

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**XÂY DỰNG ỨNG DỤNG PHÂN TÍCH CÁC ĐỘ  
ĐO, 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

<b>1. GIỚI THIỆU .....</b>	<b>1</b>
<b>2. NỘI DUNG .....</b>	<b>1</b>
2.1. Quy trình áp dụng xây dựng hệ thống .....	1
2.1.1. Mô tả dataset.....	1
2.1.2. Thư viện và công cụ sử dụng .....	1
2.1.3. Phân tích thiết kế hệ thống .....	2
2.2. Quy trình thực hiện.....	4
2.3. Đánh giá kết quả.....	4
2.3.1. Import và xử lý dữ liệu đầu vào .....	5
2.3.2. Cụm chức năng phân tích các độ đo.....	5
2.3.3. Cụm chức năng phân tích bằng các thuật toán khám phá cộng đồng .....	6
2.3.4. Cụm chức năng dự đoán liên kết trong mạng xã hội.....	7
<b>3. KẾT LUẬN .....</b>	<b>8</b>
3.1. Ưu điểm.....	8
3.2. Khuyết điểm .....	8
<b>PHỤ LỤC PHÂN CÔNG NHIỆM VỤ .....</b>	<b>9</b>

## 1. 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 liên 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 đo, 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.

## 2. NỘI DUNG

### 2.1. Quy trình áp dụng xây dựng hệ thống

#### 2.1.1. 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 nhất), 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](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.

#### 2.1.2. Thư viện và công cụ sử dụng

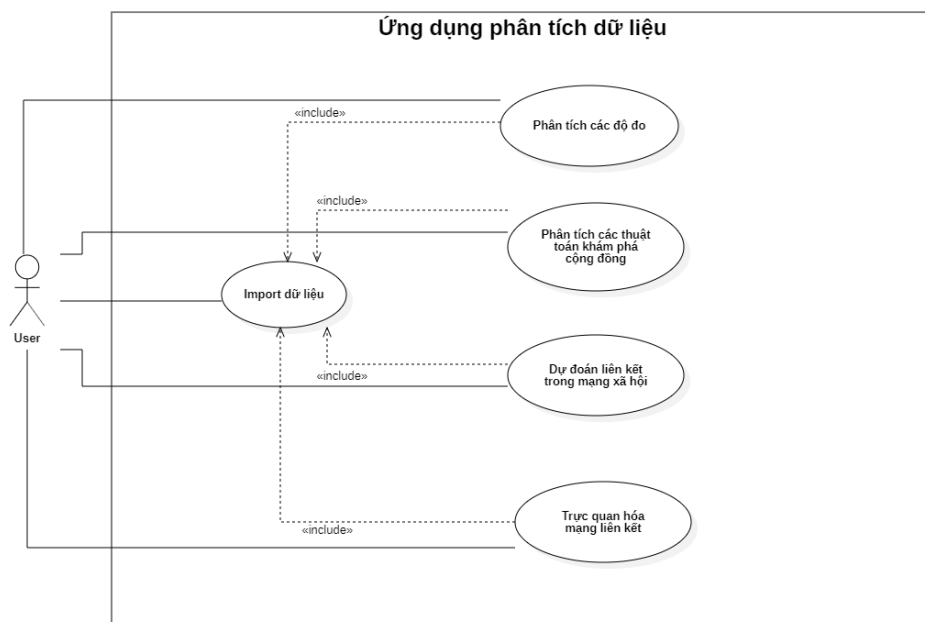
– Tkinter : là thư viện GUI tiêu chuẩn cho python.

