**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**XÂY DỰNG ỨNG DỤNG PHÂN TÍCH CÁC ĐỘ DO, KHAI PHÁ DỮ LIỆU CỦA MỘT MẠNG XÃ HỘI (FACEBOOK NETWORK)**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | | |
| STT | Họ tên | MSSV |
| 1 | Huỳnh Văn Pháp | 19521987 |
| 2 | Ngô Bảo Thiên | 19522260 |

**TP. HỒ CHÍ MINH – 5/2022**

**MỤC LỤC**

[1. GIỚI THIỆU 1](#_Toc103773152)

[2. NỘI DUNG 1](#_Toc103773153)

[2.1. Quy trình áp dụng xây dựng hệ thống 1](#_Toc103773154)

[2.1.1. Mô tả dataset 1](#_Toc103773155)

[2.1.2. Thư viện và công cụ sử dụng 1](#_Toc103773156)

[2.1.3. Phân tích thiết kế hệ thống 2](#_Toc103773157)

[2.2. Quy trình thực hiện 4](#_Toc103773158)

[2.3. Đánh giá kết quả 4](#_Toc103773159)

[2.3.1. Import và xử lí dữ liệu đầu vào 5](#_Toc103773160)

[2.3.2. Cụm chức năng phân tích các độ đo 5](#_Toc103773161)

[2.3.3. Cụm chức năng phân tích bằng các thuật toán khám phá cộng đồng 6](#_Toc103773162)

[2.3.4. Cụm chức năng dự đoán liên kết trong mạng xã hội 7](#_Toc103773163)

[3. KẾT LUẬN 8](#_Toc103773164)

[3.1. Ưu điểm 8](#_Toc103773165)

[3.2. Khuyết điểm 8](#_Toc103773166)

[PHỤ LỤC PHÂN CÔNG NHIỆM VỤ 9](#_Toc103773167)

# GIỚI THIỆU

Đề tài của đồ án là xây dựng ứng dụng phân tích mạng xã hội là quá trình điều tra các cấu trúc của mạng xã hội thông qua danh sách mạng lưới liên kết (bạn bè) nhằm mục đích phân tích vai trò, phân loại các nút trong mạng xã hội, phát hiện cộng đồng trong mạng xã hội và dự đoán các liêt kết mới của các nút trong mạng xã hội

Việc thực hiện đề tài gồm các quá trình sau : Tìm bộ dữ liệu để phân tích → Lên danh sách các chức năng cần thực hiện→Tìm hiểu các thư viện cần sử dụng để thực hiện đề tài → Phân tích thiết kế hệ thống qua : Class Diagram và Use Case Diagram bằng StarUML →Tiến hành lập trình và hoàn thiện ứng dụng.

* Kết quả đã đạt được:

+ Hoàn thành được các chức năng đã đề ra : phân tích các độ do, khai phá, dự đoán liên kết.

+ Có thể sử dụng ứng dụng để phân tích các bộ dữ liệu của nhiều mạng xã hội khác nếu có dataset mạng lưới bạn bè của mạng xã hội đó.

Trong báo cáo này, nhóm tập trung trình bày ba nội dung chính: (1) Quy trình áp dụng xây dựng hệ thống, (2) Quy trình thực hiện, (3) Đánh giá kết quả.

# NỘI DUNG

## Quy trình áp dụng xây dựng hệ thống

### Mô tả dataset

Tập dữ liệu này bao gồm danh sách kết nối (‘Bạn bè’) giữa những người với nhau từ Facebook. Dữ liệu được thu thập từ những người tham gia khảo sát sử dụng ứng dụng Facebook bởi Stanford University (<https://snap.stanford.edu/>) trích dẫn từ : J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

Tập dữ liệu này có 2 cột không có tiêu đề : Cột 1 (Node thứ nhât), Cột 2 (Node thứ hai có liên kết là bạn bè với Node thứ nhất)

Link: <https://snap.stanford.edu/data/facebook_combined.txt.gz>, trong quá trình thực hiện đề tài nhóm đã đổi kiểu dữ liệu file từ \*txt sang \*csv để dễ sử dụng.

