**ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**Báo cáo đồ án môn học**

**MẠNG XÃ HỘI**

*Chủ Đề : Phân tích mạng xã hội trực tuyến Meetup.com*

Giáo viên hướng dẫn : Thái Bảo Trân

Lớp : IS353.J21

Thực hiện : PhạmVăn Hữu -15520293

TP HCM, Ngày 12 tháng 6 năm 2019

**MỤC LỤC**

Contents

[**I. Lý do chọn đề tài 5**](#_Toc11059211)

[**II. Thông tin dữ liệu 5**](#_Toc11059212)

[**III. Phân tích mạng xã hội nhóm với nhóm trên Meetup ở Nashville 9**](#_Toc11059213)

[**1. Xây dựng và vẽ đồ thị 9**](#_Toc11059214)

[**1.1. Nhập thư viện và đọc dữ liệu 9**](#_Toc11059215)

[**1.2. Xem thông tin Dataset 10**](#_Toc11059216)

[**1.3. Xem các thống kê cơ bản của dữ liệu 10**](#_Toc11059217)

[**1.4. Vẽ đồ thị mạng với các loại layouts khác nhau 11**](#_Toc11059218)

[**2. Phân tích đồ thị 16**](#_Toc11059219)

[**2.1. Tính các đặc trưng của mạng 16**](#_Toc11059220)

[**2.2. Tìm nhóm trung tâm trong mạng 19**](#_Toc11059221)

[**IV. Phân tích các nhóm thuộc danh mục ‘Tech’ trên Meetup ở Nashville 20**](#_Toc11059222)

[**1. Xây dựng và vẽ đồ thị 20**](#_Toc11059223)

[**1.1. Nhập thư viện và đọc dữ liệu 20**](#_Toc11059224)

[**1.2. Tạo đồ thị các nhóm thuộc danh mục ‘Tech’ trên Meetup ở Nashville 20**](#_Toc11059225)

[**1.3. Vẽ đồ thị các nhóm thuộc danh mục ‘Tech’ 21**](#_Toc11059226)

[**2. Phân tích đồ thị 22**](#_Toc11059227)

[**2.1. Tính các đặc trưng của mạng 22**](#_Toc11059228)

[**2.2. Tìm nhóm trung tâm trong mạng 22**](#_Toc11059229)

[**V. Ứng dụng-Xây dựng một hệ tư vấn đơn giản (Recommender System) 23**](#_Toc11059230)

[**1. Đặt vấn đề 23**](#_Toc11059231)

[**2. Ý tưởng 23**](#_Toc11059232)

[**3. Thực hiện ý tưởng 24**](#_Toc11059233)

[**3.1. Import thư viên và lấy những thông tin cần thiết từ dataset 24**](#_Toc11059234)

[**3.2. Gộp dữ liệu hai bảng trên 25**](#_Toc11059235)

[**3.3. Lấy giá trị trung bình số lần tham gia sự kiện 25**](#_Toc11059236)

[**3.4. Lấy giá trị tổng số lần tham gia sự kiện 26**](#_Toc11059237)

[**3.5. Thực hiện đề xuất nhóm tương tự 26**](#_Toc11059238)

[**VI. Kết luận 28**](#_Toc11059239)

[**1. Ưu điểm 28**](#_Toc11059240)

[**2. Nhược điểm 28**](#_Toc11059241)

[**3. Hướng phát triển đồ án 29**](#_Toc11059242)

[**VII. Bảng phân công công việc 29**](#_Toc11059243)

[**VIII. Tài liệu tham khảo 29**](#_Toc11059244)

# Lý do chọn đề tài

Meetup là một trang web mạng xã hội nhằm mục đích mang mọi người đến với nhau để cùng làm việc, khám phá, dạy và học về những điều mà có thể giúp cuộc sống của họ trở nên sống động có ý nghĩa hơn.

Meetup cho phép các người dùng có thể tìm và tham gia vào các nhóm (nhóm được tạo ra từ tập hợp những người có cùng chung sở thích,nghề nghiệp ,lĩnh vực ,..).Kể từ năm 2017,Meetup đã có hơn 32 triệu người dùng với 280 nghìn nhóm tại 182 quốc gia.Các thành viên xác định nhóm hay hoạt động mà họ quan tâm nhất để có thể sử dụng mạng xã hội này hiệu quả.

