

**ĐỒ ÁN CUỐI KỲ**

**Môn: Mạng Xã Hội**

**Đề tài báo cáo** : Local Higher-order network clustering

**Giảng viên** : Ths. Thái Bảo Trân

**Nhóm sinh viên thực hiện**:

**Họ và tên** **MSSV**

Lê Đức Vinh 17521262 (SĐT Liên hệ: 0986847270)

Ngô Anh Vũ 17521272

Lê Như Toàn 17521138

Mục lục

[**I.** **TỔNG QUAN ĐỀ TÀI** 5](#_Toc72085036)

[**1.** **Giới thiệu chung** 5](#_Toc72085037)

[**2.** **Phát biểu bài toán** 5](#_Toc72085038)

[**3.** **Thách thức của bài toán** 5](#_Toc72085039)

[**4.** **Đối tượng và phạm vi** 5](#_Toc72085040)

[**5.** **Mục tiêu** 6](#_Toc72085041)

[**II.** **MÔ HÌNH GIẢI BÀI TOÁN (FRAMEWORK)** 7](#_Toc72085042)

[**1.** **Tiền xử lý dữ liệu** 7](#_Toc72085043)

[**2.** **Đặc trưng** 7](#_Toc72085044)

[**3.** **Phương pháp máy học** 8](#_Toc72085045)

[**III.** **CÀI ĐẶT THỰC NGHIỆM** 9](#_Toc72085046)

[**1.** **Dữ liệu thực nghiệm** 9](#_Toc72085047)

[**2.** **Độ đo đánh giá** 10](#_Toc72085048)

[**3.** **Phương pháp thực nghiệm** 12](#_Toc72085049)

[**4.** **Kết quả thực nghiệm** 17](#_Toc72085050)

[**IV.** **KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 20](#_Toc72085051)

[**1.** **Kết luận** 20](#_Toc72085052)

[**2.** **Hướng phát triển** 20](#_Toc72085053)

[**V.** **DEMO** 21](#_Toc72085054)

[**VI. Thuật toán** 24](#_Toc72085055)

[**1.**  **Degree Centrality** 24](#_Toc72085056)

[a. Giải thuật: 24](#_Toc72085057)

[b. Code: 25](#_Toc72085058)

[c. Kết quả: 25](#_Toc72085059)

[**2.**  **Betweeness Centrality** 26](#_Toc72085060)

[a. Giải thuật: 26](#_Toc72085061)

[b. Code 26](#_Toc72085062)

[c. Kết Quả 26](#_Toc72085063)

[**3.**  **Closeness Centrality** 26](#_Toc72085064)

[a. Giải thuật: 26](#_Toc72085065)

[b. Code 26](#_Toc72085066)

[c. Kết Quả 26](#_Toc72085067)

[**4.**  **Clustering coefficient** 27](#_Toc72085068)

[a. Giải thuật: 27](#_Toc72085069)

[b. Code 27](#_Toc72085070)

[c. Kết Quả: 27](#_Toc72085071)

[**5.**  **Harmonic Centrality** 28](#_Toc72085072)

[a. Giải thuật: 28](#_Toc72085073)

[c. Kết Quả: 28](#_Toc72085074)

[**6.**  **Eigenvector Centrality** 28](#_Toc72085075)

[a. Giải thuật: 28](#_Toc72085076)

[b. Code 28](#_Toc72085077)

[c. Mô hình 29](#_Toc72085078)

[**7.**  **Page Rank** 29](#_Toc72085079)

[a. Giải thuật: 29](#_Toc72085080)

[b. Code 29](#_Toc72085081)

[c. Mô hình 30](#_Toc72085082)

[***Đánh giá quá trình thực hiện đồ án*** 30](#_Toc72085083)

[**TÀI LIỆU THAM KHẢO** 31](#_Toc72085084)

|  |  |  |  |
| --- | --- | --- | --- |
| Mốc thời gian | Công việc | Người thực hiện | Tỉ lệ % hoàn thành |
| 22/3-31/3/2021 | Tìm hiểu sơ bộ và xác định đề tài mà nhóm sẽ thực hiện | Lê Đức Vinh, Ngô Anh vũ, Lê Như Toàn | 100% |
| 1/4-14/4/2021 | Tìm hiểu mô hình giải bài toán (framework) | Lê Đức Vinh, Ngô Anh vũ, Lê Như Toàn | 100% |
| 15/4-31/4-2021 | Nghiên cứu và chọn ra các dataset để có thể tiến hành thực nghiệm | Lê Đức Vinh, Lê Như Toàn | 100% |
| 1/5-10/5/2021 | Chạy demo và thu thập kết quả   * Toàn: lọc dataset * Vinh: chạy demo trên Gephi * Vũ: chạy demo code sử dụng ngôn ngữ PHP | Lê Đức Vinh, Ngô Anh Vũ, Lê Như Toàn | 100% |
| 4/6-20/6-2021 | Làm slide báo cáo cuối kì | Ngô Anh Vũ, Lê Như Toàn | 100% |
| 15/6-20/6/2021 | Quay video demo | Lê Đức Vinh | 100% |
| 11/5-16/5 | Viết báo cáo | Lê Đức Vinh, Ngô Anh Vũ, Lê Như Toàn | 100% |

# **TỔNG QUAN ĐỀ TÀI**

## **1. Giới thiệu chung**

* Paper được viết bởi Hao Yin, Austin R. Benson, Jure Leskovec, David F. Gleich, được công bố năm 2017.
* Thuật toán chính được sử dụng là MAPPR, được phát triển trên thuật toán APPR và phương pháp phân cụm đồ thị toàn cục (global graph clustering).
* Thuật toán có nhiều điểm cải tiến hơn so với phương pháp phân cụm toàn cục:
  + Tập trung vào node cần quan tâm thay vì quét tất cả các node như phương pháp global.
  + Tốc độ tìm các seed node tốt nhanh
  + Thay đổi phù hợp với mục tiêu 1 cách nhanh chóng
  + Có thể áp dụng trên đồ thị có hướng

## **2. Phát biểu bài toán**

Các phương pháp phân cụm đồ thị cục bộ hiện nay cho phép phân cụm tập trung xung quanh node gốc ban đầu với tốc độ nhanh hơn và không phụ thuộc kích thước đầu vào. Tuy nhiên, các phương pháp phân cụm đồ thị cục bộ tại thời điểm đó không được thiết kế để tính đến các cấu trúc bậc cao quan trọng đối với mạng, cũng như chúng không thể xử lý hiệu quả các mạng có hướng. Chính vì vậy, phương pháp này được phát triển để khắc phục các vấn đề trên.

