TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ MÔN XỬ LÝ DỮ LIỆU LỚN**

**LINK PREDICTION BASED ON GRAPH NEURAL NETWORKS**

*Người hướng dẫn*: **TS BÙI THANH HÙNG**

*Người thực hiện*: **MAI HOÀNG MINH – 51703130**

**NGUYỄN VĂN TINH – 51703199**

Lớp **: 17050303**

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**



**ĐỒ ÁN CUỐI KÌ MÔN XỬ LÝ DỮ LIỆU LỚN**

**LINK PREDICTION BASED ON GRAPH NEURAL NETWORKS**

*Người hướng dẫn*: **TS BÙI THANH HÙNG**

*Người thực hiện*: **MAI HOÀNG MINH – 51703130**

**NGUYỄN VĂN TINH – 51703199**

Lớp **: 17050303**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

LỜI CẢM ƠN

Được sự phân công của TS Bùi Thanh Hùng em đã tìm hiểu và thực hiện các yêu cầuvề đề tài LINK PREDICTION BASED ON GRAPH NEURAL NETWORKS .

Qua đồ án lần này chúng em đã hiểu và nắm rõ hơn về cấc vấn đề liên quan đến đề tài. Để được có thành quả như vậy chúng em xin gửi lời cảm ơn tới các thầy cô giáo khoa CNTT trường đại học Tôn Đức Thắng và đặc biệt là TS Bùi Thanh hùng đã hướng dẫn chúng em hoàn thành đồ án này.

Chúng em đã cố gắng hoàn thành đồ án một cách tốt nhất nhưng sẽ có những thiếu sót mà chúng em không nhận thấy. Chúng em mong nhận được những phản hồi và ý kiến của các thầy cô để bài đồ án sau chúng em có thể hoàn thành tốt hơn, chúng em xin chân thành cảm ơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Bùi Thanh Hùng;. 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 đồ án 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 đồ án 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 (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Mai Hoàng Minh*

Cán bộ hướng dẫn khoa học : ....................................................................... (Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Cán bộ phản biện 1: ....................................................................... (Ghi rõ họ, tên, học hàm, học vị và đơn vị công tác)

Cán bộ phản biện 2: ........................................................................ (Ghi rõ họ, tên, học hàm, học vị và đơn vị công tác) Luận văn thạc sĩ được bảo vệ tại HỘI ĐỒNG CHẤM BẢO VỆ LUẬN VĂN THẠC SĨ TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG, ngày . . . . . tháng . . . . năm . . . . theo Quyết định số ………./20 /TĐT-QĐ-SĐH ngày …./…./…….

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

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

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

Hầu hết các nền tảng truyền thông xã hội, bao gồm cả Facebook và twitter, có thể được cấu trúc dưới dạng đồ thị. Những người dùng đã đăng ký được kết nối với nhau trong một vũ trụ mạng. Và để làm việc trên các mạng và đồ thị này, chúng ta cần một bộ phương pháp tiếp cận, công cụ và thuật toán khác (thay vì các phương pháp học máy truyền thống). Vì vậy, trong bài viết này, chúng ta sẽ giải quyết một vấn đề mạng xã hội với sự trợ giúp của đồ thị và học máy

Trước tiên, chúng tôi triển khai và áp dụng nhiều phương pháp dự đoán liên kết cho từng mạng bản ngã có trong tập dữ liệu SNAP Facebook và tập dữ liệu SNAP Twitter, cũng như cho các mạng ngẫu nhiên khác nhau được tạo bằng networkx, sau đó tính toán và so sánh ROC AUC, Độ chính xác trung bình và thời gian chạy của mỗi phương thức.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc57808825)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc57808826)

[TÓM TẮT iv](#_Toc57808827)

[MỤC LỤC 1](#_Toc57808828)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4](#_Toc57808829)

[PHẦN I 5](#_Toc57808830)

[CHƯƠNG 1 5](#_Toc57808831)

[PHÂN TÍCH BÀI TOÁN 5](#_Toc57808832)

[1.1 Giới thiệu về bài toán 5](#_Toc57808833)

[1.2 Phân tích yêu cầu của bài toán 5](#_Toc57808834)

[1.3 Hướng giải quyết 6](#_Toc57808835)

[CHƯƠNG 2 7](#_Toc57808836)

[2.1 Giới thiệu chung về các phương pháp giải quyết bài toán 7](#_Toc57808837)

[Mô hình tổng quát giải quyết bài toán link prediction 7](#_Toc57808838)

[2.2 Phương pháp 1 Node2Vec 7](#_Toc57808839)

[2.3 Phương pháp 2 : spectral clustering 8](#_Toc57808840)

[2.4 Phương pháp 3 : Adamic-Adar, Jaccard Coefficient, Preferential Attachment 12](#_Toc57808841)

[2.4 Phương pháp 4 - Variational Graph Autoencoders 14](#_Toc57808842)

[CHƯƠNG 3 17](#_Toc57808843)

[THỰC NGHIỆM 17](#_Toc57808844)

[3.1 Dữ liệu 17](#_Toc57808845)

[3.2 Xử lý dữ liệu 18](#_Toc57808846)

[3.3 Công nghệ sử dụng 18](#_Toc57808847)

[3.4 Cách đánh giá 19](#_Toc57808848)

[3.5 Kết quả đạt được 19](#_Toc57808849)

[CHƯƠNG 4 20](#_Toc57808850)

[KẾT LUẬN 20](#_Toc57808851)

[4.1 Kết luận 20](#_Toc57808852)

[4.2 Hướng phát triển 20](#_Toc57808853)

[PHẦN 2 21](#_Toc57808854)

[Bài toán 1: 21](#_Toc57808855)

[Bài toán 2: 21](#_Toc57808856)

[Bài toán 3: 21](#_Toc57808857)

[TÀI LIỆU THAM KHẢO 22](#_Toc57808858)

[PHỤ LỤC 23](#_Toc57808859)

[TỰ CHẤM ĐIỂM 24](#_Toc57808860)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 2.1 Mô hình tổng quát giải quyết bài toán link prediction in social network 7](#_Toc57835366)

[Hình 2.4 Kiến trúc của variational graph autoencoder 15](#_Toc57835367)

[Hình 3.5 biểu đồ so sánh giữa các phương pháp theo độ đo Roc trên tập dữ liệu twitter 22](#_Toc57835368)

[Hình 3.6 biểu đồ so sánh giữa các phương pháp theo độ đo runtime trên tập dữ liệu twitter 22](#_Toc57835369)

[Hình 3.7 biểu đồ so sánh giữa các phương pháp theo độ đo Average precision trên tập dữ liệu facebook 23](#_Toc57835370)

[Hình 3.8 biểu đồ so sánh giữa các phương pháp theo độ đo roc trên tập dữ liệu facebook 23](#_Toc57835371)

[Hình 3.9 biểu đồ so sánh giữa các phương pháp theo độ đo runtime trên tập dữ liệu facebook 24](#_Toc57835372)

**DANH MỤC BẢNG**

PHẦN I

CHƯƠNG 1

PHÂN TÍCH BÀI TOÁN

1.1 Giới thiệu về bài toán

Bài toán này chứa một loạt các thử nghiệm học máy để dự đoán liên kết trong các mạng xã hội.

Hầu hết các nền tảng truyền thông xã hội, bao gồm cả twitter, có thể được cấu trúc dưới dạng đồ thị. Những người dùng đã đăng ký được kết nối với nhau trong một vũ trụ mạng. Và để làm việc trên các mạng và đồ thị này, chúng ta cần một bộ phương pháp, công cụ và thuật toán khác (thay vì các phương pháp học máy truyền thống).

Vì vậy, trong bài viết này, chúng ta sẽ giải quyết một vấn đề mạng xã hội với sự trợ giúp của đồ thị và học máy. Trước tiên, chúng ta sẽ hiểu các khái niệm và thành phần cốt lõi của dự đoán liên kết trước khi tham gia một nghiên cứu điển hình của Facebook và triển khai nó bằng Python!

1.2 Phân tích yêu cầu của bài toán

Mạng xã hội về cơ bản là đại diện cho các mối quan hệ giữa các thực thể xã hội, chẳng hạn như con người, tổ chức, chính phủ, đảng phái chính trị, v.v.

Sự tương tác giữa các thực thể này tạo ra lượng dữ liệu không thể tưởng tượng được dưới dạng bài đăng, tin nhắn trò chuyện, tweet, lượt thích, bình luận, lượt chia sẻ, v.v. Điều này mở ra cơ hội và các trường hợp sử dụng mà chúng ta có thể thực hiện.