### Thư viện và công cụ sử dụng

* Tkinter : là thư viện GUI tiêu chuẩn cho python.
* pandas : là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu
* networkx : là một module của python, hỗ trợ khả năng vẽ đồ thị, thao tác với dữ liệu, và đọc đồ thị đa cấp.
* matplotlib : là một thư viện sử dụng để vẽ các đồ thị trong Python.
* numpy : là thư viện lõi phục vụ cho khoa học máy tính của Python.
* pyvis : là một thư viện Python cho phép tạo đồ thị mạng tương tác trong một vài dòng mã.
* StarUML : là công cụ để phân tích thiết kế hệ thống theo OOP

### Phân tích thiết kế hệ thống

Use Case Diagram:

Diagram

Description automatically generated

*Hình 1: Sơ đồ hệ thống (Use Case Diagram)*

* Mô tả : Người dùng muốn sử dụng các chức năng của ứng dụng đầu tiên phải import vào file dữ liệu sau đó có thể chọn các nhóm chức năng để sử dụng bao gồm :

+ Phân tích các độ đo : Degree Centrality(In, Out), Betweenness Centrality, Closeness Centrality, Clustering Coefficient.

+ Phân tích bằng các thuật toán khám phá cộng đồng: Girvan NewMan, Page Rank

+ Dự đoán liên kết trong mạng xã hội: CommonNeighbors, Adamic/Adar, Katz, Preferential Attachment

Class Diagram :

A picture containing text

Description automatically generated

*Hình 2: Sơ đồ lớp (Class Diagram)*

* Mô tả :

+ pre\_data (class) : thực hiện việc tiền xử lí dữ liệu (tạo graph, node …).

+ d\_network (class) : thực hiện các hàm trực quan hóa đồ thị.

+ output\_data (class) : thực hiện việc xử lí và trả về dữ liệu.

+ backend (class) : điều hướng các sự kiện từ ui (class).

+ dataview (class) : trực quan hóa các kiểu hiển thị dữ liệu.

+ view\_discovery\_algorithm (class) : trực quan hóa thuật toán girvan newman

+ view (class) : tạo đối tượng để chạy chương trình

## Quy trình thực hiện

Diagram

Description automatically generated

*Hình 3: Quy trình thực hiện đề tài*

## Đánh giá kết quả

Graphical user interface, application

Description automatically generated

*Hình 4: Ứng dụng sau khi đã hoàn thành*

### Import và xử lí dữ liệu đầu vào

Đã hoàn thành được việc import và xử lí file \*csv đầu vào tạo graph, các biến dữ liệu để lưu kết quả để thực hiện các công việc tính toán.

Xử lí thêm được việc import file từ máy tính mà không cần cố dịnh đường dẫn file gốc.

Graphical user interface, application

Description automatically generated

*Hình 5: Chọn dữ liệu đầu vào từ máy tính*

### Cụm chức năng phân tích các độ đo

Graphical user interface, text, application

Description automatically generated

*Hình 6: Cụm chức năng phân tích các độ đo*

Các chức năng :

* Chọn độ đo : Chọn độ đo để phân tích.
* Trực quan hóa mạng liên kết : Hiển thị đồ thị theo giá trị độ do đã chọn.
* Độ đo : Tính toán giá trị độ đo theo độ đo đã chọn và hiển thi ra danh sách theo số lượng node đã chọn.
* Tần số và xác suất xuất hiện : Hiển thị danh sách các giá trị độ đo đã chọn của đồ thị và tần số, xác suất xuất hiện của giá trị đó.

### Cụm chức năng phân tích bằng các thuật toán khám phá cộng đồng

**Thuật toán Page Rank:**

Graphical user interface, application

Description automatically generated

*Hình 7: Cụm chức năng thuật toán Page Rank*

* Các chức năng :

+ Trực quan hóa mạng liên kết : Hiển thị đồ thị theo giá trị Page Rank

+ Độ do : Tính toán giá trị Page Rank và hiển thi ra danh sách theo số lượng node đã chọn.

+ Tần số và xác suất xuất hiện : Hiển thị danh sách các giá trị Page Rank của đồ thị và tần số, xác suất xuất hiện của giá trị đó.

**Thuật toán Girvan NewMan:**

Graphical user interface, text, application, email

Description automatically generated

*Hình 8: Cụm chức năng thuật toán Girvan Newman*

* Các chức năng:

+ Khai phá cộng đồng : Tìm ra những cộng đồng có liên kết chặt chẽ.

+ Trực quan hóa mạng liên kết : Hiển thị đồ thị theo các cộng đồng đã tách được.

### Cụm chức năng dự đoán liên kết trong mạng xã hội

Graphical user interface, application

Description automatically generated

*Hình 9: Cụm chức năng dự đoán liên kết trong mạng xã hội*

* Các chức năng :

+ Chọn thuật toán: Chọn loại thuật toán dùng để phân tích.