## **3. Thách thức của bài toán**

* Giả thuyết và các thuật toán trong phương pháp tiếp cận cục bộ hầu hết đều phát triển dùng độ dẫn như một thước đo chất lượng cụm, tuy nhiên độ dẫn này chỉ được định nghĩa cho các mạng vô hướng đơn giản.
* Sử dụng phương pháp phân cụm cục bộ dành cho các đồ thị bậc cao có hướng vẫn còn là một thách thức mở.
* Các phép đo chất lượng cụm hiện tại chỉ đơn giản là đếm các cạnh riêng lẻ và không xem xét các cạnh này có liên kết để tạo thành các cấu trúc mạng nhỏ được gọi là motif mạng. Các cấu trúc kết nối bậc cao như vậy rất quan trọng đối với việc tổ chức các mạng phức tạp vẫn còn là một câu hỏi mở làm thế nào để kết hợp các motif mạng vào các khuôn khổ phân cụm cục bộ.

## **4. Đối tượng và phạm vi**

* **Đối tượng**
* Đối tượng nghiên cứu của bài toán này là những đồ thị mạng bậc cao, có thể có hướng như:
  + Các cộng đồng trên web
  + Mạng xã hội
  + Mạng sinh học (mạng lưới tương tác protein - protein)
* **Phạm vi**
* Nghiên cứu được thực hiện vào năm 2017 với các dữ liệu tập trung vào các lĩnh vực mạng xã hội, web và mạng sinh học.
* Paper giải quyết vấn đề phát hiện cộng động trên các đồ thị bậc cao

## **5. Mục tiêu**

* Mục tiêu của phân cụm đồ thị - còn được gọi là phát hiện cộng đồng hoặc phân vùng đồ thị, là xác định các cụm node được liên kết dày đặc với node gốc cho trước ban đầu tạo thành một đồ thị mạng

# **II. MÔ HÌNH GIẢI BÀI TOÁN (FRAMEWORK)**



## **Tiền xử lý dữ liệu**

* Mục đích: chuyển đổi đồ thị từ dạng vô hướng hoặc có hướng không trọng số thành một đồ thị vô hướng có trọng số theo một motif cho trước.
* Các bước thực hiện:
  + Input: một đồ thị không có trọng số, một motif M
  + Tạo thành ma trận kề motif WM có chứa node i và j là tần suất xuất hiện của các node i và j theo motif M.
  + Output: Một đồ thị vô hướng có trọng số

## **Đặc trưng**

* **APPR**
* Input: Đồ thị vô hướng có trọng số có tập đỉnh Vw, tập cạnh Ew, tập thể hiện của motif M W, seed node u, tham số α, dung sai ε
* Gán vector (v) có giá trị bằng 0 với các đỉnh v với v là các node còn lại trong đồ thị ngoại trừ u
* Rank(u) = 1, rank(v) = 0
* Ma trận bậc của node v dw(v) là tổng số cạnh với điều kiện cạnh này là cạnh của motif thuộc tập thể hiện của motif M và chứa node v

dw(v) =

* Trong khi r(v)/dw(v) ≥ ε với mọi v Vw (tập các node V) thì
  + p = r(v) - dw(v) (phần dư của node v)
  + (v) = (v) + (1 – α)p (điểm ppr của node v)
  + r(v) = dw(v) (rank mới của node v)
  + Với mỗi x mà (v, x) Ew thì
    - r(x) = r(x) + αp

W(v, x) là số thể hiện của motif M chứa v và x

r(x) là rank của node x

* Output: vector APPR có trọng số p
* **Sweep**
* Input: Đồ thi G, vector APPR có trọng số p, node gốc u
* Gán Dw = diag(We) với Dw là ma trận đường chéo
* 𝜎i = lấy mục nhỏ nhất thứ I của
* Tìm tập hợp tiền tố có độ dẫn motif trung bình nhỏ nhất
* S = arg minl với Sl = { 𝜎1, 𝜎2, … , 𝜎l }
* Output: tập các node chứa node gốc với độ dẫn nhỏ nhất

## **Phương pháp máy học**

* Sử dụng thuật toán MAPPR là một phương pháp máy học học không có giám sát, số cụm đầu ra không được biết trước để phân cụm
* Các bước chính trong thuật toán MAPPR
  + B1: Xây dựng một đồ thị có trọng số W trong đó Wij là thể hiện của motif M chứa node i và j
  + B2: Tính vector PPR gần đúng cho đồ thị trọng số này
  + B3: Sử dụng trình quét (sweep) để xuất ra cụm có độ dẫn tối thiểu

# **III. CÀI ĐẶT THỰC NGHIỆM**

## **Dữ liệu thực nghiệm**

* Các dataset được sử dụng trong quá trình thực nghiệm:
* Dữ liệu thực tế:

Dữ liệu vô hướng:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type | Node | Edges | Communities | Description |
| Com-DBLP | Undirected | 317,080 | 1,049,866 | 13,477 | DBLP collapboration network |
| Com-amazon | Undirected | 334,863 | 925,872 | 75,149 | Amazon product network |
| Com-Youtube | Undirected | 1,134,890 | 2,987,624 | 8,385 | Youtube online social network |
| Com-LiveJournal | Undirected | 3,997,962 | 34,681,189 | 287,512 | LiveJournal online social network |
| Com-Orkut | Undirected | 3,072,441 | 117,185,083 | 6,288,363 | Orkut online social network |
| Com-Friendster | Undirected | 65,608,366 | 1,806,067,135 | 957,154 | Friendster online social network |

Mô tả dataset:

* **com-Amazon** đại diện cho việc cùng mua hàng thường xuyên trên trang bán lẻ trực tuyến Amazon. Các cộng đồng là các thành phần được kết nối của đồ thị con do các nút trong một danh mục sản phẩm tạo ra ( Yang và Leskovec , 2012 ) .
* **com-DBLP** đại diện cho quyền đồng tác giả trên DBLP. Các cộng đồng là các thành phần được kết nối của tiểu mục do các cá nhân đã xuất bản tại một hội nghị cụ thể hoặc trên một tạp chí cụ thể đưa ra (Yang và Leskovec , 2012 ) .
* **com-Youtube** thể hiện tình bạn giữa những người dùng trên mạng xã hội trực tuyến YouTube. Các cộng đồng là thành phần được kết nối của các nhóm do người dùng xác định (Mislove et al. , 2007) .
* **com-LiveJournal** đại diện cho tình bạn giữa những người dùng trên mạng xã hội trực tuyến LiveJournal. Các cộng đồng là thành phần được kết nối của các nhóm do người dùng xác định ( Mislove et al. , 2007 ) .
* **com-Orkut** thể hiện tình bạn giữa những người dùng trên mạng xã hội trực tuyến Orkut. Các cộng đồng là thành phần được kết nối của các nhóm do người dùng xác định ( Mislove et al. , 2007 ) .
* **com-Friendster** đại diện cho tình bạn giữa những người dùng trên mạng xã hội trực tuyến Friendster. Các cộng đồng là thành phần kết nối của các nhóm do người dùng xác định ( Yang và Leskovec , 2012 ) .