Điều đó đưa chúng tôi đến Phân tích mạng xã hội (SNA). Chúng ta có thể định nghĩa nó là sự kết hợp của một số hoạt động được thực hiện trên mạng xã hội. Các hoạt động này bao gồm thu thập dữ liệu từ các trang mạng xã hội trực tuyến và sử dụng dữ liệu đó để đưa ra các quyết định kinh doanh.

Những lợi ích của phân tích mạng xã hội có thể rất bổ ích. Dưới đây là một số lợi ích chính:

* Giúp bạn hiểu khán giả của mình hơn
* Được sử dụng để phân khúc khách hàng
* Được sử dụng để thiết kế Hệ thống khuyến nghị
* Phát hiện tin tức giả, trong số những thứ khác

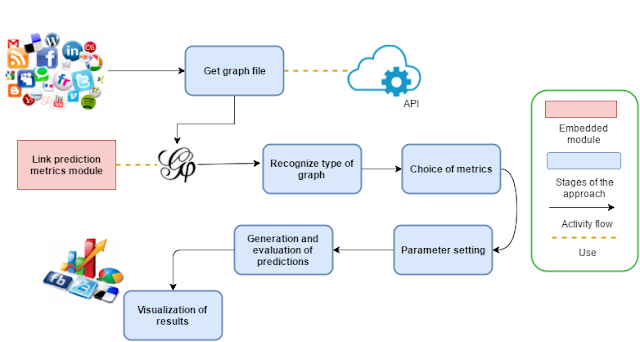
1.3 Hướng giải quyết

Trước tiên, chúng tôi triển khai và áp dụng nhiều phương pháp dự đoán liên kết cho từng mạng cái tôi có trong tập dữ liệu SNAP Twitter và SNAP Facebook, cũng như cho các mạng ngẫu nhiên khác nhau được tạo bằng networkx, sau đó tính toán và so sánh ROC AUC, Average Precision và thời gian chạy của mỗi phương thức.

CHƯƠNG 2

2.1 Giới thiệu chung về các phương pháp giải quyết bài toán

Mô hình tổng quát giải quyết bài toán link prediction in social network :



Hình 2.1 Mô hình tổng quát giải quyết bài toán link prediction in social network

2.2 Phương pháp 1 Node2Vec

Node2Vec cũng là 1 mô hình Node Embedding dựa trên ý tưởng của DeepWalk và Word2Vec. Điểm khác biệt của Node2Vec là ngoài việc sử dụng random walk như thông thường, mô hình giới thiệu thêm 2 thông số P và Q để điều chỉnh lại bước đi ngẫu nhiên trên đồ thị.

Chiến lược lấy mẫu của Node2vec, chấp nhận 4 đối số:

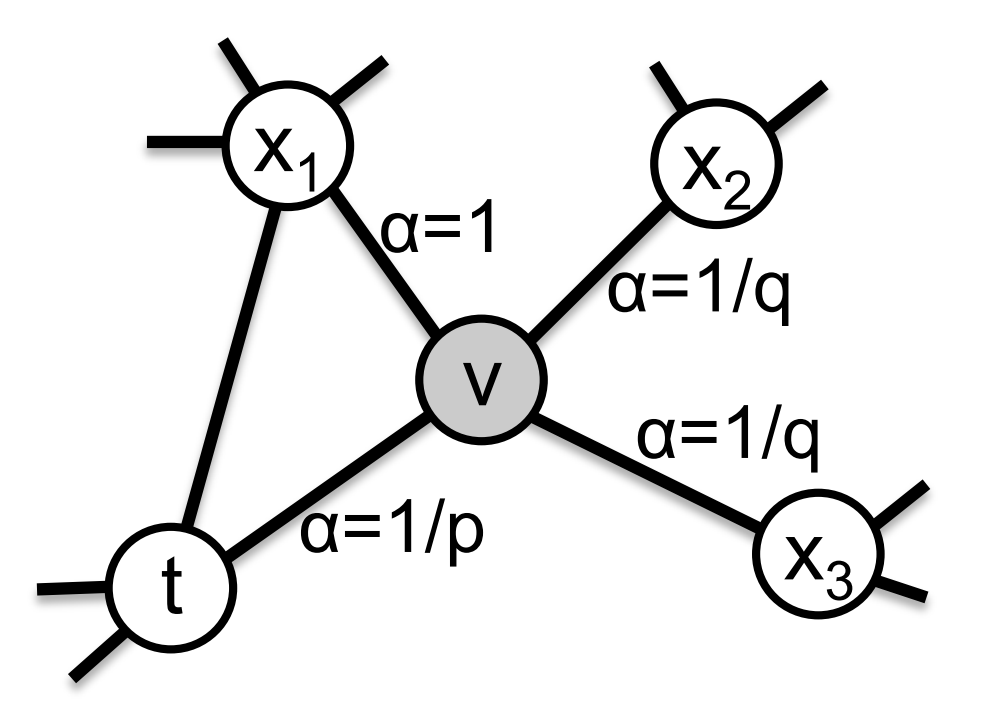
* Number of walks: Số lần đi bộ ngẫu nhiên được tạo từ mỗi nút trong biểu đồ
* Walk length: Có bao nhiêu nút trong mỗi lần đi bộ ngẫu nhiên
* P: Trả về hyperparameter
* Q: In-out hyperparameter

và cả các tham số bỏ qua tiêu chuẩn (kích thước cửa sổ ngữ cảnh, số lần lặp lại, v.v.)

Thuật toán tạo bước đi ngẫu nhiên sẽ đi qua từng nút trong biểu đồ và sẽ tạo ra < Number of walks > random walk, có độ dài <walk length>.

Q và P, được giải thích rõ hơn bằng hình ảnh.

Hãy xem xét bạn đang đi bộ ngẫu nhiên và vừa chuyển đổi từ nút <t> sang nút <v> trong sơ đồ sau



Xác suất để chuyển đổi từ <v> sang bất kỳ neighbors nào của nó là:

<edge weight> \* <α> (chuẩn hóa), trong đó <α> phụ thuộc vào hyperparameters. (siêu tham số).

P kiểm soát xác suất để quay lại <t> sau khi truy cập <v>.

Q kiểm soát xác suất để khám phá các phần chưa được khám phá của đồ thị.

Nói một cách trực quan, đây giống như tham số phức tạp trong tSNE, nó cho phép bạn nhấn mạnh cấu trúc cục bộ / toàn cục của đồ thị.

Đừng quên rằng trọng lượng cũng được tính đến, vì vậy xác suất di chuyển cuối cùng là một hàm của:

1. Nút trước trong bước đi

2. P và Q

3. Edge weight

2.3 Phương pháp 2 : spectral clustering

Phân cụm là một trong những kỹ thuật được sử dụng rộng rãi nhất để phân tích dữ liệu khám phá. Mục tiêu của nó là chia các điểm dữ liệu thành nhiều nhóm sao cho các điểm trong cùng một nhóm tương tự nhau và các điểm trong các nhóm khác nhau không giống nhau.

Phân cụm quang phổ ngày càng trở nên phổ biến do thực hiện đơn giản và hiệu suất đầy hứa hẹn trong nhiều phân nhóm dựa trên đồ thị. Nó có thể được giải một cách hiệu quả bằng phần mềm đại số tuyến tính tiêu chuẩn, và thường hoạt động tốt hơn các thuật toán truyền thống như thuật toán k-mean.

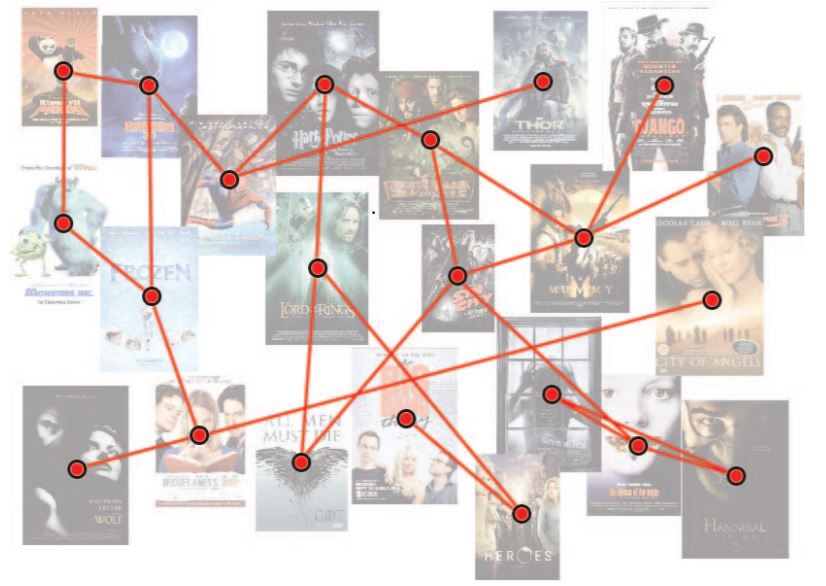
Tạo một đồ thị tương tự giữa N đối tượng của chúng ta để phân cụm.

