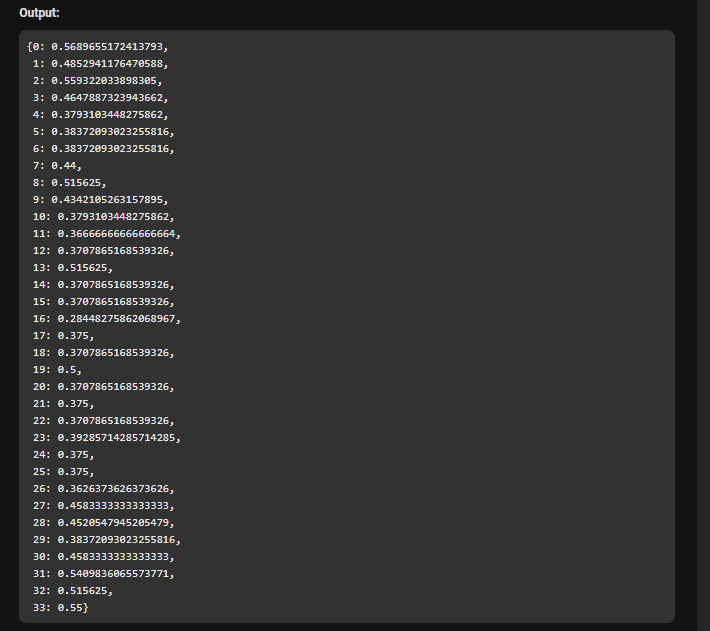


Hình 4.2. Output của hàm degree\_centrality

4.2.2 Minh họa thuật toán Closeness Centrality

Đồ thị input G gồm các nút quan trọng có cạnh nối với các nút khác trong đồ thị. Độ gần của mỗi nút được tính bằng tổng độ dài đường dẫn từ nút đã cho đến tất cả các nút khác. Nhưng đối với một nút không thể tiếp cận tất cả các nút khác, tính độ gần của nút đó vẫn sẽ được tính theo thuật toán Closeness Centrality đã được đề cập trong chương 2.

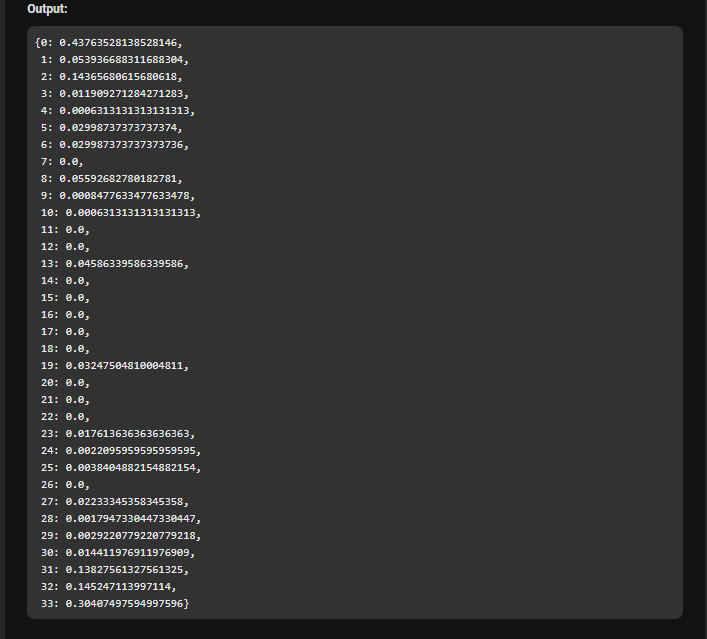
Tương tự như thuật toán Degree Centrality, em cũng truyền input G làm argument cho hàm closeness\_centrality() có sẵn trong module nx và gán kết quả trả về của hàm này cho biến clo. Sau khi in biến deg ra màn hình console, kết quả là độ gần của mỗi nút trong đồ thị G.