Dữ liệu có hướng:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type | Node | Edges | Communities | Description |
| Email-EU | Directed | 1,005 | 25,571 | 42 | E-mail network |
| Wiki-topcats | Directed | 1,791,489 | 28,522,807 | 17,364 | Wikipedia hyperlinks |

Mô tả các dataset:

* **email-Eu-core** là một mạng lưới thư điện tử giữa các thành viên của một tổ chức nghiên cứu Châu Âu. Thành viên bộ phận của các nhà nghiên cứu xác định cộng đồng mạng thực.
* **wiki-cat** là mạng siêu liên kết của Wikipedia tiếng Anh. Các danh mục bài viết là các cộng đồng mạng thực (chúng tôi chỉ xem xét 100 danh mục cho phân tích của chúng tôi) ( Klymko và cộng sự , 2014 ) .
* Dữ liệu mạng tổng hợp:

|  |  |  |
| --- | --- | --- |
| Dataset | |V| | |E| |
| CA-CONDMAT | 15,147 | 75,623 |
| EMAIL-ENRON | 18,561 | 156,139 |
| FB-HARVARD1 | 13,319 | 793,410 |
| WEB-GOOGLE | 393,582 | 2,905,337 |

## **Độ đo đánh giá**

* Độ đo đánh giá chủ yếu của thuật toán là F1 Score.
* Cách thực hiện:
* Với mỗi ground truth community, dùng mỗi node như là 1 node nguồn để được 1 tập
* Chọn tập F1 điểm cao nhất mà khôi phục ground truth
* Tính trung bình F1 trên tất cả cộng đồng
* Giá trị F1 đánh giá mức độ hội tụ của một cộng đồng tốt như thế nào.
* Công thức tính F1:

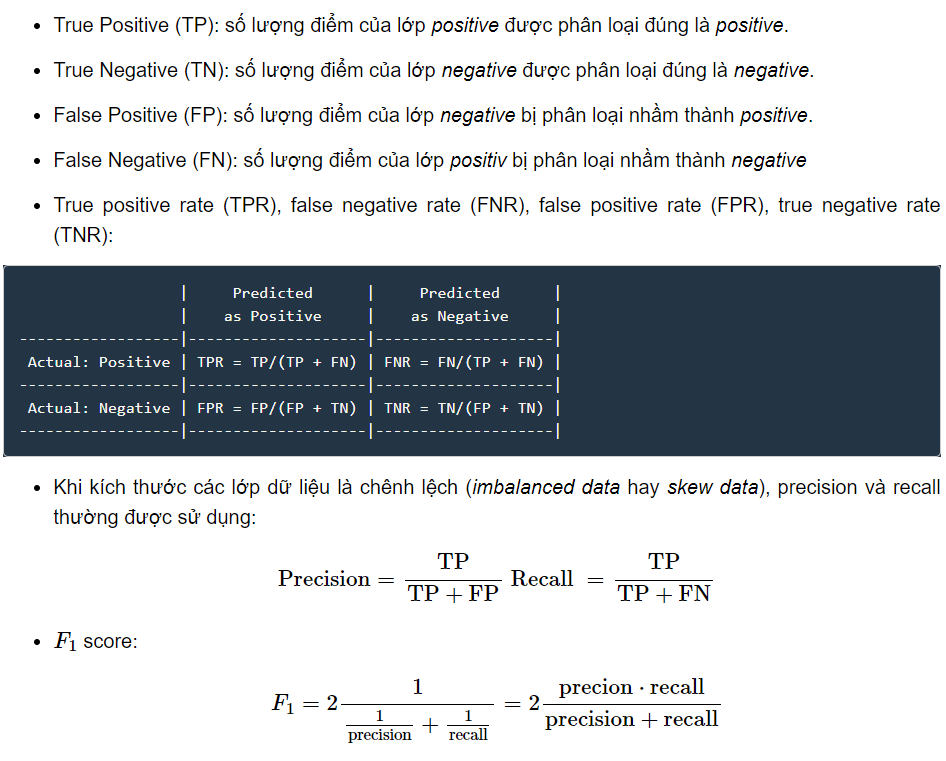
Trong đó:

* Precision được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm mô hình dự đoán là Positive.
* Recall được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm thật sự là Positive (hay tổng số điểm được gán nhãn là Positive ban đầu).

Công thức tính Precision và Recall:

Precision =

Recall =



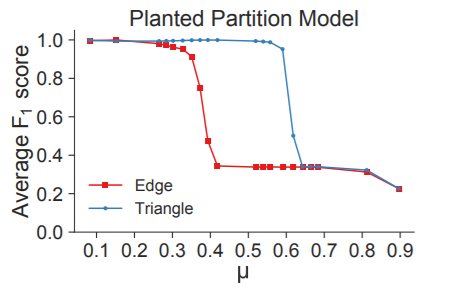
## **Phương pháp thực nghiệm**

* Phương pháp thực nghiệm trải qua hai bước chính:
* Đánh giá hiệu suất của thuật toán MAPPR trên mạng bằng một nhóm hoặc một cộng đồng sự thật nền tảng (ground- truth[1]) bao gồm cả mạng ảo tổng hợp (synthetic networks).
* Đánh giá hiệu suất của thuật toán MAPPR trên mạng thế giới thực (real-world networks).
* Quy trình đánh giá đối với APPR và MAPPR dựa trên các cạnh như sau:
* Đối với mỗi cộng đồng ground-truth, sử dụng mỗi nút làm hạt giống để có được một tập hợp và sau đó chọn ra tập hợp có điểm F1 cao nhất để khôi phục ground-truth
* Tiếp theo, tính trung bình của điểm F1 trên tất cả các cộng đồng được phát hiện về độ phát hiện chính xác của phương pháp. Phép đo này cho biết cộng đồng có thể được phục hồi tốt hay không.