– 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

### 2.1.3. Phân tích thiết kế hệ thống

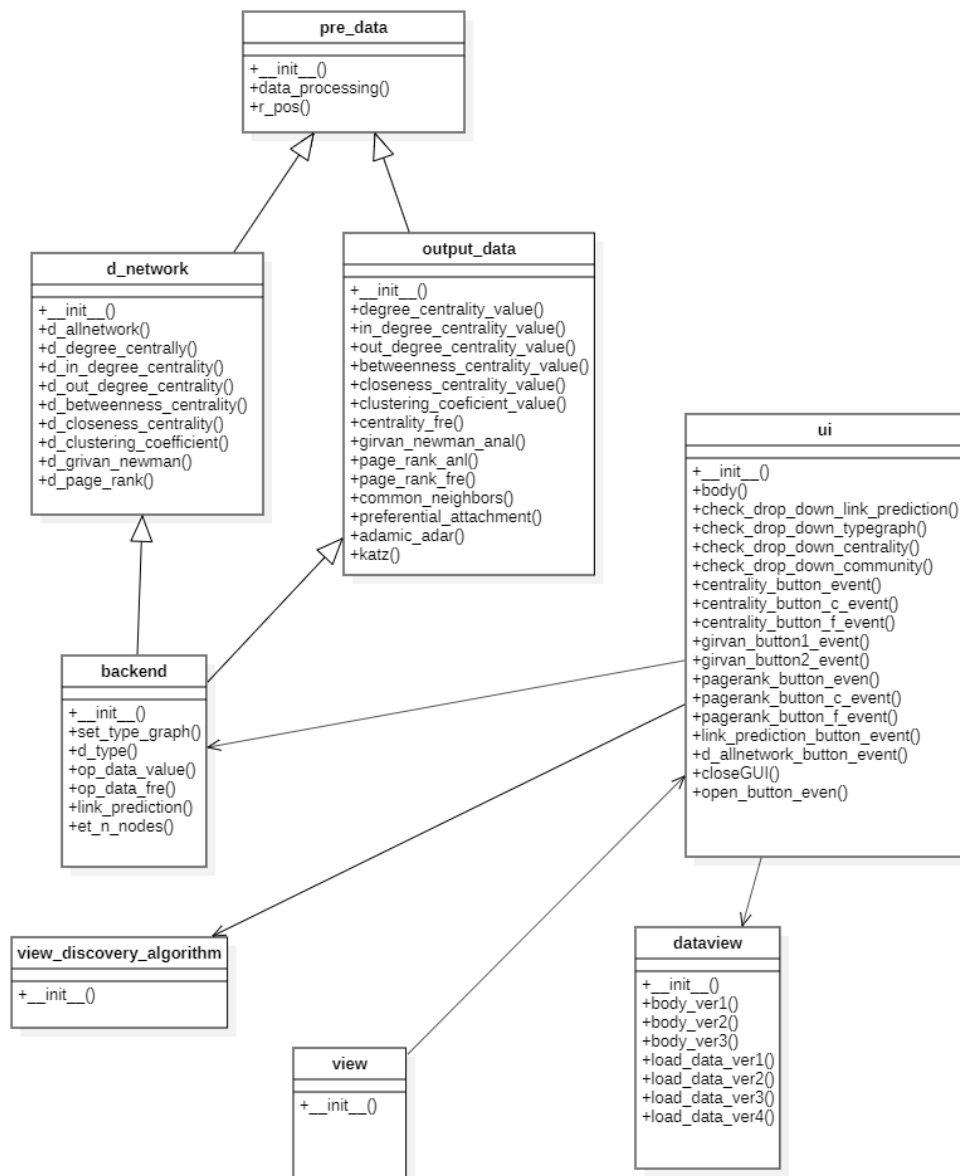
Use Case Diagram:



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

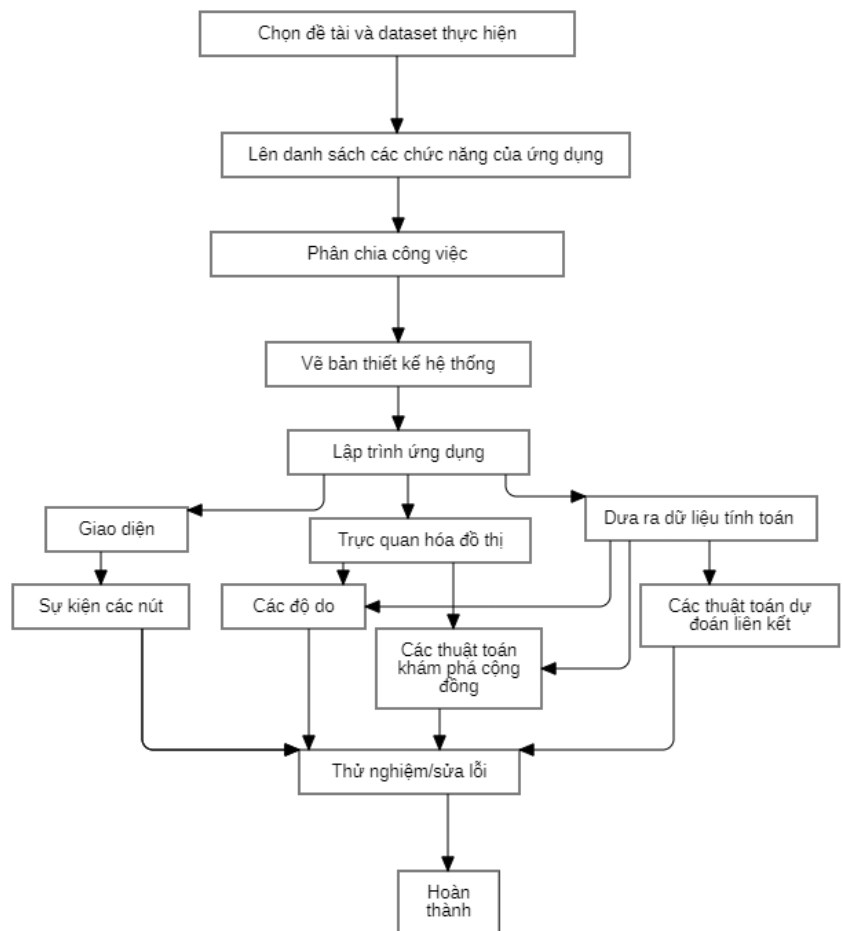
## Class Diagram :



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

## 2.2. Quy trình thực hiện



Hình 3: Quy trình thực hiện đề tài

## 2.3. Đánh giá kết quả

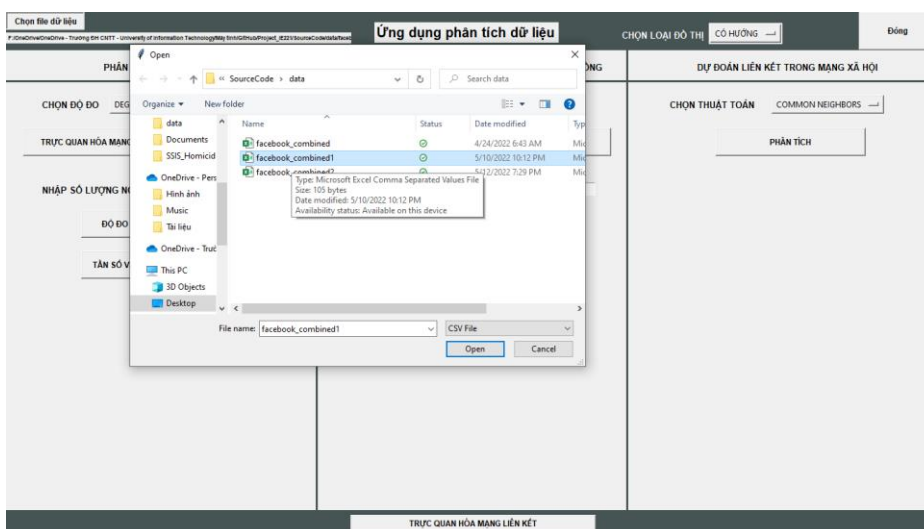
Chọn file dữ liệu			Ứng dụng phân tích dữ liệu		CHỌN LOẠI ĐỒ THỊ	CÓ HƯỚNG	Đóng
PHÂN TÍCH CÁC ĐỘ ĐO		PHÂN TÍCH BẢNG CÁC THUẬT TOÁN KHÁM PHÁ CỘNG ĐỒNG		DỰ ĐOÁN LIÊN KẾT TRONG MẠNG XÃ HỘI			
CHỌN ĐỘ ĐO		CHỌN THUẬT TOÁN		CHỌN THUẬT TOÁN			
DEGREE CENTRALITY		PAGE RANK		COMMON NEIGHBORS			
TRỰC QUAN HÓA MẠNG LIÊN KẾT ĐỘ ĐO DEGREE CENTRALITY		TRỰC QUAN HÓA MẠNG LIÊN KẾT THEO PAGE RANK		PHÂN TÍCH			
NHẬP SỐ LƯỢNG NODE HIỂN THỊ		NHẬP SỐ LƯỢNG NODE HIỂN THỊ					
ĐỘ ĐO DEGREE CENTRALITY		PAGE RANK					
TẦN SỐ VÀ XÁC SUẤT XUẤT HIỆN		TẦN SỐ VÀ XÁC SUẤT XUẤT HIỆN					
TRỰC QUAN HÓA MẠNG LIÊN KẾT							