Hình 4.3. Output của hàm closeness\_centrality

4.2.3 Minh họa thuật toán Betweeness Centrality

Đồ thị G gồm các nút quan trọng kết nối các nút khác. Đối với các đồ thị có số lượng nút lớn, giá trị trung tâm giữa là rất cao. Vì vậy, chúng ta có thể bình thường hóa giá trị bằng cách chia cho số lượng cặp nút (không bao gồm nút hiện tại). Đối với Đồ thị có Hướng, số lượng cặp nút là (| N | -1) \* (| N | -2), trong khi đối với Đồ thị vô hướng, số lượng cặp nút là (1/2) \* (| N | -1) \* (| N | -2). Tất cả công việc tính toán theo thuật toán betweeness centrality đều được tính theo hàm betweenness\_centrality(), kết quả được gán vào biến bet và in ra màn hình kết quả sau:

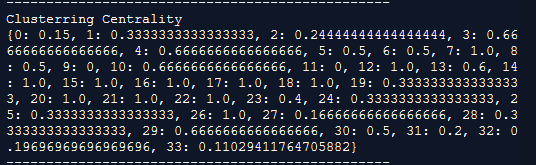


Hình 4.3. Output của hàm betweenness\_centrality

4.2.4 Minh họa thuật toán Clusterring Centrality

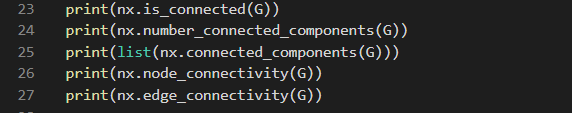
Em tính trung bình trên tất cả hệ số phân cụm cục bộ của các nút riêng lẻ, sau đó tính tổng hệ số phân cụm cục bộ của tất cả các nút chia cho tổng số nút.

Tương tự như các thuật toán ở tên, em sử dụng hàm clustering() của module nx để tính hệ số phân cụm cục bộ của các nút riêng lẻ độ phân trong đồ thị input G. Kết quả đạt được là:



Hình 4.4. Hệ số phân cụm cục bộ của các nút riêng lẻ trong đồ thị input G

Dựa vào các hệ số trên, ta có thể biết được đồ thị đã cho có tính phân cụm. Thư viện Networkx cung cấp một số chức năng tích hợp để kiểm tra các tính năng kết nối khác nhau của đồ thị. Cụ thể trong đoạn code sau:

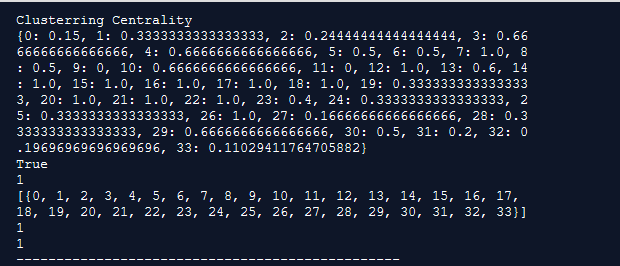


Hình 4.5. Đoạn code demo thuật toán Clusterring Centrality

Trong đoạn code trên, hàm is\_connected() kiểm tra xem đồ thị G có được kết nối hay không. Hàm trả về True nếu đồ thị được kết nối, nếu không sẽ trả về False. Trong trường hợp đồ thị input G, hàm này trả về giá trị False, tức đồ thị G không được kết nối và có tính phân cụm.

Hàm number\_connected\_components() trả về số lượng các thành phần kết nối khác nhau trong đồ thị G và hàm list(nx.connected\_components) trả về danh sách các nút trong các thành phần được kết nối khác nhau.

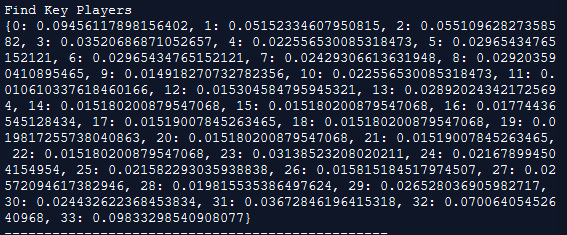
Từ đó ta có các hàm node\_connectivity() và edge\_ connectivity() lần lượt trả về số lượng nút được loại bỏ để đồ thị bị ngắt kết nối và trả về số cạnh cần loại bỏ để đồ thị bị ngắt kết nối. Các kết quả tính toán trên phục vụ cho thuật toán Clusterring Centrality, kết quả cuối cùng được in ra màn hình như sau:



Hình 4.6 Kết quả cuối cùng của thuật toán Clusterring Centrality

4.2.5 Minh họa thuật toán tìm key player

Em sử dụng hàm pagerank() được dựng sẵn trong module nx. Hàm này thường sẽ cần biến alpha là 0.8 hoặc 0.9. Ở đây em chọn alpha = 0.8. Hàm pagerank() sẽ gán một số điểm quan trọng cho mỗi nút. Các nút quan trọng là những nút có nhiều liên kết đến các nút khác. Kết quả trả về là danh sách các page rank của mỗi nút trong đồ thị:



Hình 4.7 Danh sách trả về của hàm pagerank

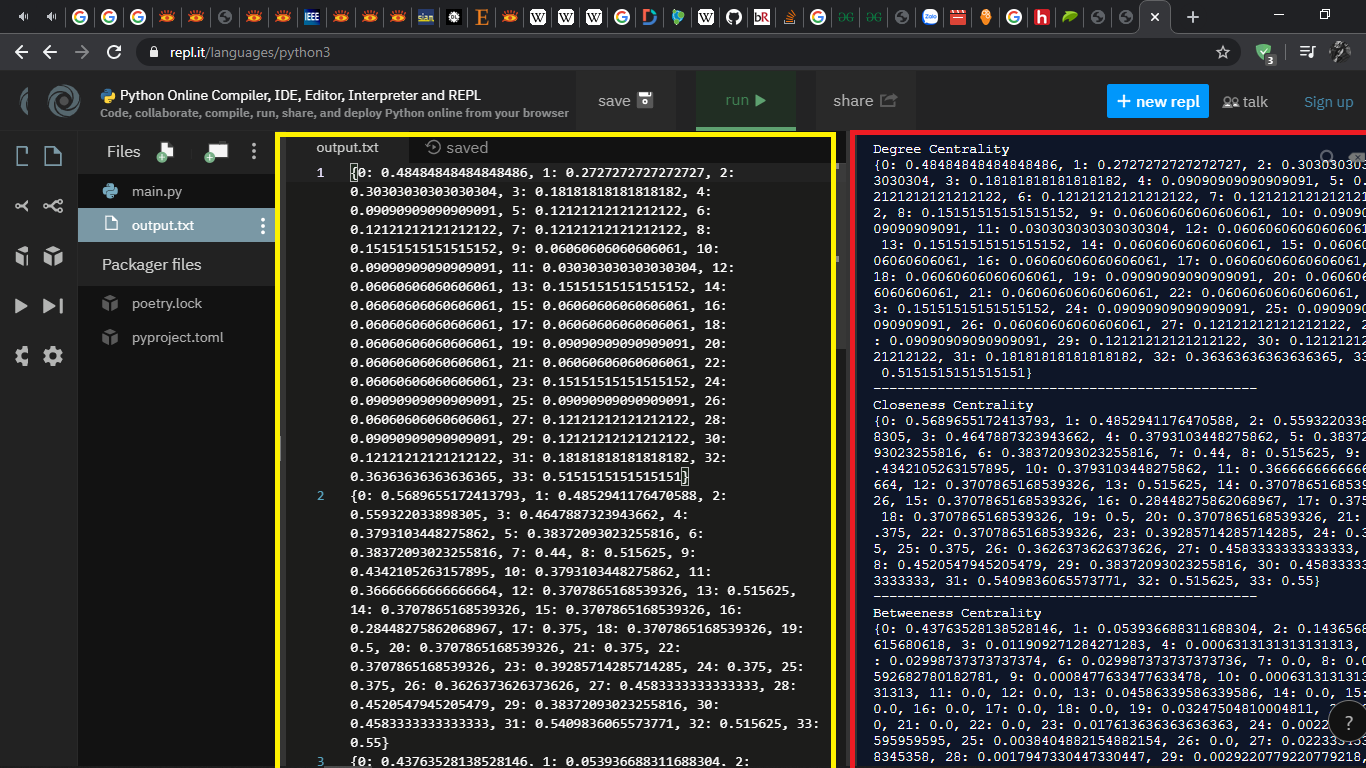
Từ danh sách trả về trên, ta tìm max của danh sách đó sẽ nhận được key player của một đồ thị.

4.3 Thực thi source code và kết quả của source code

Trong phần code em có sử dụng thư viện networkx nên trước khi chạy code này trên command line, người dùng phải cài đặt thư viện networkx theo cú pháp:

$ pip install networkx

Để tránh phải cài đặt lâu, em và cả người dùng file main.py có thể sử dụng website <https://repl.it/languages/python3/> để thực thi file main.py. Đầu tiên ta mở trình duyệt web và truy cập theo đường link ở trên. Sau đó copy toàn bộ code trong file main.py và paste vào file main.py trong website này và bấm Run để thực thi code. Kết quả sau khi thực thi file main.py được thể hiện trong file output.txt và màn hình console.



Hình 4.8 Kết quả sau khi thực thi file main.py

Trong hình 4.8, khung màu vàng thể hiện nội dung được ghi trong file output.txt và khung màu đỏ là màn hình console.