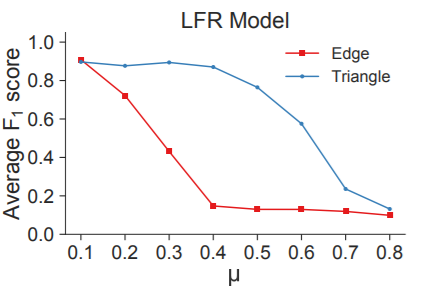
\* [1] ground truth là một mạng thực tế.

***Thực nghiệm trên các mạng khác nhau***

* Khôi phục tập ground-truth trong hai mô hình đồ thị tổng hợp ngẫu nhiên là Planted partition model và LFR model để đánh giá hiệu suất của thuật toán MAPPR.
  + **Với Planted partition model:** Tạo ra một đồ thị vô hướng không có trọng số với kn1 nút. Các nút được chia thành k cụm và mỗi cụm có kích thước là n1, p là xác xuất tồn tại một cạnh và q là xác xuất tồn tại một cạnh từ hai nút bất kỳ trong các cụm khác nhau. Sau đó, đối với APPR edge-based tiến hành sửa các giá trị n1 = 50, k = 10, p = 0.5 và chọn các giá trị ngẫu nhiên cho q để tính μ(mức trộn) theo công thức µ = [(k − 1)q]/[p + (k − 1)q], đối với MAPPR sử dụng motif là tam giác và tính độ dẫn tam giác. Cuối cùng so sánh hiệu xuất APPR edge-based và MAPPR triangle-based.



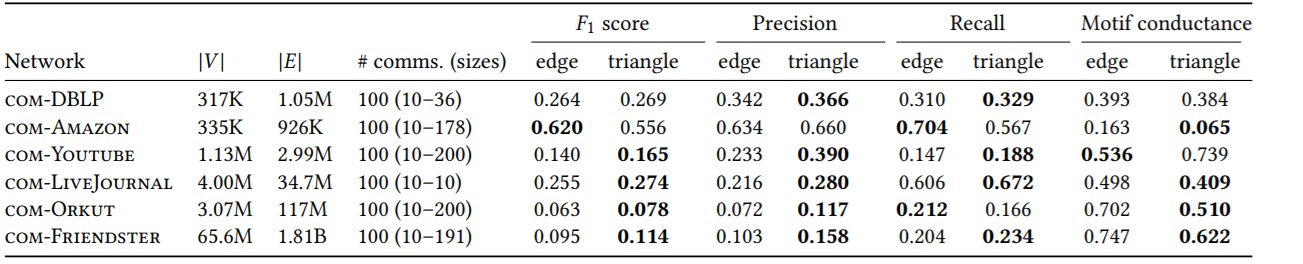
* + - Phương pháp MAPPR triangle-based tốt hơn đáng kể so với phương pháp APPR edge-based khi μ ∈ [ 0,4 , 0,6 ]
    - Phương pháp edge-based làm giảm hiệu suất khi mức trộn tăng dần bởi vì số lượng các cạnh trong cộng đồng và bên ngoài cộng đồng là gần bằng nhau còn đối với Motif tam giác thì số lượng Motif tam giác trong cộng đồng vẫn lớn hơn số lượng Motif tam giác ngoài cộng đồng nên phương pháp triangle-based cũng có thể khôi phục cộng đồng tốt hơn
  + **Với LFR model:** Cũng tạo ra một đồ thị tổng hợp ngẫu nhiên với planted communities. Sau đó sửa các giá trị n(số lượng nút) = 1000, 500 nút thuộc cụm 1 và 500 nút thuộc cụm 2, k thuộc [43,50], AvgDegree = 20 và MaxDegree = 50.
  + Tiến hành so sánh hiệu xuất APPR edge-based và MAPPR triangle-based với giá trị μ thuộc [0.1,0.4] và μ > 0.6.



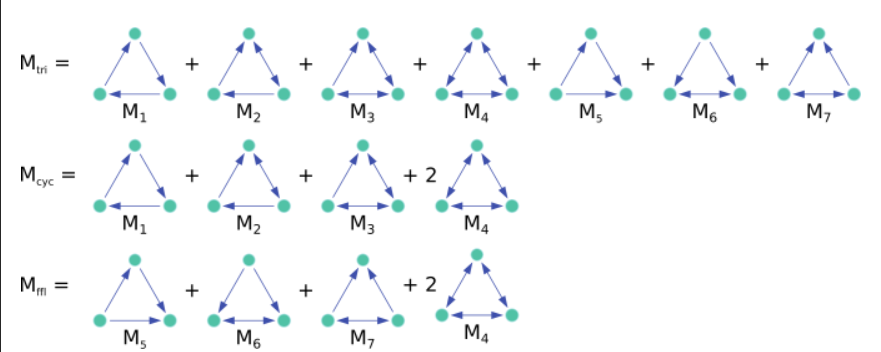
* + - Hiệu suất của phương pháp edge-based giảm dần khi tăng mức trộn μ từ 0,1 đến 0,4 , trong khi phương pháp triangle-based duy trì điểm F1 xấp xỉ 0,9
    - Đối với các mức trộn lớn đến 0,6, điểm F1 cho phương pháp triangle-based vẫn lớn hơn ba lần so với phương pháp edge-based và trong gần như toàn bộ các mức trộn, sử dụng Motif tam giác cải thiện hiệu suất phục hồi cộng đồng.
* AvgDegree: Bậc trung bình

***Khôi phục tập ground-truth trên các mạng thế giới thực để đánh giá hiệu suất của MAPPR***

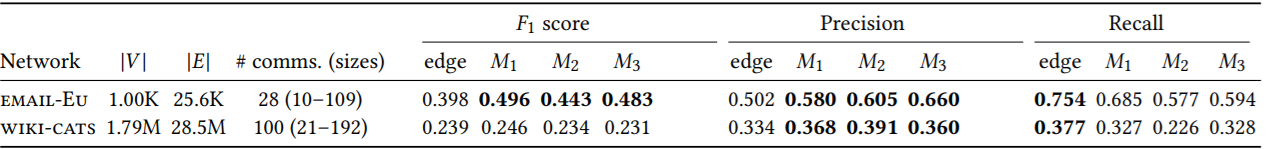
* Đối với đồ thị vô hướng:
  + Sử dụng các tập ground-truth được lấy từ các công ty lớn như com-Amazon, com-DBLP, com-Youtube, com-LiveJournal, com-Orkut để so sánh MAPPR và APPR
  + Đối với mỗi mạng, kiểm tra 100 cộng đồng có quy mô dao động từ 10 đến 200 nút.