Hình 4: Ứng dụng sau khi đã hoàn thành

### 2.3.1. Import và xử lý dữ liệu đầu vào

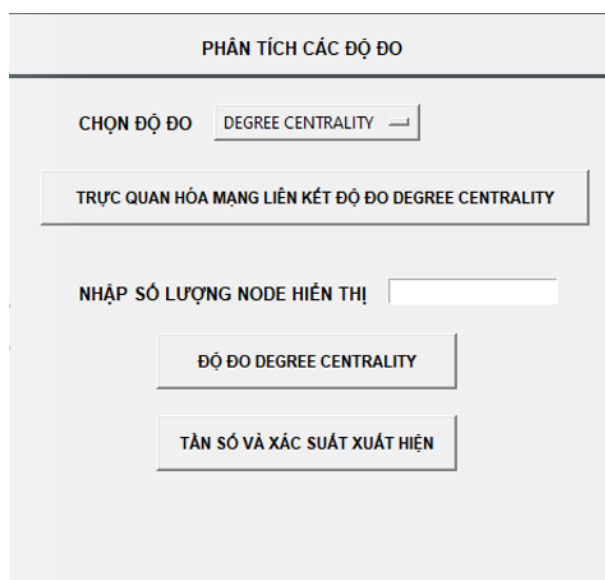
Đã hoàn thành được việc import và xử lý file \*csv đầu vào tạo graph, các biến dữ liệu để lưu kết quả để thực hiện các công việc tính toán.

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



Hình 5: Chọn dữ liệu đầu vào từ máy tính

### 2.3.2. Cụm chức năng phân tích các độ đo



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 đo đã chọn.

- Độ đo : Tính toán giá trị độ đo theo độ đo đã chọn và hiển thị 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.

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

#### Thuật toán Page Rank:

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 đo : Tính toán giá trị Page Rank và hiển thị 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.



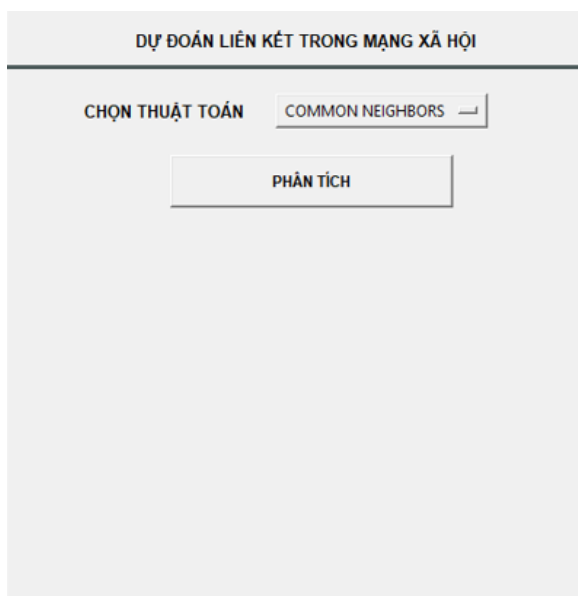
## Thuật toán Girvan Newman:



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 : Hiện thị đồ thị theo các cộng đồng đã tách được.

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



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.

### **3. KẾT LUẬN**

#### **3.1. 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.

#### **3.2. Nhược 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**

- [1] [Modelling Influence in a Social Network: Metrics and Evaluation](#); Tác giả : Behnam Hajian, Tony White; July 2011
- [2] [Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network](#); Tác giả : Junlong Zhang, Yu Luo; 2017
- [3] [Defense resource allocation in road dangerous goods transportation network: A SelfContained Girvan-Newman Algorithm and Mean Variance Model combined approach](#), Tác giả : Wencheng Huang, Linqing Li, Hongyi Liua, Rui Zhanga, Minhao Xua; 2016
- [4] [The Algorithm of Link Prediction on Social Network](#), 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.