Để hiểu hơn về cấu trúc mạng xã hội Meetup là như thế nào?Mối quan hệ giữa các thành viên trong nhóm?Các nhóm có mối quan hệ gì với nhau?Các nhóm và các thành viên trọng mạng có ảnh hưởng với nhau như thế nào?Xong,từ những hiểu biết rút ra từ việc phân tích mạng xã hội Meetup.com chúng ta có thể xây dựng một hệ thống khuyến nghị các nhóm hay các hoạt động đến các thành viên trong mạng dựa trên sở thích hay sự quan tâm về một lĩnh vực nào đó của họ.

# Thông tin dữ liệu

* Tên dataset : Nashville Meetup NetworkX
* Data sources:
  + member-to-group-edges.csv : Danh sách cạnh cho việc xây dựng đồ thị *member-to-group* với trọng số là số lượng các sự kiện mà thành viên đã tham dự trong mỗi nhóm.
  + group-edges.csv : Danh sách cạnh cho việc xây dựng đồ thị *group-to-group* với trọng số là số thành viên chung giữa 2 nhóm.
  + member-edges.csv : Danh sách cạnh cho việc xây dựng đồ thị *member-to-member* với trọng số là số nhóm chung giữa 2 thành viên.
  + rsvps.csv : Dữ liệu thô về việc tham dữ các sự kiện của các thành viên trong một nhóm ,được tổng hợp từ *member-to-group-edges.csv.*
  + meta-Groups.csv: Thông tin cho từng nhóm, bao gồm tên và danh mục.
  + meta-thành viên.csv: Thông tin cho từng thành viên, bao gồm tên và địa điểm
  + meta-event.csv: Thông tin cho từng sự kiện, bao gồm tên và thời gian.
* Nguồn dataset: <https://www.kaggle.com/stkbailey/nashville-meetup>
* Thông số dataset *member-to-group-edges.csv*:
  + Số thuộc tính: 3
  + Số dòng: 45600
  + Dung lượng file: 943.3 KB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 | member\_id | ID của thành viên |
| 2 | group\_id | ID của nhóm |
| 3 | weight | Trọng số - số lượng các sự kiện mà thành viên đã tham dự trong mỗi nhóm |

* Thông số dataset *group-edges.csv*:
  + Số thuộc tính: 4
  + Số dòng: 6693
  + Dung lượng file: 163.24 KB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 |  | Thứ tự |
| 2 | group1 | ID Nhóm 1 |
| 3 | group2 | ID Nhóm 2 |
| 4 | weight | Trọng số - số thành viên chung giữa 2 nhóm. |

* Thông số dataset *member-edges.csv*:
  + Số thuộc tính: 4
  + Số dòng: 1180000
  + Dung lượng file: 32.87 MB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 |  | Thứ tự |
| 2 | member1 | ID thành viên 1 |
| 3 | member2 | ID thành viên 2 |
| 4 | weight | Trọng số - số nhóm chung giữa 2 thành viên. |

* Thông số dataset *rsvps.csv*:
  + Số thuộc tính: 4
  + Số dòng: 127000
  + Dung lượng file: 4.37 MB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 |  | Thứ tự |
| 2 | event\_id | ID sự kiện |
| 3 | member\_id | ID thành viên |
| 4 | group\_id | ID nhóm |

* Thông số dataset *meta-groups.csv*:
  + Số thuộc tính: 7
  + Số dòng: 602
  + Dung lượng file: 59.66 KB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 | group\_id | ID nhóm |
| 2 | group\_name | Tên nhóm |
| 3 | num\_members | Số lượng thành viên |
| 4 | category\_id | ID Danh mục |
| 5 | category\_name | Tên danh mục |
| 6 | organizer\_id | ID người tổ chức |
| 7 | group\_urlname | Tên url của nhóm |

* Thông số dataset *meta-members.csv*:
  + Số thuộc tính: 7
  + Số dòng: 24600
  + Dung lượng file:1.22 MB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 | member\_id | ID thành viên |
| 2 | name | Tên thành viên |
| 3 | hometown | Quê quán |
| 4 | city | Thành phố |
| 5 | state | Tiểu Bang |
| 6 | lat |  |
| 7 | lon |  |