Tính toán k eigenvector đầu tiên của ma trận Laplacian để xác định một vectơ đặc trưng cho mỗi đối tượng.

Chạy k-mean trên các tính năng này để tách các đối tượng thành k lớp.

Bước 1 :

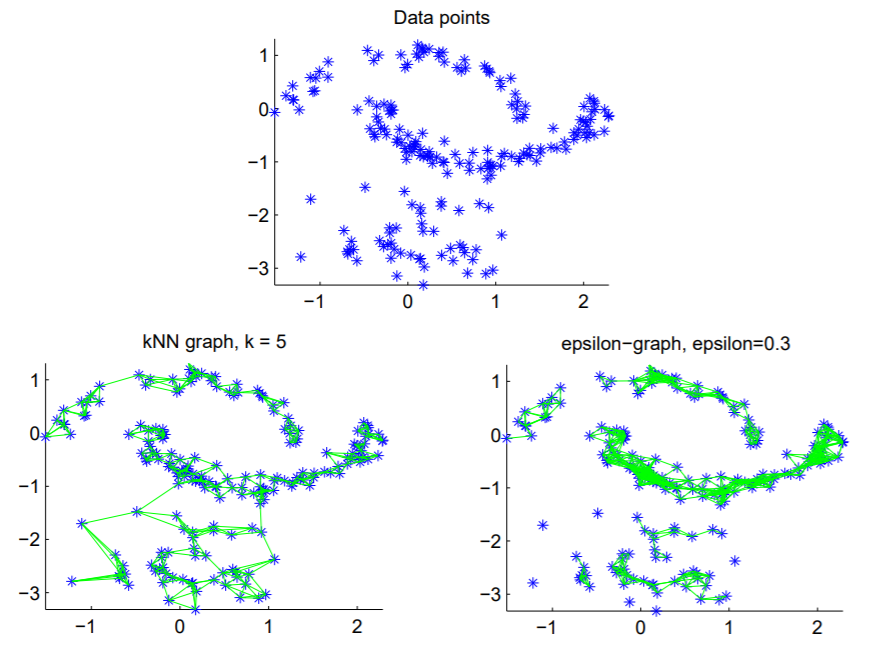
Một cách hay để biểu diễn một tập hợp các điểm dữ liệu x1 ,. . . x N có dạng đồ thị tương tự G = (V, E).



Có nhiều cách khác nhau để xây dựng một biểu đồ thể hiện mối quan hệ giữa các điểm dữ liệu:

ε-neighborhood graph : Mỗi đỉnh được kết nối với các đỉnh rơi bên trong một quả cầu bán kính ε trong đó ε là giá trị thực phải được điều chỉnh để bắt cấu trúc cục bộ của dữ liệu.

k-nearest neighbor graph : Mỗi đỉnh được kết nối với k lân cận gần nhất của nó với k là số nguyên kiểm soát các mối quan hệ cục bộ của dữ liệu.



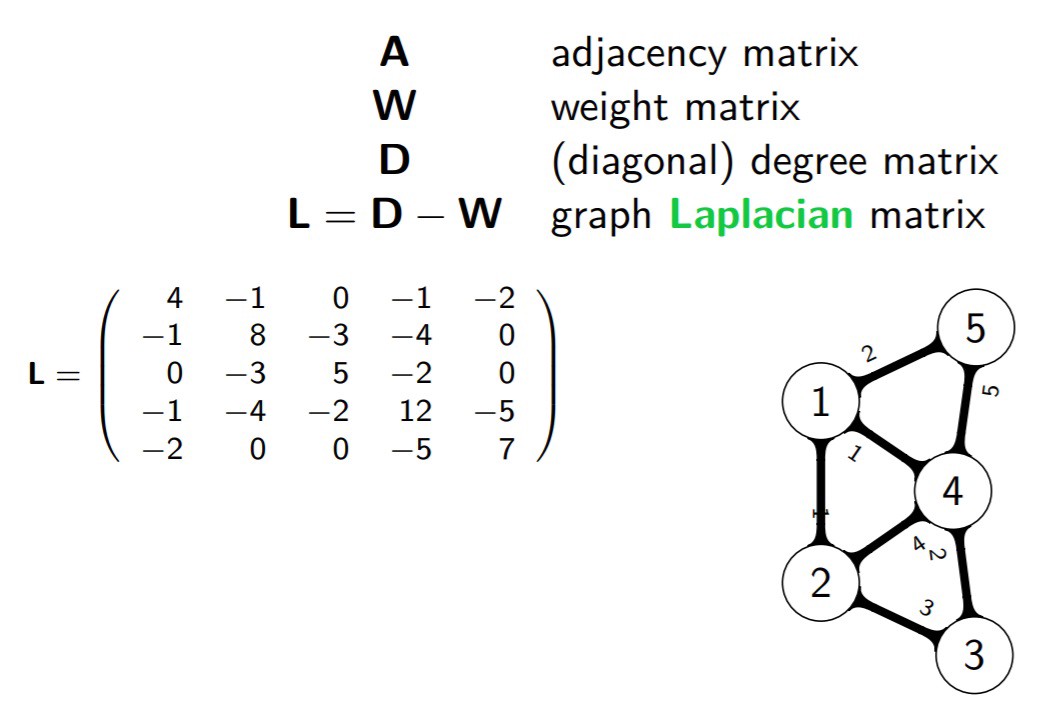
Trong ví dụ trên, chúng tôi đã vẽ 3 cụm: hai “moons” và 1 Gaussian..

Trong đồ thị ε-lân cận, chúng ta có thể thấy rằng rất khó để chọn một tham số hữu ích ε. Với ε = 0,3 như trong hình, các điểm trên mặt trăng giữa đã được kết nối rất chặt chẽ, trong khi các điểm trong Gaussian hầu như không được kết nối với nhau. Vấn đề này luôn xảy ra nếu chúng ta có dữ liệu "ở các tỷ lệ khác nhau", tức là khoảng cách giữa các điểm dữ liệu là khác nhau ở các vùng khác nhau của không gian. Tuy nhiên, đồ thị lân cận k-gần nhất, có thể kết nối các điểm "trên các tỷ lệ khác nhau". Chúng ta có thể thấy rằng các điểm trong Gaussian mật độ thấp được kết nối với các điểm trong moon mật độ cao.

Bước 2 :

Bây giờ chúng ta đã có đồ thị của mình, chúng ta cần hình thành ma trận Laplacian liên quan của nó.

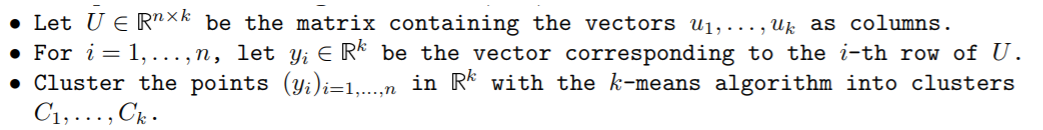
N.B: Các công cụ chính để phân cụm quang phổ là ma trận Laplacian đồ thị.



Tất cả những gì chúng ta phải làm bây giờ là tính toán các eigenvectors u\_ j của L.

Bước 3 :

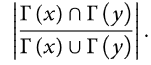
Run k-means :



Cách chọn K :bằng cách chiếu các điểm vào một phép nhúng phi tuyến tính và phân tích các giá trị riêng của ma trận Laplacian, người ta có thể suy ra số lượng các cụm có trong dữ liệu. Khi đồ thị tương tự không được kết nối hoàn toàn, tính đa giá trị riêng λ = 0 cho chúng ta ước lượng của k.