Đến đây là kết thúc phần demo của 5 thuật toán degree centrality, closeness centrality, betweenness centrality, clustering centrality và thuật toán tìm key players. Và cũng gần như đã kết thúc nội dung chính của bài báo cáo. Dưới đây là chương 5 để tóm tắt những gì em đã báo cáo và cũng là lời kết thúc của bài báo cáo này.

CHƯƠNG 5 – KẾT LUẬN

Trong bài báo cáo này đã đề cập đến các phương pháp đo lường trong mô hình hóa mạng xã hội, trong đó em nhận thấy thuật toán tìm key players trong một mạng xã hội là hấp dẫn nhất và có ứng dụng quan trọng trong lĩnh vực xây dựng và phát triển mạng xã hội. Cụ thể thuật toán này đã tổng hợp các phương pháp đo lường đã được đề cập trước đó như phương pháp tính độ giữa, phương pháp tính mức độ hay phương pháp tính trung tâm về độ gần gũi. Từ đó tìm ra được key players góp phần quan trọng trong một mô hình mạng xã hội.

Ngoài ra, bài báo cáo đã đề cập đến mảng phát hiện cộng đồng trong mạng xã hội. Việc đánh giá các thuật toán để tìm ra thuật toán nào tốt hơn trong việc phát hiện cấu trúc cộng đồng trong một mạng xã hội vẫn là một câu hỏi mở. Nó phải được dựa trên các phân tích về các mạng có cấu trúc đã biết. Nhưng qua quá trình nghiên cứu về các thuật toán phát hiện cấu trúc cộng đồng như em đã đề cập ở trong bài báo cáo này thì em nhận thấy thuật toán LPA (Label Propagation Community Detection) là một thuật toán thú vị và tính toán nhanh. Hơn nữa, đối với vốn kiến thức của bản thân em trong lý thuyết đồ thị, thì việc hiểu và vận dụng thuật toán này trở nên đơn giản nhất so với các thuật toán còn lại.

Quá trình demo các thuật toán em tập trung vào các thuật toán đo lường trong mô hình hóa mạng xã hội được đề cập trong chương 2 do các thuật toán này đơn giản hơn so với các thuật toán phát hiện cấu trúc cộng đồng trong chương 3. Em sử dụng ngôn ngữ lập trình Python3 với thư viện networkx với các hàm tích hợp sẵn đã hỗ trợ em rất nhiều trong việc minh họa các thuật toán trên.

Mô hình hóa mạng xã hội là một lĩnh vực cực kì hấp dẫn, có ứng dụng và tầm ảnh hưởng to lớn trong công nghệ xây dựng và phát triển mạng xã hội hiện nay. Trong quá trình nghiên cứu, em có một số ý tưởng trong việc phát triển mô hình này như việc tạo một đồ thị quản lý các key players và tìm các key players trong các key players nhỏ hơn trong mô hình mạng xã hội phức tạp, từ đó dễ quản lý sâu hơn các key players. Sau quá trình nghiên cứu về đề tài các ứng dụng của lý thuyết đồ thị trong mô hình mạng xã hội, em đã có cái nhìn tổng quan hơn về lý thuyết đồ thị, các ứng dụng của lý thuyết đồ thị trong thực tế cùng với các phương pháp tính toán tổ hợp trong các thuật toán ứng dụng với đồ thị.

**TÀI LIỆU THAM KHẢO**

**Tiếng Anh**

1. Wu, Q., Qi, X., Fuller, E., & Zhang, C.-Q. (2013), *“Follow the Leader”: A Centrality Guided Clustering and Its Application to Social Network Analysis*, The Scientific World Journal, 2013, pp. 1–9.
2. Iezzi, D. F. (2012), *Centrality Measures for Text Clustering. Communications in Statistics* - Theory and Methods, 41(16-17), pp. 3179–3197.
3. Leydesdorff, L. (2007), *Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. Journal of the American Society for Information Science and Technology,* Amsterdam School of Communications Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands, 58(9), pp. 1303–1319.
4. Tatti, N. (2019), *Density-Friendly Graph Decomposition*, ACM Transactions on Knowledge Discovery from Data, 13(5), pp. 1–29.
5. Berzinji, A., Kaati, L., & Rezine, A. (2012), *Detecting Key Players in Terrorist Networks*, 2012 European Intelligence and Security Informatics Conference*.*
6. Zaslavsky, T. (1982), *Signed graphs*, Discrete Applied Mathematics, 4(1), pp. 47–74.
7. Fortunato, S. (2010), *Community detection in graphs*, Physics Reports, 486(3-5), pp. 75–174.
8. Linton C. Freeman (2004), *The development of social network analysis,* Empirical Press Vancouver, BC Canada