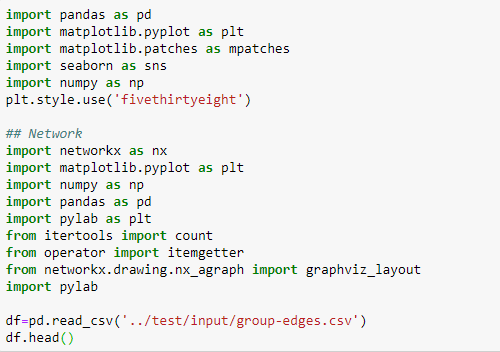
* Thông số dataset *member-to-group-edges.csv*:
  + Số thuộc tính: 4
  + Số dòng: 19300
  + Dung lượng file:1.46 MB
  + Bảng thông tin của thuộc tính

|  |  |  |
| --- | --- | --- |
| STT | Tên | Ý nghĩa |
| 1 | event\_id | ID sự kiện |
| 2 | group\_id | ID nhóm |
| 3 | name | Tên sự kiện |
| 4 | time | Thời gian diễn ra |

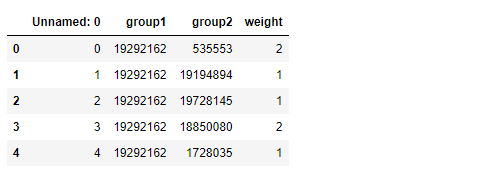
# Phân tích mạng xã hội nhóm với nhóm trên Meetup ở Nashville

## Xây dựng và vẽ đồ thị

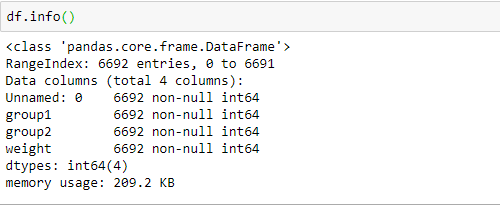
### Nhập thư viện và đọc dữ liệu



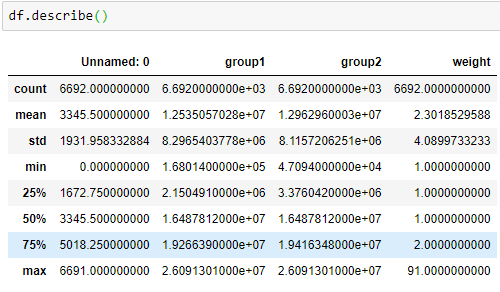
Kết quả:



### Xem thông tin Dataset



### Xem các thống kê cơ bản của dữ liệu

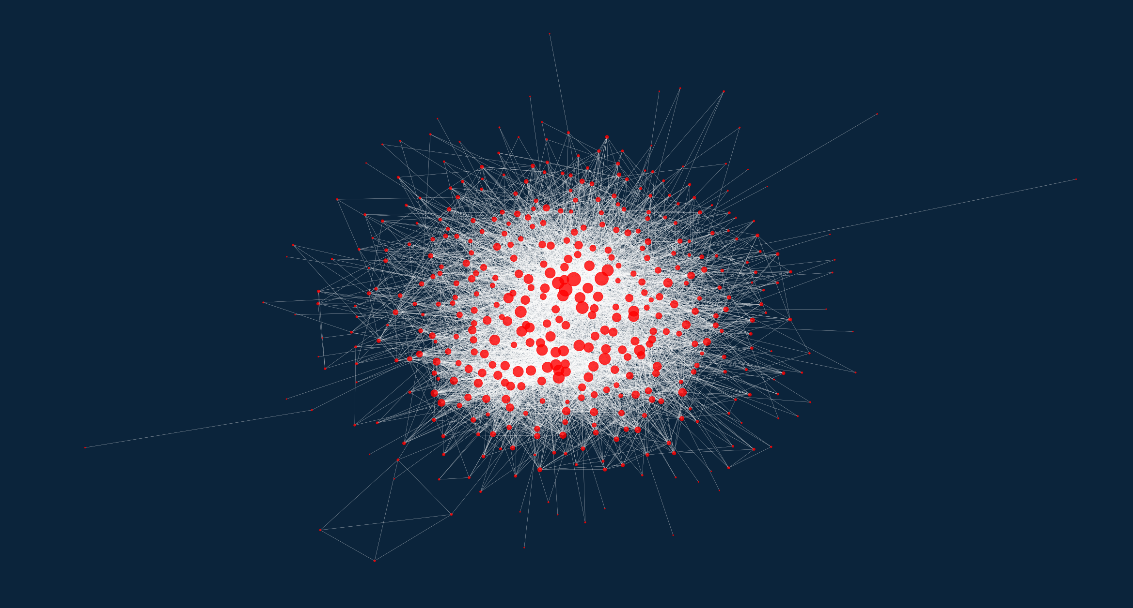


### Vẽ đồ thị mạng với các loại layouts khác nhau

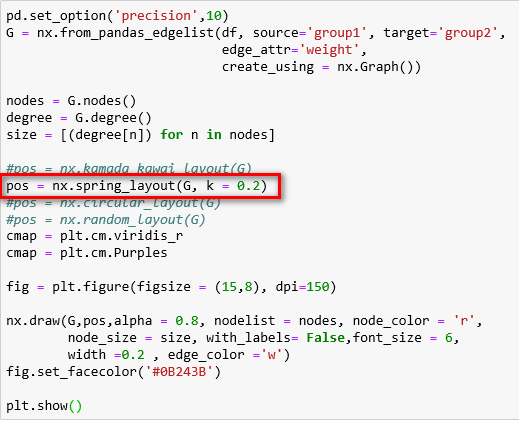
Đối với dữ liệu các nhóm trên Meetup.com ở khu vực Nashville thì mỗi đỉnh sẽ là các nhóm độc lập trên Meetup và các liên kết giữa hai nhóm sẽ hình thành từ số thành viên chung của hai nhóm đó.Trọng số là số lượng thành viên chung của hai nhóm.