2.4 Phương pháp 3 : Adamic-Adar, Jaccard Coefficient, Preferential Attachment

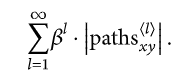
Jaccard’s Coefficient (hệ số Jaccard), còn được gọi là chỉ số Jaccard hoặc hệ số tương tự Jaccard, là một thước đo thống kê được sử dụng để so sánh độ giống nhau của các tập mẫu. Nó thường được ký hiệu là J (x, y) trong đó x và y đại diện cho hai nút khác nhau trong một mạng. Trong dự đoán liên kết, tất cả các lân cận của một nút được coi như một tập hợp và dự đoán được thực hiện bằng cách tính toán và xếp hạng mức độ giống nhau của tập lân cận của mỗi cặp nút. Phương pháp này dựa trên phương pháp Common Neighbors và độ phức tạp của nó cũng vậy O(Nk2). Biểu thức toán học của phương pháp này như sau :



Do giả định rằng nút có mức độ cao có nhiều khả năng nhận được các liên kết mới hơn , phần đính kèm ưu tiên được giới thiệu như một phương pháp dự đoán. Mức độ của cả hai nút trong một cặp cần được xem xét để dự đoán. Giống như các nước láng giềng thông thường, đây cũng là một phương pháp dự đoán cơ bản thường được sử dụng làm cơ sở để đo lường hiệu suất của các phương pháp dự đoán khác. Phương pháp này sẽ tính toán điểm tương tự cho từng cặp nút trong mạng thay vì chỉ các nút lân cận; do đó sự phức tạp của đính kèm ưu đãi là O(N2k2).. Phương pháp này có thể được thể hiện :



Adamic/Adar Index, Ban đầu nó được thiết kế để đo lường mối quan hệ giữa các trang chủ cá nhân. Như coong thức càng có nhiều bạn z, điểm thấp sẽ được gán cho. Do đó, hàng xóm chung của một cặp nút có ít nút lân cận đóng góp nhiều hơn vào giá trị điểm Adamic / Adar (AA) so với giá trị này với số lượng lớn các mối quan hệ. Trong thế giới thực mạng xã hội có thể hiểu như sau: nếu một người quen chung của hai người có nhiều bạn hơn thì khả năng anh ta giới thiệu hai người với nhau sẽ ít hơn trường hợp anh ta chỉ có ít bạn. . Nó cho kết quả tốt trong việc dự đoán tình bạn theo trang chủ cá nhân và Đồ thị cộng tác Wikipedia, nhưng trong thử nghiệm dự đoán sự cộng tác của tác giả, nó cho thấy một dự đoán kém chính xác. Nó là một phương pháp khác dựa trên láng giềng chung; sự phức tạp cũng là O(Nk2). Nó được tính là :

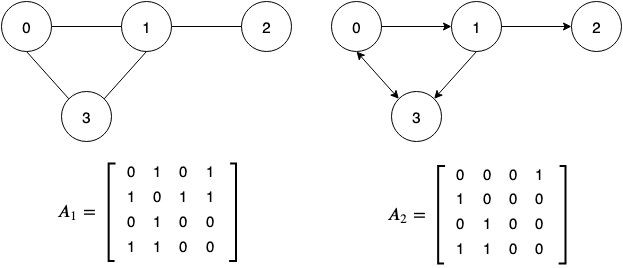


trong đó z là lân cận chung của nút x và nút y.

2.4 Phương pháp 4 - Variational Graph Autoencoders

Adjacency Matrix :

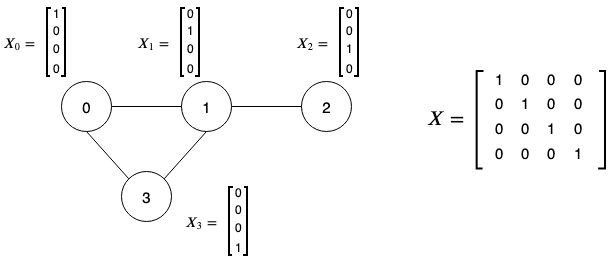
Chúng tôi sử dụng ma trận kề A để biểu diễn đồ thị đầu vào. Thông thường chúng ta giả sử ma trận kề là nhị phân. Giá trị 1 tại hàng i và cột j có nghĩa là có một cạnh giữa đỉnh i và đỉnh j. Giá trị 0 tại hàng m và cột n có nghĩa là không có cạnh giữa đỉnh m và đỉnh n.



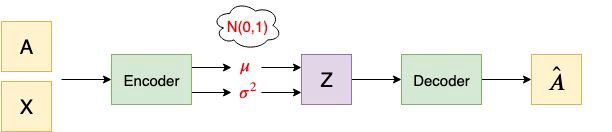
Một ví dụ về ma trận kề khi đồ thị đầu vào có hướng và vô hướng. Ở phía bên trái khi đồ thị bị vô hướng, ma trận kề là đối xứng. Đỉnh 0 và đỉnh 1 được nối với nhau nên A [0,1] = 1 và A [1,0] = 1. Ở phía bên phải khi đồ thị có hướng, ma trận kề là không đối xứng. Đỉnh 0 chỉ đến đỉnh 1, có nghĩa là đỉnh 1 tổng hợp thông tin từ đỉnh 0, do đó A [1,0] = 1.

Feature Matrix

Chúng tôi sử dụng ma trận tính năng X để biểu diễn các tính năng của mỗi nút từ đồ thị đầu vào. Hàng i của ma trận đối tượng X đại diện cho các nhúng đối tượng cho đỉnh i.



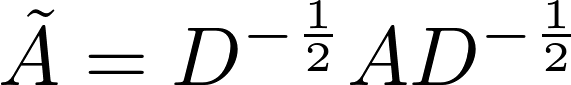
một ví dụ về ma trận tính năng. Mỗi nút i có ma trận đặc trưng riêng X\_i và chúng ta có thể kết hợp chúng với nhau để có được ma trận đặc trưng X cho toàn đồ thị, trong đó mỗi hàng i đại diện cho ma trận đặc trưng của nút i.



Hình 2.4 Kiến trúc của variational graph autoencoder

Bộ mã hóa (mô hình suy luận) của VGAE bao gồm các mạng tích chập đồ thị (GCN). Nó lấy ma trận kề A và ma trận đặc trưng X làm đầu vào và tạo ra biến tiềm ẩn Z làm đầu ra. Lớp GCN đầu tiên tạo ra ma trận tính năng chiều thấp hơn. Nó được định nghĩa là:

Image for post



Dấu ngã là ma trận kề chuẩn hóa đối xứng.

Lớp GCN thứ hai tạo μ và logσ², trong đó

Image for post

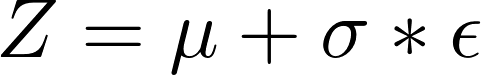
Image for post

Bây giờ nếu chúng ta kết hợp phép toán của GCN hai lớp với nhau, chúng ta nhận được:

Image for post

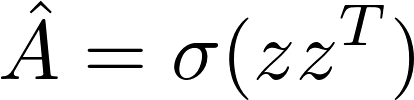
tạo ra μ và logσ².

Sau đó, chúng ta có thể tính Z bằng thủ thuật tham số hóa:



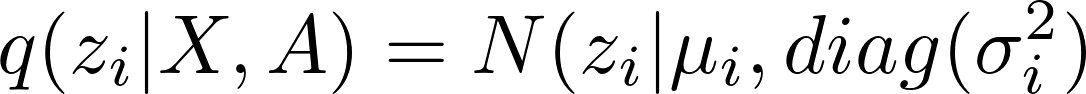
trong đó ε ~ N (0,1).

Bộ giải mã (mô hình tổng quát) được xác định bằng tích bên trong giữa biến tiềm ẩn Z. Đầu ra của bộ giải mã của chúng tôi là ma trận kề A-hat được tái tạo, được định nghĩa là:



trong đó σ (•) là hàm sigmoid logistic.

Tóm lại, bộ mã hóa được biểu diễn dưới dạng:



và bộ giải mã được biểu thị là:

Image for post

CHƯƠNG 3

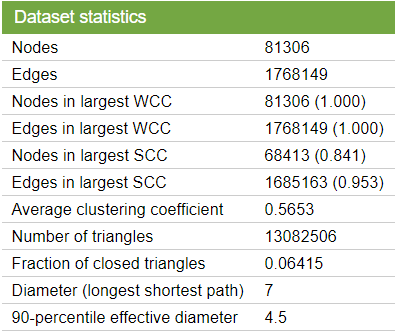
THỰC NGHIỆM

3.1 Dữ liệu

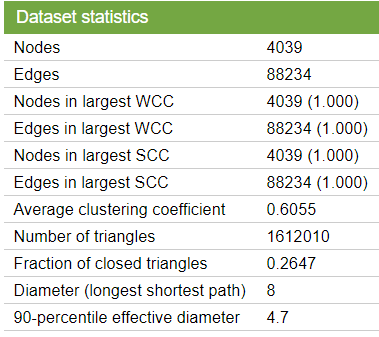
Dữ liệu được lấy từ : <https://snap.stanford.edu/> .