+ Phân tích : Phân tích và đưa ra thứ tự các node có thể xuất hiện liên kết dựa theo thuật toán đã chọn.

# KẾT LUẬN

## Ưu điểm

Thông qua đề tài này, nhóm đã tạo được một ứng dụng phân tích dữ liệu. Đồng thời biết cách sử dụng các kỹ thuật lập trình python trong một dự án.

Về mặt kết quả sau khi hoàn thành ứng dụng:

* Ứng dụng đã phân tích và tính toán chính xác các kết quả sau khi đã thử nghiệm.
* Sử dụng để phân tích được nhiều dữ liệu khác nhau chứ không mỗi riêng dataset ban đầu.
* Hoàn thành chức năng phân tích các độ đo : Degree Centrality(In, Out), Betweenness Centrality, Closeness Centrality, Clustering Coefficient.
* Hoàn thành chức năng khai phá cộng đồng bằng các thuật toán: Page Rank, Girvan NewMan.
* Hoàn thành chức năng dự đoán liên kết bằng các thuật toán : Common Neighbors, Adamic/Adar, Katz, Preferential Attachment.
* Trong tương lai có thể hoàn thiện và thêm nhiều chức năng phân tích khác.

## Khuyết điểm

Giao diện ứng dụng còn sơ sài chưa được bắt mắt, chưa responsive trên mọi máy tính khi sử dụng ứng dụng. Chỉ phân tích được với những bộ dữ liệu thích hợp.

**TÀI LIỆU THAM KHẢO**

[Modelling Influence in a Social Network: Metrics and Evaluation](https://scholar.google.com.vn/scholar_url?url=https://pdfs.semanticscholar.org/0372/cdead2a09c4a2176e3b07be0d3fbc3e80ad7.pdf&hl=vi&sa=X&ei=gZaEYpL7HKuO6rQP296DwAY&scisig=AAGBfm1V6BB0ASAu_NeSDY7ZnzBSBU4lVw&oi=scholarr); Tác giả : Behnam Hajian, Tony White; July 2011

[Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network](https://scholar.google.com.vn/scholar_url?url=https://ojs.aaai.org/index.php/SOCS/article/download/18200/17991&hl=vi&sa=X&ei=xZaEYsbQOOeEywSBrZ7ICQ&scisig=AAGBfm0gOqR1SPwwyKF8NE-MZj8T42JZVg&oi=scholarr); Tác giả : Junlong Zhang, Yu Luo; 2017

[Defense resource allocation in road dangerous goods transportation network: A SelfContained Girvan-Newman Algorithm and Mean Variance Model combined approach](https://www.researchgate.net/profile/Wencheng-Huang/publication/353109256_Defense_resource_allocation_in_road_dangerous_goods_transportation_network_A_Self-_Contained_Girvan-Newman_Algorithm_and_Mean_Variance_Model_combined_approach/links/60e7a2a9b8c0d5588ce2f25d/Defense-resource-allocation-in-road-dangerous-goods-transportation-network-A-Self-Contained-Girvan-Newman-Algorithm-and-Mean-Variance-Model-combined-approach.pdf), Tác giả : Wencheng Huang, Linqing Li, Hongyi Liua, Rui Zhanga, Minhao Xua; 2016

[The Algorithm of Link Prediction on Social Network](https://www.hindawi.com/journals/mpe/2013/125123/), Tác giả : Liyan Dong; 2013

**PHỤ LỤC**

Hình 1: Sơ đồ hệ thống (Use Case Diagram)

Hình 2: Sơ đồ lớp (Class Diagram)

Hình 3: Quy trình thực hiện đề tài

Hình 4: Ứng dụng sau khi đã hoàn thành

Hình 5: Chọn dữ liệu đầu vào từ máy tính

Hình 6: Cụm chức năng phân tích các độ đo

Hình 7: Cụm chức năng thuật toán Page Rank

Hình 8: Cụm chức năng thuật toán Girvan Newman

Hình 9: Cụm chức năng dự đoán liên kết trong mạng xã hội

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Huỳnh Văn Pháp | Lựa chọn đề tài, phân công nhiệm vụ, tìm nguồn dữ liệu, lên danh sách các chức năng, vẽ sơ đồ phân tích thiết kế, lập trình sự kiện cho các tương tác, viết báo cáo. |
| 2 | Ngô Bảo Thiên | Thiết kế và lập trình giao diện, hỗ trợ code sự kiện cho các tương tác, làm slide thuyết trình. |