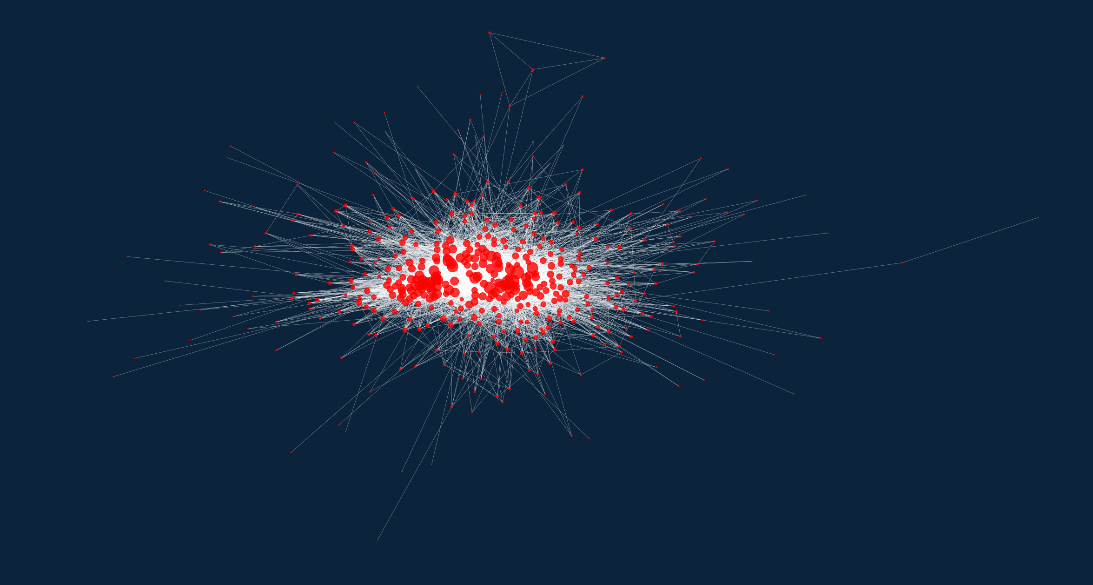
#### Kamada\_kawai\_layout





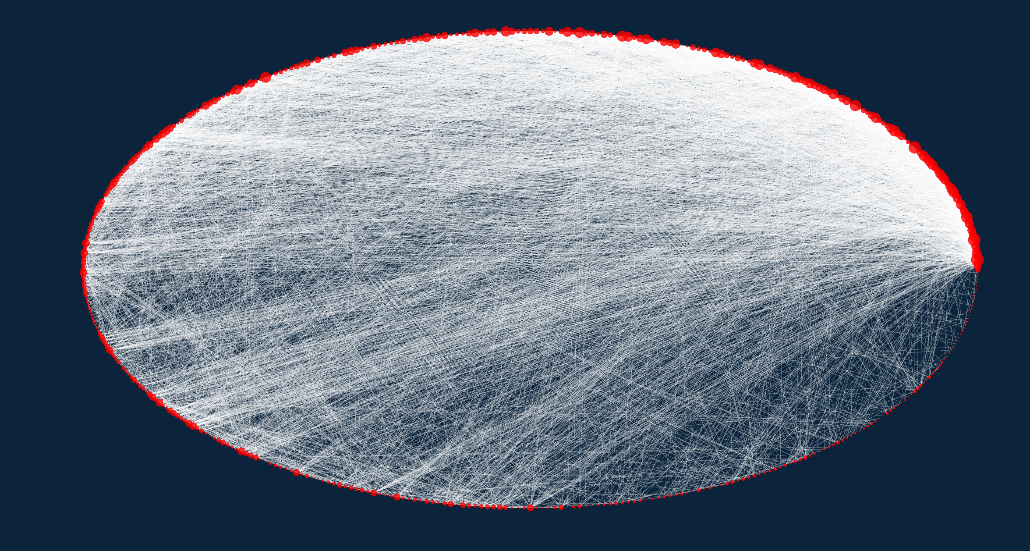
#### Spring\_layout





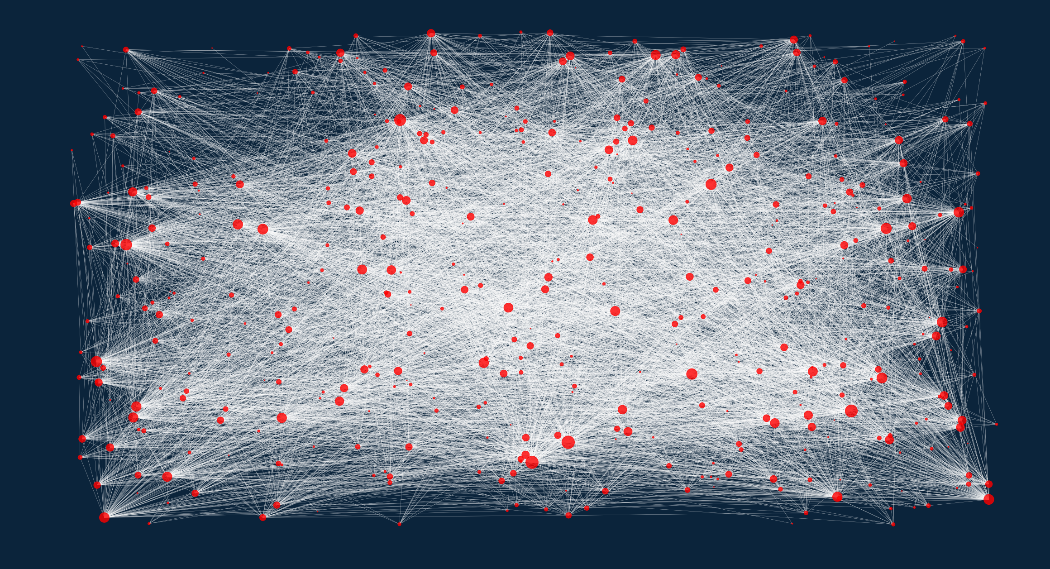
#### Circular\_layout





#### Random\_layout

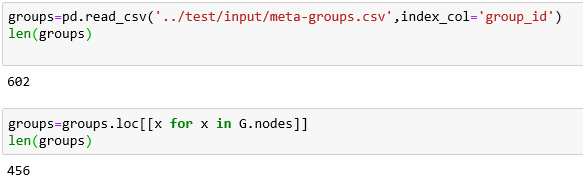




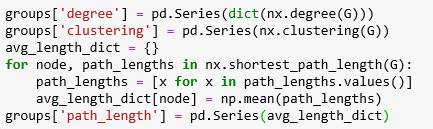
## ****Phân tích đồ thị****

### Tính các đặc trưng của mạng

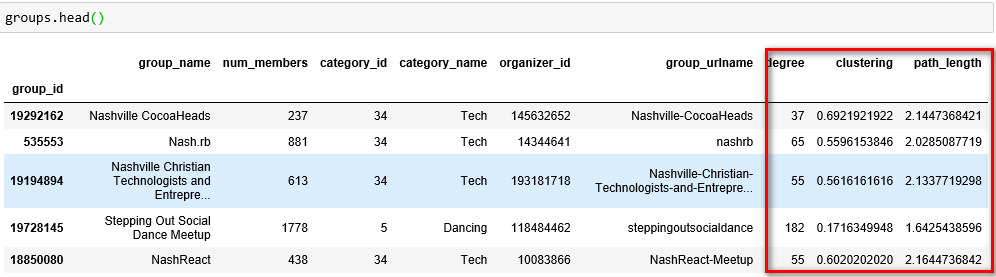
Đọc dữ liệu file *meta-groups.csv* để lấy ra thông tin những nhóm tồn tại trong đồ thị.Từ đó ta tính các đặc trưng như Degree, Clustering,Path length và thêm các giá trị đó vào DataFrame