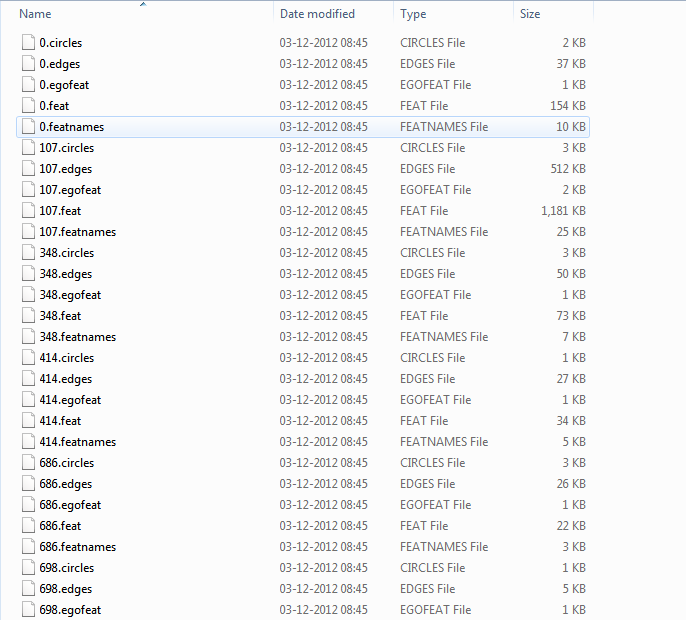
Nguồn : J. McAuley and J. Leskovec. [Learning to Discover Social Circles in Ego Networks](http://i.stanford.edu/~julian/pdfs/nips2012.pdf). NIPS, 2012.

Tập dữ liệu này bao gồm ‘circles’ (hoặc lists) từ Twitter. Dữ liệu Twitter đã được thu thập từ các nguồn công khai. Tập dữ liệu bao gồm các node features (profiles), vòng kết nối và ego network.



Tập dữ liệu này bao gồm circles (hoặc  'friends lists') từ Facebook. Dữ liệu Facebook được thu thập từ những người tham gia khảo sát bằng ứng dụng Facebook này. Tập dữ liệu bao gồm các node features (profiles), vòng kết nối và ego network.





Node.edge chứa thông tin của tất cả các cạnh có trong mạng

**facebook\_combined.txt** chứa thông tin kết hợp về các nút và cạnh

Những gì chúng tôi yêu cầu là các tệp nodeId.feat và nodeId.featnames để thêm thông tin tại mỗi nút. NodeId.featnames chứa tất cả các tính năng / thuộc tính trong một tập hợp các nút cụ thể. và nodeId.feat là một tệp chứa mọi hàng dưới dạng tập hợp các nút và mọi cột là tên đặc điểm của nút được cung cấp trong tệp ‘nodeId.featnames’. Mỗi cột cho một nút cụ thể được ký hiệu bằng ‘0’ hoặc ‘1’, trong đó ‘0’ có nghĩa là nút không có tính năng cụ thể và ‘1’ có nghĩa là nút có tính năng đó.

3.2 Xử lý dữ liệu

Quá trình xử lý dataset Facebook :

* Store all ego graphs in pickle files as (adj, features) tuples
* Read edge-list
* Add ego user (directly connected to all other nodes)
* read features into dataframe
* Add features from dataframe to networkx nodes
* Get adjacency matrix in sparse format (sorted by g.nodes())
* Get features matrix (also sorted by g.nodes())
* Save adj, features in pickle file

Quá trình xử lý dataset Twitter :

* Read combined edge-list
* Parse edgelist into directed graph
* Get adjacency matrix
* Save adjacency matrix
* Generate visualization
* Draw networkx graph -- scale node size by log(degree+1)
* Create black border around node shapes
* Save network stats to .txt file

3.3 Công nghệ sử dụng

Ngôn ngữ lập trình : Python

Các thư viện :

* [TensorFlow](https://www.tensorflow.org/install/) (1.0 or later)
* [networkx](https://networkx.github.io/)
* [gensim](https://radimrehurek.com/gensim/install.html)
* [scikit-learn](http://scikit-learn.org/stable/)
* [scipy](https://www.scipy.org/_)
* [pandas](https://pandas.pydata.org/)

Các công cụ **:**

* [jupyter notebook](http://jupyter.org/install.html)

3.4 Cách đánh giá

**ROC curve** : Đường cong ROC (đường đặc tính hoạt động của máy thu) là đồ thị thể hiện hiệu suất của một mô hình phân loại ở tất cả các ngưỡng phân loại. Đường cong này vẽ hai tham số:

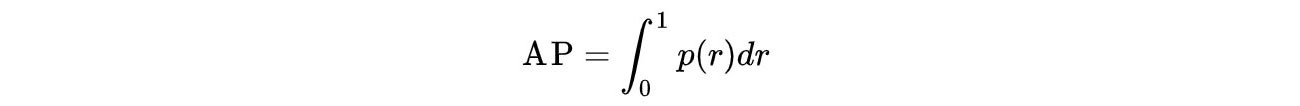
True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

False Positive Rate (FPR) is defined as follows:

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

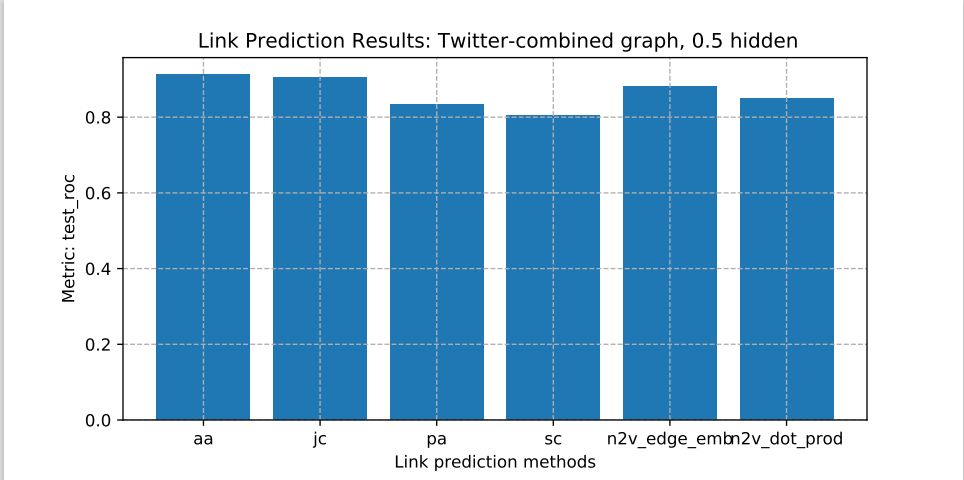
**AUC** đo toàn bộ diện tích hai chiều bên dưới toàn bộ đường cong ROC (tính tích phân) từ (0,0) đến (1,1).

**Average precision** is a measure that combines recall and precision for ranked retrieval results. For one information need, the average precision is the mean of the precision scores after each relevant document is retrieved.