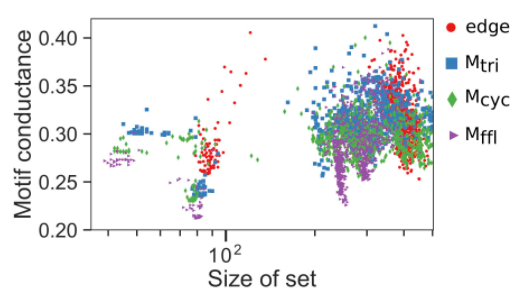
* Đối với đồ thị có hướng:
  + Sử dụng 3 motif có hình dáng khác nhau là Mtri, Mcyc, Mffl Với 1 trường hợp của Mtri, 2 trường hợp của Mcyc và 2 trường hợp của Mffl



* + Và 2 mạng thế giới thực có hướng là EMAIL-EU và WIKI-CAST. Sau đó sử dụng M1 để so sánh hiệu xuất của phương pháp APPR edge-based và APPR triangle-based theo kết quả bảng bên dưới.



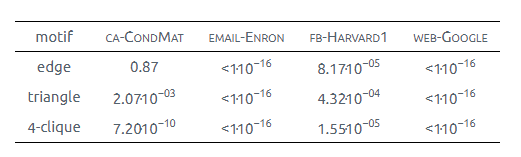
* + - Bảng thể hiện các Motif như sau (M1, the undirected triangle; M2, the cycle; and M3, the feed-forward loop)
    - Đối với cả hai mạng, sử dụng mô-típ M tri giúp cải thiện điểm F1 so với phương pháp edge-based. Sự cải thiện rất mạnh mẽ trong email-Eu-core (cải thiện tương đối 25%).
    - Cả ba mô-típ đều dẫn đến những cải tiến đáng kể trong mạng lưới này. Trong cả hai mạng, các mô típ cung cấp thêm độ chính xác nhưng làm mất đi khả năng thu hồi.



* + Phân bố kích thước của tập node và độ dẫn trong email-Eu-core
  + Đối với những cộng đồng nhỏ khoảng từ 70 đến 100 nodes thì độ dẫn motif nhỏ hơn độ dẫn cạnh và kết quả cho ra cộng đồng là gần như tương tự nhau
  + Đối với những cộng đồng lớn thì độ dẫn motif cho ra kết quả cộng đồng nhỏ hơn độ dẫn cạnh

***Tìm nhiều nút hạt giống tốt:***

* Một node u được gọi là cực tiểu cục bộ (Local minimum) nếu như:
  + Với NU bao gồm node u và hàng xóm của u, Nv tương tự
* Để kiểm tra xem Local minumun có phải là seed node cho APPR hay không, ta tiến hành như sau
  + trước tiên ta tính toán toàn bộ các cụm APPR bầng cách sử dụng mọi node trong mỗi mạng làm seed node
  + Sử dụng kiểm định Mann Whitney U một phía để kiểm tra giả thuyết rỗng rằng Global minimum có độ dẫn motif nhỏ hơn so với Local minimum, kết quả chứng minh giả thuyết này bị bác bỏ với mức ý nghĩa p < 0.003 với tất cả các Clique và mạng được quan sát ngoại trừ nhóm 2-clique (edge) trong mạng ca-CondMat
* Nói cách khác, Local minimum là seed node tốt hơn so với Global minimum

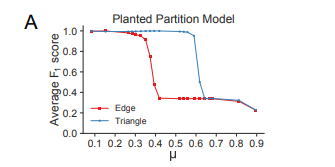
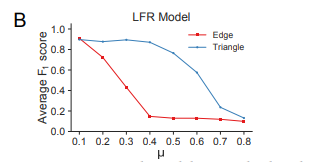


Sử dụng những local minimum này để cấu trúc hồ sơ cộng đồng mạng cho các motif khác nhau (NPC là độ dẫn tối ưu đối với cụm có kích thước cố định ). Hình dạng của biểu đồ thể hiện cấu trúc cụm của mạng. Việc sử dụng seed node local minimum là đủ để nắm bắt các xu hướng chính cuẩ biểu đồ NPC. Cực tiểu cục bộ cho các nhóm bậc thấp hơn cũng có thể đươc sử dụng làm hạt giống tốt khi tìm kiếm các cụm.

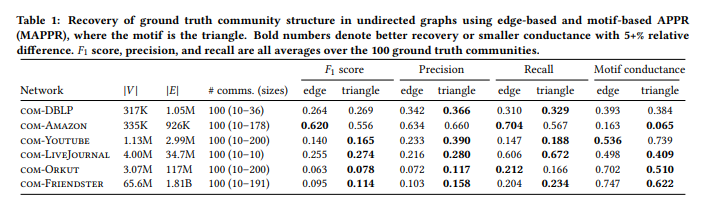
## **Kết quả thực nghiệm**

* **Khôi phục mạng nhân tạo**: khi thử MAPPR để khôi phục cho 2 mô hình đồ thị ngẫu nhiên the planted partion model và LFR model. Ta thấy rằng dùng motif tam giác thì tăng thì sẽ tang phạm vi tham số của ground truth communities mà ta có thể khôi phục.
  + *Planted partion model*:
    - Sử dụng tam giác với MAPPR có hiệu quả tốt hơn đáng kể so với phương pháp APPR dựa trên cạnh khi μ ∈ [0,4, 0,6].
    - Trong chế độ này, đối với bất kỳ nút nhất định nào, số lượng các cạnh intra-community và inter-community ​​là gần như nhau. Do đó, phương pháp edge-based làm giảm hiệu suất. Tuy nhiên, số lượng tam giác intra-

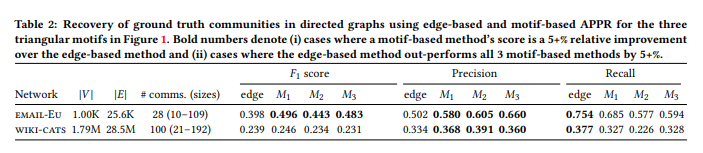
community vẫn lớn hơn số tam giác inter-community, vì vậy phương pháp triangle-based có thể khôi phục planted partion.