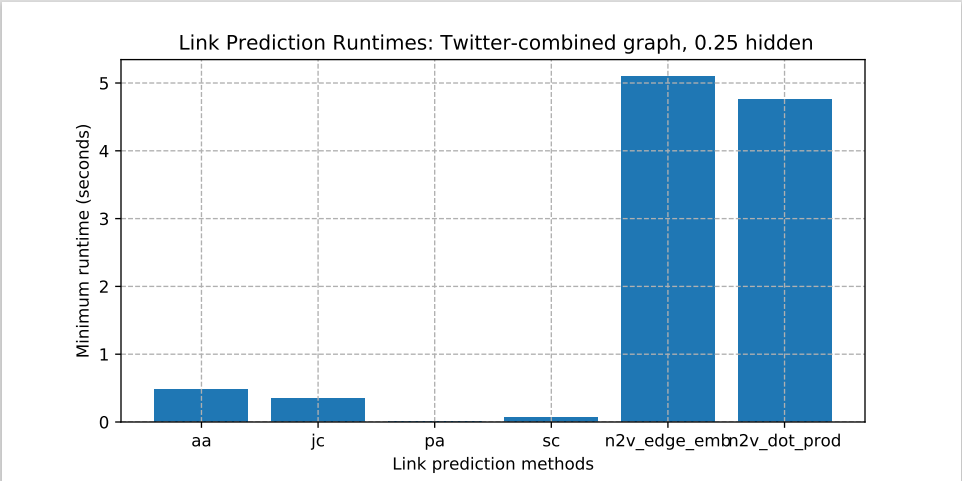


**Run time :** Thời gian thực thi của các giải thuật

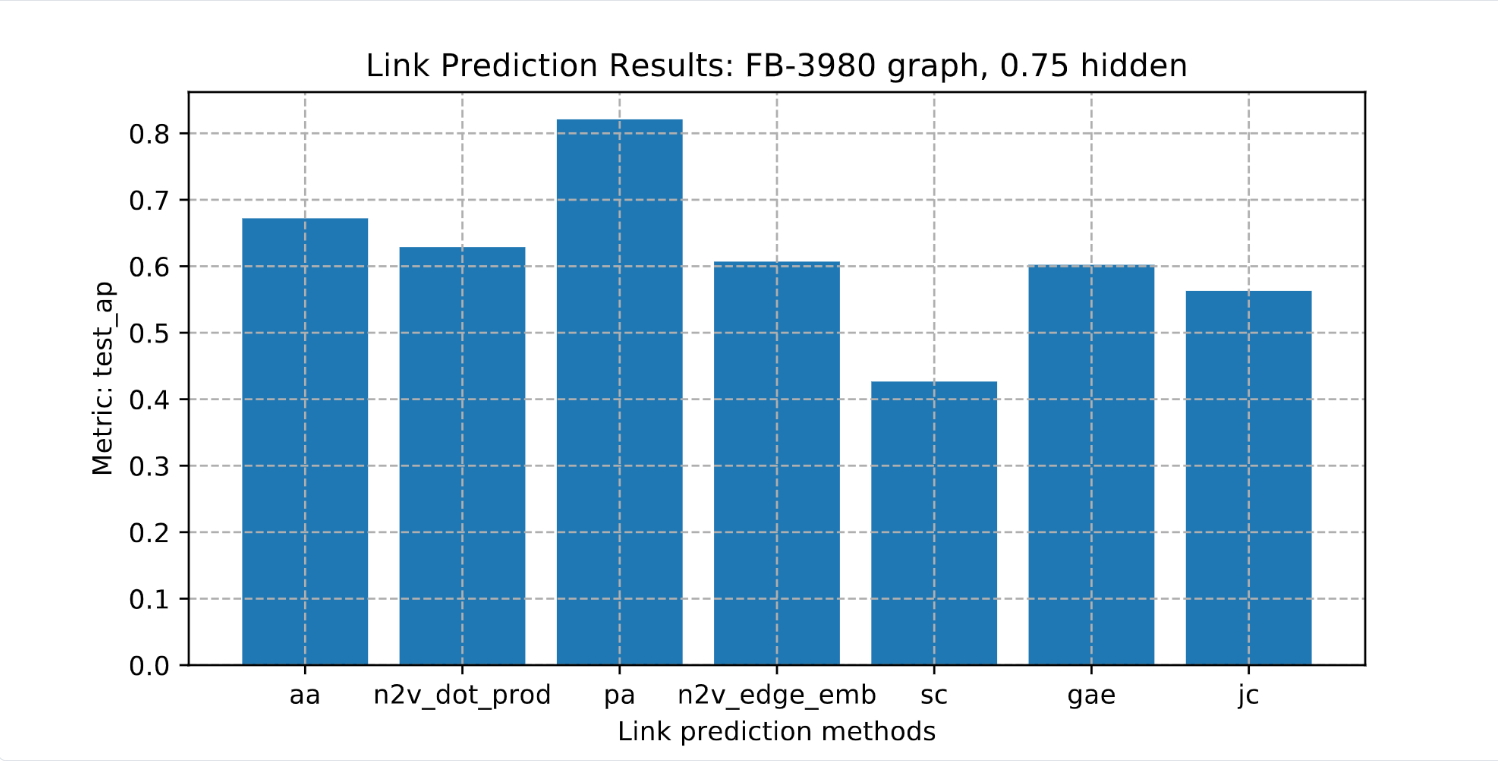
3.5 Kết quả đạt được



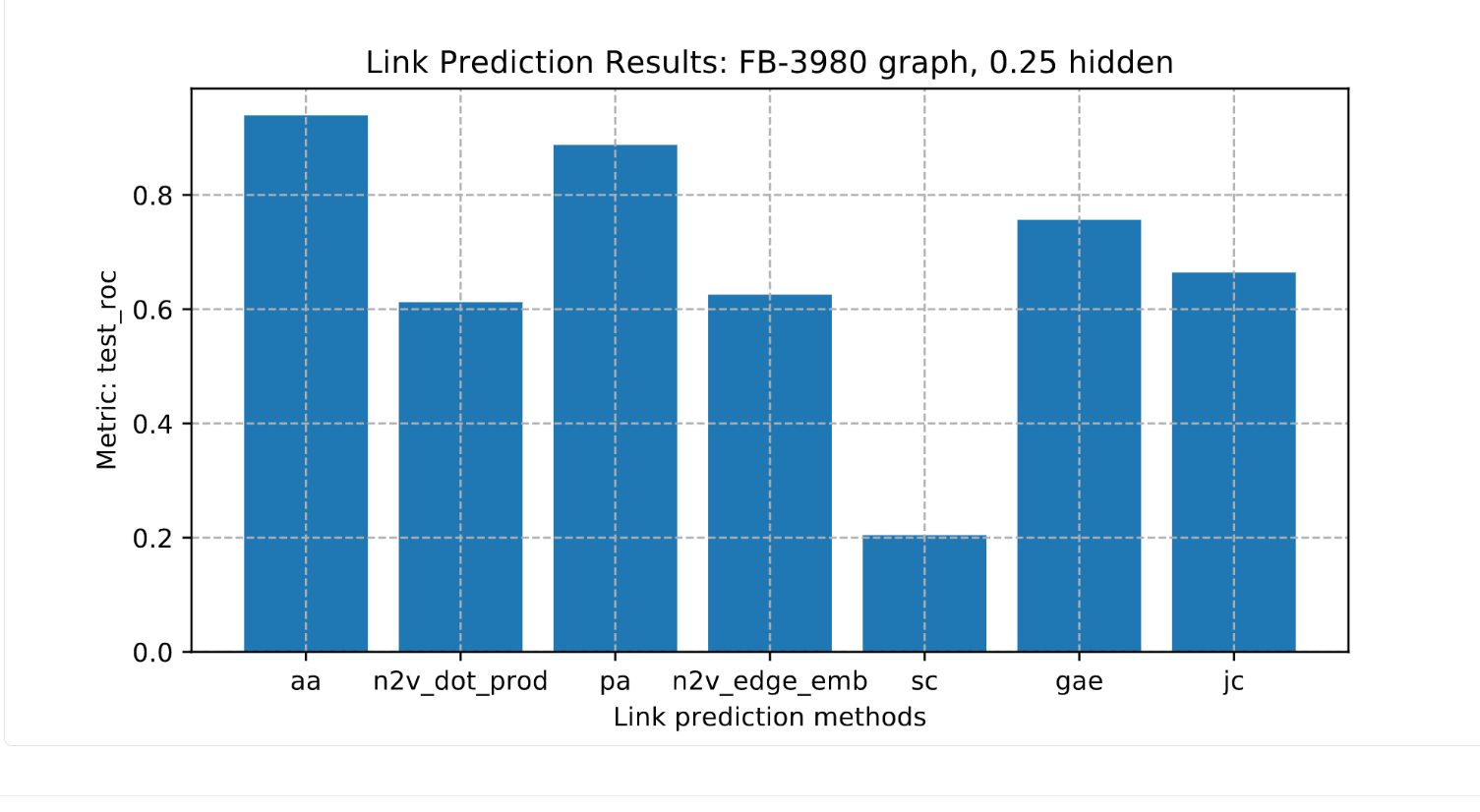
Hình 3.5 biểu đồ so sánh giữa các phương pháp theo độ đo Roc trên tập dữ liệu twitter



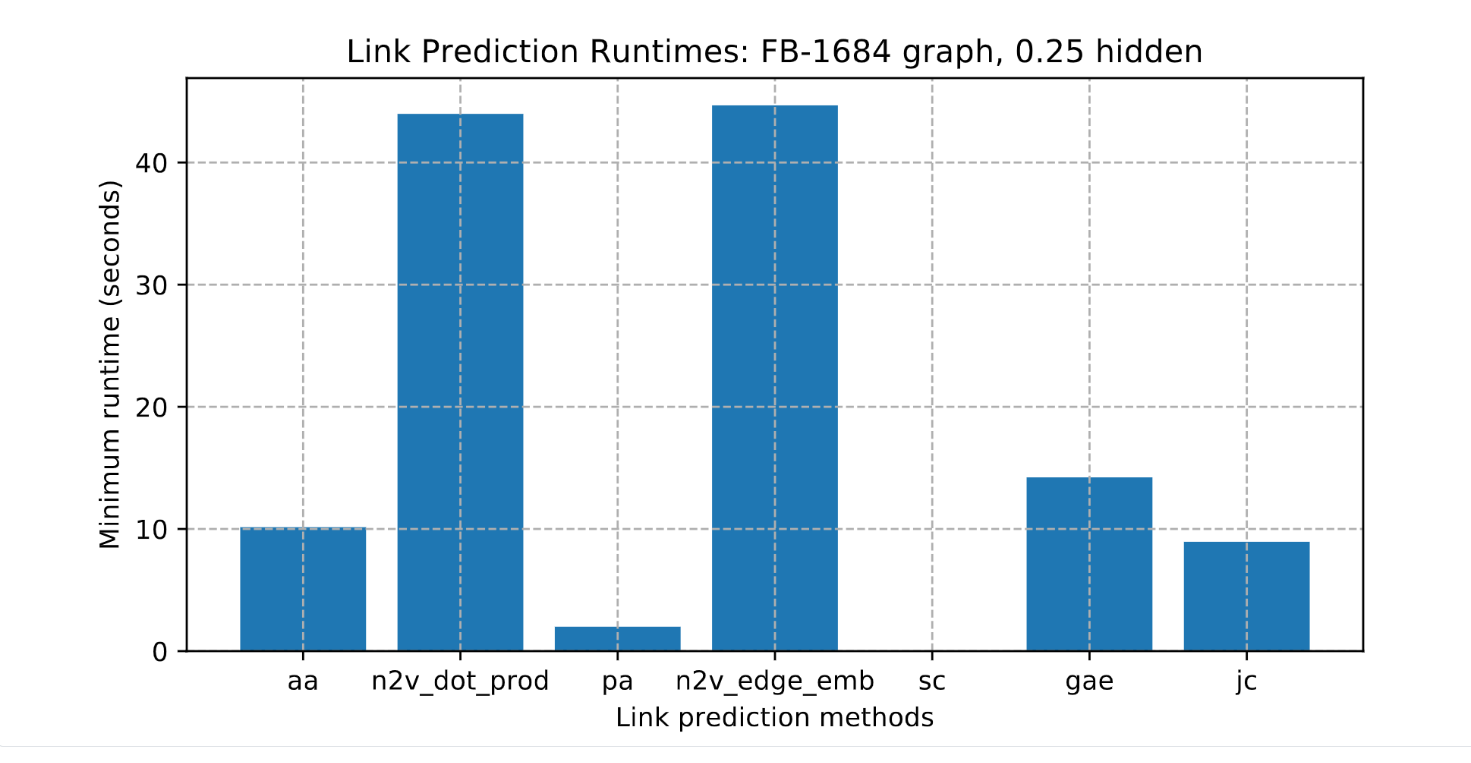
Hình 3.6 biểu đồ so sánh giữa các phương pháp theo độ đo runtime trên tập dữ liệu twitter



Hình 3.7 biểu đồ so sánh giữa các phương pháp theo độ đo Average precision trên tập dữ liệu facebook



Hình 3.8 biểu đồ so sánh giữa các phương pháp theo độ đo roc trên tập dữ liệu facebook



Hình 3.9 biểu đồ so sánh giữa các phương pháp theo độ đo runtime trên tập dữ liệu facebook

CHƯƠNG 4

KẾT LUẬN

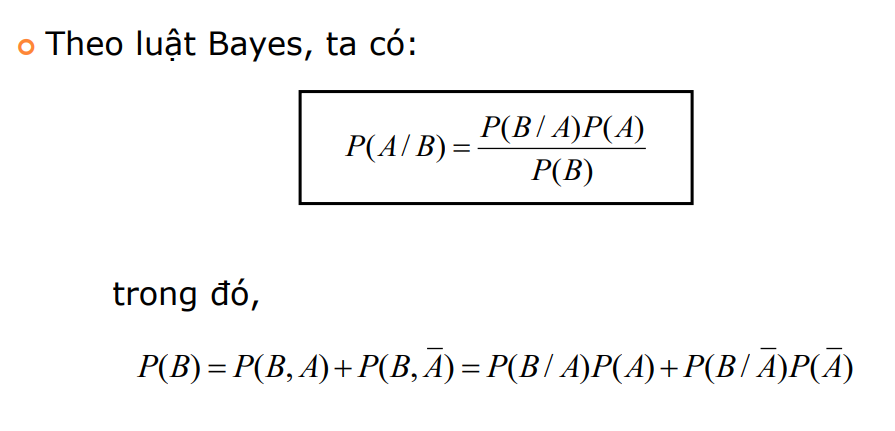
4.1 Kết luận

Sau khi hoàn thành đồ án :

4.2 Hướng phát triển

PHẦN 2

Bài toán 1:



Gọi G là khách hang có tín dụng tốt.

Gọi OD là khách hàng rút quá số tiền gửi.

P(OD|G) =0.01 => P(OD~|G) =0.99 . P(OD|G~) = 0.24 => P(OD~|G~) = 0.76 . P(G) = 0.84 => P(G~) = 0.16.

1. Trong tháng đầu tiên, khách hàng này rút quá số tiền gửi. Hỏi ngân hàng sẽ xác định lại khả năng để khách hàng trở này trở thành tín dụng tốt là bao nhiêu?

P(G|OD) = (P(OD|G)\*P(G))/P(OD)

= (P(OD|G)\*P(G))/( P(OD|G)\*P(G) + P(OD|G~)\*P(G~))

=(0.01\*0.84) / (0.01\*0.84 + 0.24 \*0.16).

=0.18.

2. Đến cuối tháng thứ 2, nếu khách hàng này không rút quá số tiền gửi thì ngân hàng sẽ xác định lại khả năng khách hàng này trở thành tín dụng tốt là bao nhiêu? P(G) = 0.18 = >P(G~) = 0.82

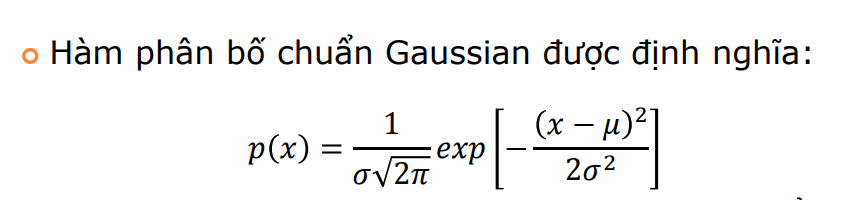
P(G|OD~) = (P(OD~|G)\*P(G))/P(O~)

= (P(OD~|G)\*P(G))/( P(OD~|G)\*P(G) + P(OD~|G~)\*P(G~))

=(0.99\*0.18)/(0.99\*0.18+0.76\*0.82)

=0.22.

Bài toán 2:



**2.1. Xây dựng một phân lớp để chẩn đoán (với xác suất lỗi nhỏ nhất) nếu biết rằng trẻ được xét nghiệm mất thính giác.**

P(x=|class = suy giảm) = = , 0

**2.2.Xét nghiệm cho 1 trẻ được kết quả là 2.4? Trường hợp này cần chẩn đoán thế nào.**

-Xác xuất xét nghiệm cho 1 trẻ với trường hợp suy giảm thính lực.

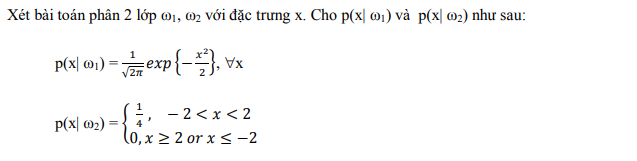
P(x=2.4|class = suy giảm) = = 0.19

-Xác xuất xét nghiệm cho 1 trẻ với trường hợp bình thường.