* + *LFR model*:
    - Hiệu suất của phương pháp edge-based giảm dần khi tăng mức trộn μ từ 0,1 đến 0,4 , trong khi phương pháp triangle-based duy trì điểm F1 xấp xỉ 0,9
    - Đối với các mức trộn lớn đến 0,6, điểm F1 cho phương pháp triangle-based vẫn lớn hơn ba lần so với phương pháp edge-based và trong gần như toàn bộ các mức trộn, sử dụng Motif tam giác cải thiện hiệu suất phục hồi cộng đồng.
* **Khôi phục những cộng đồng trong mạng thế giới thực với MAPPR:**
* Với đồ thị 1.8 tỉ cạnh, phương pháp APPR dùng ít giây nhất trên mỗi seed khi đồ thị nằm trong bộ nhớ.
* Đồ thị vô hướng:
  + - Thống kê tổng quan tập dữ liệu và thực nghiệm trong bảng 1. 5 trong 6 bảng, MAPPR đạt được điểm F1 cao hơn edge-based APPR, 3 trong 5 mạng, điểm F1 cung cấp sự cải thiện liên quan khoảng 5%.
    - Trong tất cả 6 mạng, độ chính xác trung bình của các cụm được khôi phục lớn hơn, 4 trong số đó, sự thay đổi lớn hơn 5 %. 4 trong 6 mạng, việc khôi phục theo phương pháp triangle-based cung cấp những cải thiện liên quan ít nhất 5%.



* + Đồ thị có hướng: 1 lợi điểm của MAPPR có thể dễ dàng xử lí đồ thị có hướng
    - Bảng thể hiện: M1: the undirected triangle, M2: chu kì, M3: lặp chuyển tiếp..
    - Cả 2 mạng dùng M1 cải thiện cho điểm F1 qua phương pháp edge-based. Sự cải thiện rõ rệt ở Email-EU (cải thiện liên quan 25%).
    - Sau khi tìm hiểu chi tiết hơn về Email-EU, thấy rằng 3 motif là 3 mô hình tốt hơn cho cấu trúc cộng đồng trong mạng. Ta có thể thấy rằng các cụm lớn thì những cụm edge-based lớn nhất. Vì những tập này lớn hơn kích cỡ của những cộng đồng trong mạng, những điều này nói lên rằng dùng motif tam giác thì độ chuẩn xác tốt hơn MAPPR.



# **IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **1. Kết luận**

* Việc sử dụng các tam giác motif sẽ cho khả năng khôi phục các cộng đồng tốt hơn
* Ứng dụng tốt hơn trong xếp hạng và phát hiện cộng đồng trên web, mạng xã hội, mạng sinh-tin học (phân tích tương tác giữa các protein)
* Khả năng phục hồi các cộng đồng mạng quan trọng cao hơn và khả năng tìm được nhiều node gốc tốt hơn

## **2. Hướng phát triển**

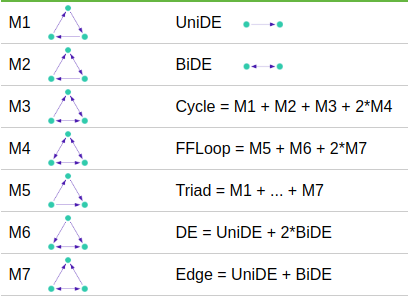
* Hiện tại việc tìm good seed node chỉ có thể áp dụng trên đồ thị vô hướng, vì vậy thuật toán cần được cải thiện thêm để có thể tìm được các good seed node trên đồ thị có hướng.

# **V. DEMO**

* **Bộ dữ liệu thử nghiệm**
* Trong phần này chúng ta sẽ chạy thử thuật toán MAPPR với bộ dữ liệu mạng com-youtube, bộ dữ liệu là một mạng vô hướng với khoảng 1.13 triệu node và 2.98 triệu cạnh thể hiện tình bạn giữa những người dùng trên mạng xã hội trực tuyến YouTube
* Bộ dữ liệu mạng để test: <http://snap.stanford.edu/data/com-Youtube.html>
* Mô tả các tham số của input
  + -i: đường dẫn đến bộ dữ liệu mạng
  + -m: motif sử dụng cho mạng này (mặc định là clique3)

Nếu là đồ thị vô hướng sử dụng một motif bất kì từ 2 - 9: "UEdge", "clique3", "clique4", ..., "clique9".