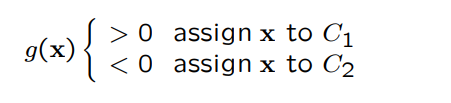
P(x=2.4|class = bình thường) = =0.097

Đưa vào xác xuất ở trên ta thấy P(x=2.4|class = bình thường) < P(x=2.4|class = suy giảm) => Chẩn đoán suy giảm.

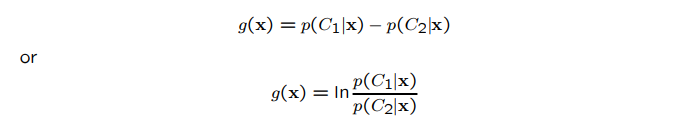
Bài toán 3:

Ta có π1 + π2 = 1 , f1(x) π1 = f2(x) π2

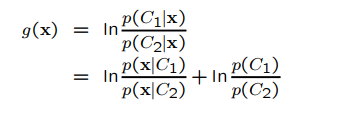
* Discriminant function.



* Đối với trường hợp P(C1) = P(C2)



* Đối với trường hợp P(C1)≠P(C2)



**3.1 Tìm luật phân phối thỏa cực tiểu hóa xác suất lỗi biết rằng P(ω1) = P(ω1) = 0.24.**

- Trường hợp -2 < x < 2 . vì P (ω1) = P (ω2) = 0,5, chúng ta có hàm phân biệt g (x) là.

=Ln

Quy tắc Bayes để phân loại sẽ là :

* Quyết định lớp nếu g(x)>0 ngược lại quyết định lớp .

Hoặc

- Quyết định lớp nếu -0.9668 < x < 0.9668 ngược lại quyết định lớp .

- Trong trường hợp x ≥ 2 hoặc x ≤ −2, ta luôn xác định ω1.

**3.2. Tìm π1 sao cho nếu P(ω1) > π1 thì phân lớp ở câu (1) luôn quyết định ω1 bất chấp x.**

Theo đề bài ta có phương trình sau.

p(x|ω1)\*π1 = p(x|ω2)\*(1 − π1 ) khi x = 2 hoặc x = -2.

⬄ \* π1 = \* (1 - π1)

⬄ π1 = 0.822.

3.3. Có tồn tại π2 sao cho nếu P(ω2) > π2 thì phân lớp ở câu (1) luôn quyết định ω2.

-Không . Vì p(x|ω2) chỉ được xác định trên khoảng -2 < x < 2, Nhưng theo phần 3.1 chúng ta đã luôn quyết định ω1 trên khoảng -2 < x < 2.

TÀI LIỆU THAM KHẢO

**Tiếng Việt**

**Tiếng Anh**

[1][Image Segmentation by Using Linear Spectral Clustering, Sandeep Reddy](https://www.omicsonline.org/open-access/image-segmentation-by-using-linear-spectral-clustering-2167-0919-1000143.php?aid=81482) [Graphs in Machine Learning, Michal Valko](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.708.6592&rep=rep1&type=pdf)

[2] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Makenew friends, but keep the old: recommending people on socialnetworking sites,” in Proceedings of the 27th international conferenceon Human factors in computing systems, ser. CHI ’09. NewYork, NY, USA: ACM, 2009, pp. 201–210. [Online]. Available:http://doi.acm.org/10.1145/1518701.1518735

[3] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixedmembership stochastic block models for relational data with applicationto protein-protein interactions,” Proceedings of Ineterational BiometricSociety-ENAR Annual Meetings, 2006.

[4] Z. Huang, X. Li, and H. Chen, “Link prediction approach to collaborativeﬁltering,” Proceedings of the 5th ACM/IEEE-CS joint conference onDigital libraries, 2005.

[5] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction usingsupervised learning,” SDM Workshop of Link Analysis, Counterterrorismand Security, 2006.[5] M. A. Hasan and M. J. Zaki, Social Network Data Analytics, C. C.Aggarwal, Ed. Springer, 2011.

[6] D. Liben-Nowell and J. Kleinber, “The link-prediction problem for socialnetworks,” Journal of the American Society for Information Science andTechnology, vol. 58, no. 7, 2007

“Facebook,” http://www.facebook.com/.

[7] “Flickr,” http://www.ﬂickr.com/.

[8] “Youtube,” <http://www.youtube.com/>.

[9] “Academia.edu,” <http://academia.edu/>.

[10] “Themarker-cafe,” <http://cafe.themarker>.

[11] https://zenodo.org/record/1408472#.X8eiIs0zZPY

J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Make

new friends, but keep the old: recommending people on social

networking sites,” in Proceedings of the 27th international conference

on Human factors in computing systems, ser. CHI ’09. New

York, NY, USA: ACM, 2009, pp. 201–210. [Online]. Available:

http://doi.acm.org/10.1145/1518701.1518735

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed

membership stochastic block models for relational data with application

to protein-protein interactions,” Proceedings of Ineterational Biometric

Society-ENAR Annual Meetings, 2006.

[3] Z. Huang, X. Li, and H. Chen, “Link prediction approach to collaborative

ﬁltering,” Proceedings of the 5th ACM/IEEE-CS joint conference on

Digital libraries, 2005.

[4] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using

supervised learning,” SDM Workshop of Link Analysis, Counterterrorism

and Security, 2006.

[5] M. A. Hasan and M. J. Zaki, Social Network Data Analytics, C. C.

Aggarwal, Ed. Springer, 2011.

[6] D. Liben-Nowell and J. Kleinber, “The link-prediction problem for social

networks,” Journal of the American Society for Information Science and

Technology, vol. 58, no. 7, 2007

J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Make

new friends, but keep the old: recommending people on social

networking sites,” in Proceedings of the 27th international conference

on Human factors in computing systems, ser. CHI ’09. New

York, NY, USA: ACM, 2009, pp. 201–210. [Online]. Available:

http://doi.acm.org/10.1145/1518701.1518735

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed

membership stochastic block models for relational data with application

to protein-protein interactions,” Proceedings of Ineterational Biometric

Society-ENAR Annual Meetings, 2006.

[3] Z. Huang, X. Li, and H. Chen, “Link prediction approach to collaborative

ﬁltering,” Proceedings of the 5th ACM/IEEE-CS joint conference on

Digital libraries, 2005.

[4] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using

supervised learning,” SDM Workshop of Link Analysis, Counterterrorism

and Security, 2006.

[5] M. A. Hasan and M. J. Zaki, Social Network Data Analytics, C. C.

Aggarwal, Ed. Springer, 2011.

[6] D. Liben-Nowell and J. Kleinber, “The link-prediction problem for social

networks,” Journal of the American Society for Information Science and

Technology, vol. 58, no. 7, 2007

J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, “Make

new friends, but keep the old: recommending people on social

networking sites,” in Proceedings of the 27th international conference

on Human factors in computing systems, ser. CHI ’09. New

York, NY, USA: ACM, 2009, pp. 201–210. [Online]. Available:

http://doi.acm.org/10.1145/1518701.1518735

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed

membership stochastic block models for relational data with application

to protein-protein interactions,” Proceedings of Ineterational Biometric

Society-ENAR Annual Meetings, 2006.

[3] Z. Huang, X. Li, and H. Chen, “Link prediction approach to collaborative

ﬁltering,” Proceedings of the 5th ACM/IEEE-CS joint conference on

Digital libraries, 2005.

[4] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using

supervised learning,” SDM Workshop of Link Analysis, Counterterrorism

and Security, 2006.

[5] M. A. Hasan and M. J. Zaki, Social Network Data Analytics, C. C.

Aggarwal, Ed. Springer, 2011.

[6] D. Liben-Nowell and J. Kleinber, “The link-prediction problem for social

networks,” Journal of the American Society for Information Science and

Technology, vol. 58, no. 7, 2007

PHỤ LỤC

TỰ CHẤM ĐIỂM

|  |  |  |  |
| --- | --- | --- | --- |
| Phần I | | | |
| Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| Chương 1 | 0.5đ | **0.5đ** |  |
| Chương 2 | 1.5đ | **1.0đ** |  |
| Chương 3 | 3.75đ | **2.0đ** |  |
| Chương 4 | 0.25đ | **0đ** | Chưa nêu được hướng phát triển trong tương lai |
| Tổng phần I | |  |  |
| Phần II | | | |
| Bài 1 | 1đ | **1đ** |  |
| Bài 2 | 1đ | **1đ** |  |
| Bài 3 | 1đ | **1đ** |  |
| Tổng phần II | |  |  |
| Điểm nhóm | | | |
| Điểm nhóm | 1đ | **1đ** |  |
| Tổng điểm | | **7.5đ** |  |