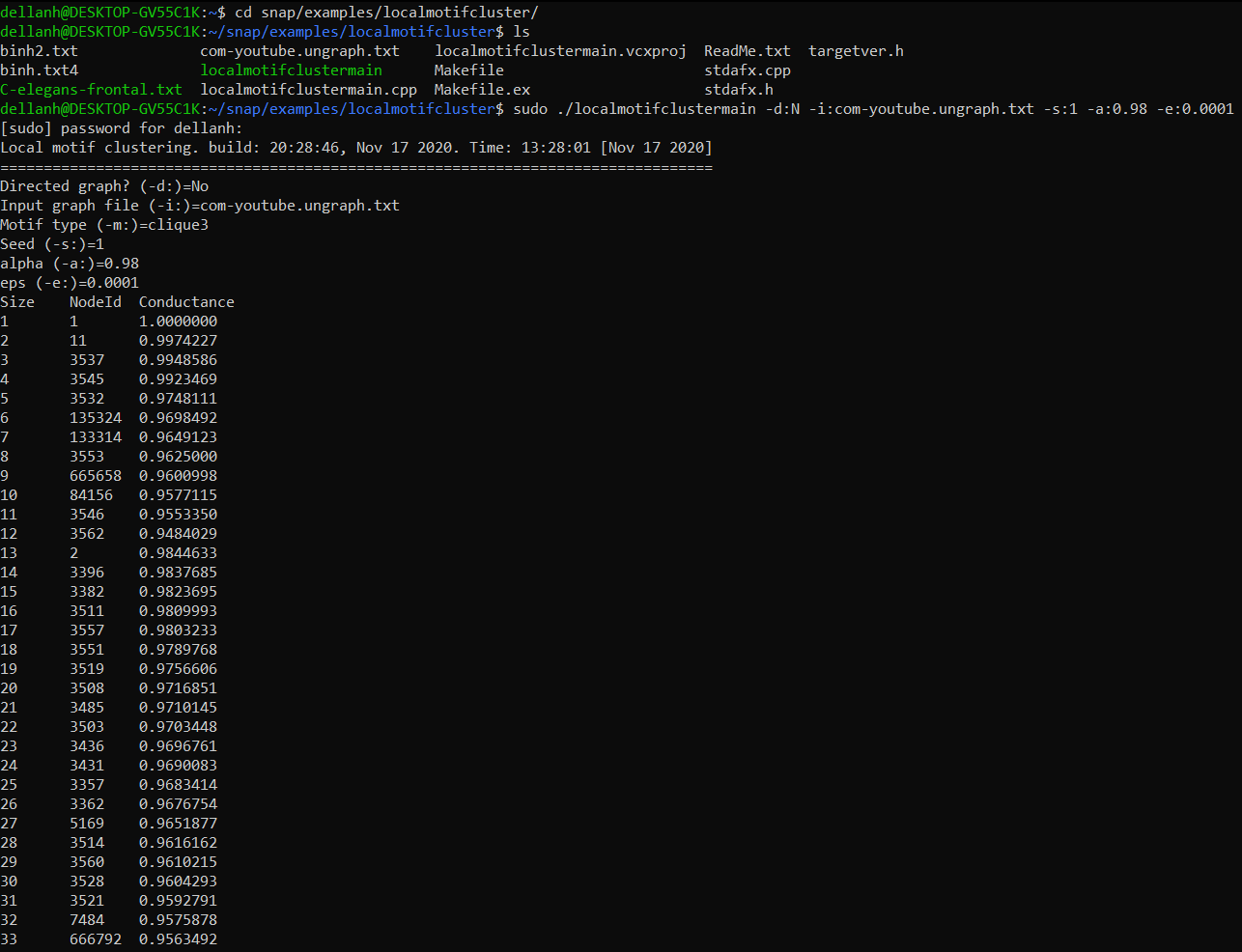
Nếu là đồ thị có hướng sử dụng một motif bất kì theo bảng dưới đây

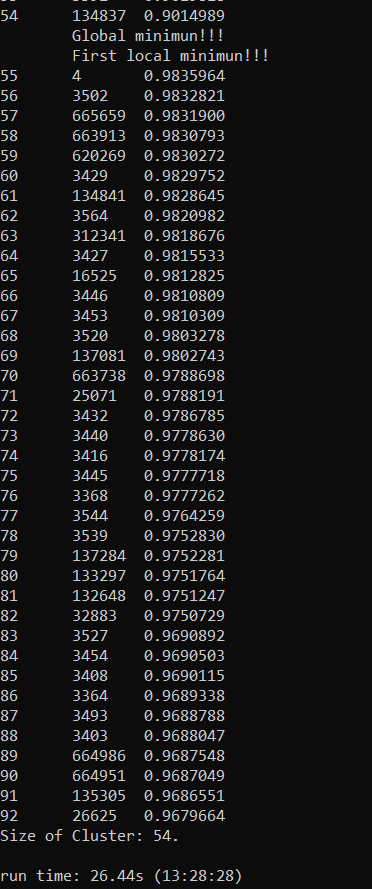


* + -d: chỉ định đồ thị là có hướng hay vô hướng N là vô hướng và Y là có hướng
  + -s: chỉ định một node id trong mạng cho trước làm node gốc (mặc định là node có id là 1)
  + -a: là xác xuất bước nhảy (mặc định a = 0.98)
  + -e: là độ sai lệch (mặc định e = 0.0001)
* **Quá trình demo**
  + Màn hình nhập các thông số ban đầu



* Màn hình kết quả



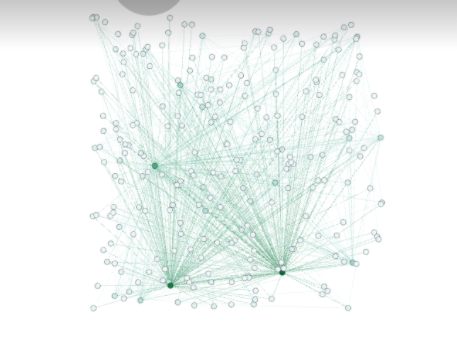


* Kết quả từ ảnh trên cho ta thấy được kích thước của cụm có độ dẫn motif tối thiểu là 54 node, ở trên là các node id và độ dẫn của node trong cụm này.

# **VI. Thuật toán**

-Mô tả dữ Liệu : Dữ liệu bao gồm các Node là người dùng youtube,mô tả quan hệ bạn bè giữa các người dùng với nhau.

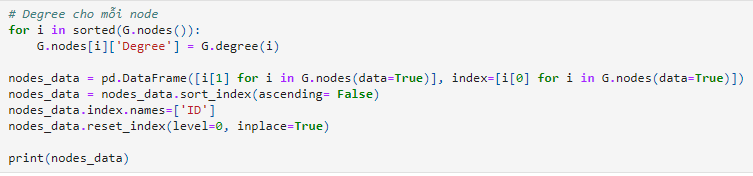
- Dataset: http://snap.stanford.edu/data/com-Youtube.html?fbclid=IwAR0aGDib8JyPAlfzi42Tvs9HcUGJCe2kg7HFIG8Xzx4n7X263QTUNurejPQ



## **1. Degree Centrality**

1. Giải thuật:

Degree Centrality( Độ trung tâm cấp bậc ): là số lượng các kết nối trực tiếp mà một node có được. 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.

1. Code:

1. Kết quả:

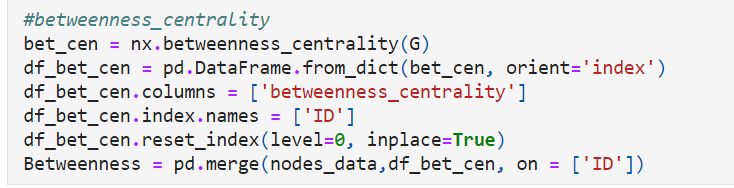


## **2. Betweeness Centrality**

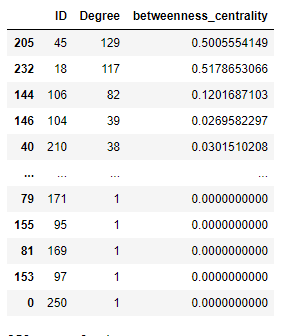
1. Giải thuật:

(Độ đo trung tâm trung gian) là Mức trung tâm xen giữa 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. Một kết nối có tính trung gian cao sẽ ảnh hưởng đến cái gì sẽ không truyền được đi trong mạng.

1. Code



1. Kết Quả

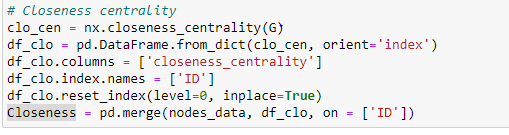


## **3. Closeness Centrality**

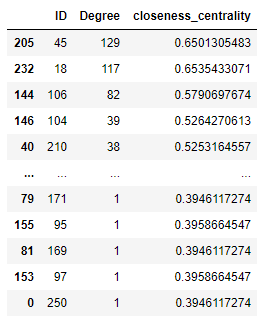
1. Giải thuật:

(Độ đo trung tâm kề cận) là Mức trung tâm gần được đo bằng khoảng cách trung bình ngắn nhất tới tất cả các node khác trong mạng

1. Code



1. Kết Quả

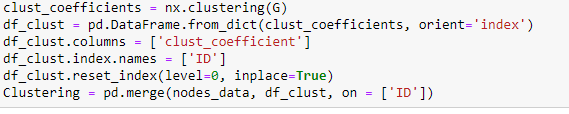


## **4. Clustering coefficient**

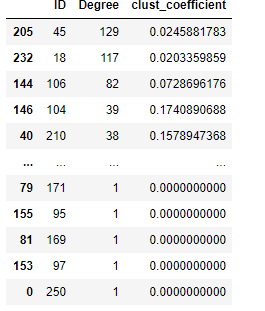
1. Giải thuật:

(Hệ số cụm) là một thước đo mức độ xu hướng kết nhóm lại với nhau của các node trong đồ thị. Những node có xu hướng tạo các nhóm liên kết chắc chẽ thường có đặc tính là có mật độ tương đối cao các mối ràng buộc. xác suất xuất hiện các mỗi ràng buộc này thường có xu hướng lớn hơn xác suất trung bình một ràng buộc được thiết lập ngẫu nhiên giữa hai node.

1. Code



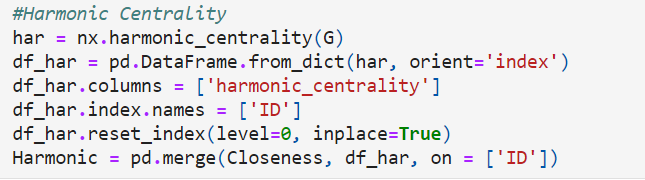
1. Kết Quả:

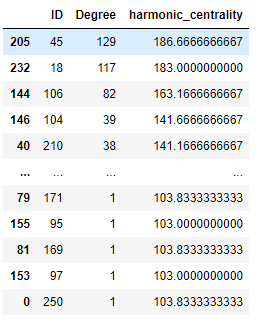


## **5. Harmonic Centrality**

1. Giải thuật:

Là thuật toán được phát hiện để thay thế closeness centrality. Thay vì tính nghịch đảo của tổng khoảng cách ngắn nhất đến một node thì thuật toán này sẽ tính nghịch đảo trên từng khoảng cách nhỏ nhất đến một node. Vì vậy, Harmonic Centrality sẽ tính được vô hạn, khắc phục hạn chế của Closeness.

1. Code
2. Kết Quả:



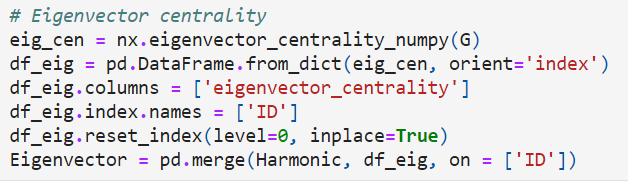
## **6. Eigenvector Centrality**

1. Giải thuật:

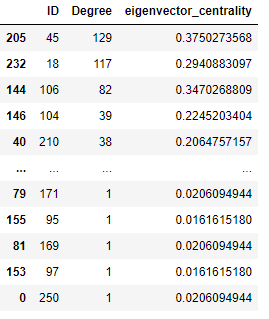
(Độ trung tâm dựa trên giá trị riêng) chỉ ra nút đó có độ trung tâm nhất trong toàn mạng. Nó được tính bằng 

với Av,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.

1. Code



1. Mô hình



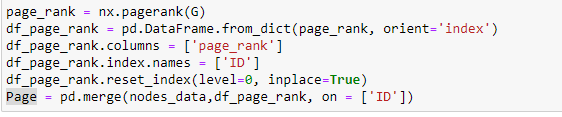
## **7. Page Rank**

1. Giải thuật:

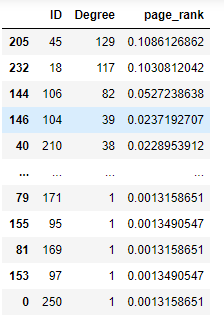
Là thuật toán phân tích các liên kết được dùng trong Google Search để xếp hạng các trang web.

* Thuật toán này chỉ định giá trị nhất định cho mỗi thành phần của một tập hợp các văn bản liên kết với nhau, ví dụ như World Wide Web.
* Mục đích "đo" tầm quan trọng tương đối của các liên kết trong tập hợp đó.
* Áp dụng cho bất kỳ tập hợp văn bản nào có trích dẫn đối ứng và liên kết cụ thể.
* giá trị (weight) mà nó gán cho bất kỳ thành phần E được gọi là PageRank của E và ký hiệu là PR(E). {\displaystyle PR(E).}

1. Code



1. Mô hình



# ***Đánh giá quá trình thực hiện đồ án***

Lê Đức Vinh : 100%

Ngô Anh Vũ : 100%

Lê Như Toàn : 100%

# **TÀI LIỆU THAM KHẢO**

* <https://ieeexplore.ieee.org/abstract/document/4031383>
* <https://caihuuthuc.wordpress.com/2020/02/23/precision-recall-va-f1-score-la-gi/>
* [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5951164](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5951164/?fbclid=IwAR1JvAMXrQSmVGp64LeM_lx5bcnHGyslHylsxwNByB2j3aXkJGros_GSo6M#R4)
* <https://science.sciencemag.org/content/353/6295/163.abstract>
* <https://journals.aps.org/pre/abstract/10.1103/PhysRevE.97.052306>

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