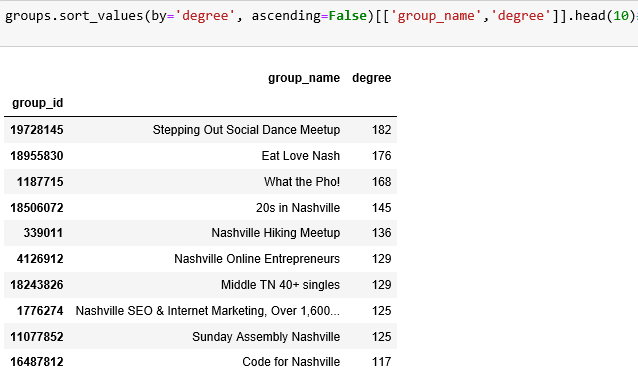
Tính các đặc trưng và thêm vào DataFrame:



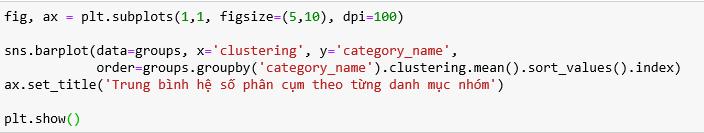
Kết quả:



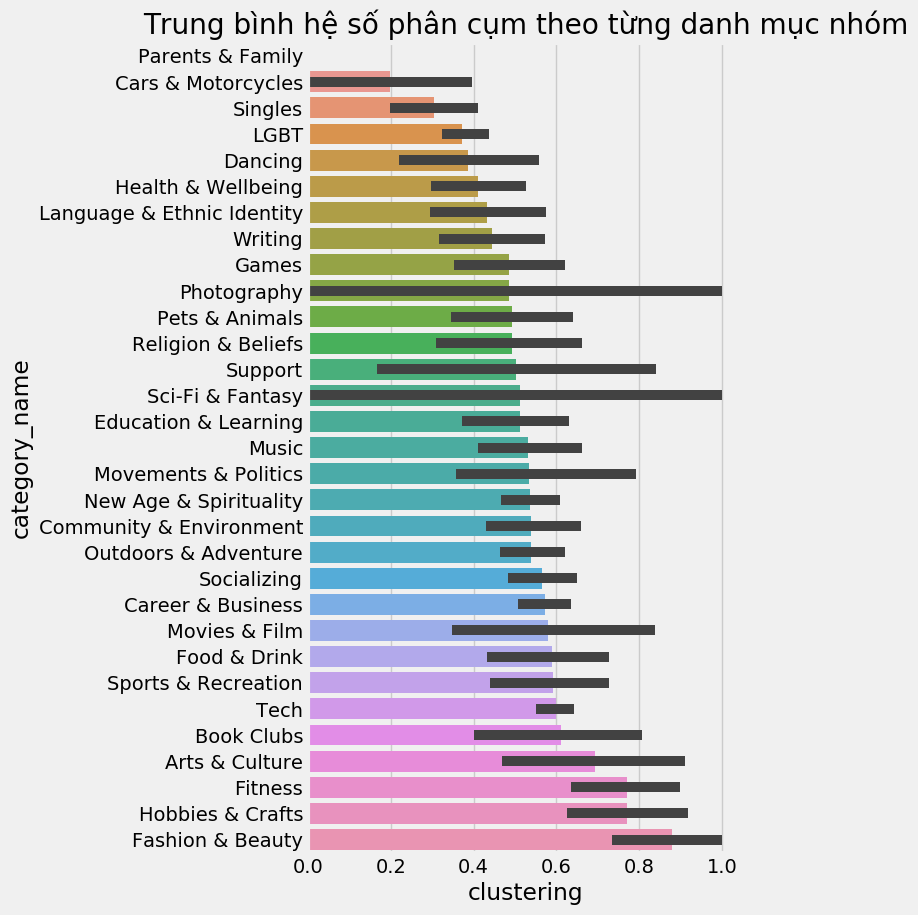
#### Top 10 nhóm có giá trị Degree cao nhất trong mạng



#### Tính trung bình hệ số phân cụm theo từng danh mục nhóm



Kết quả :

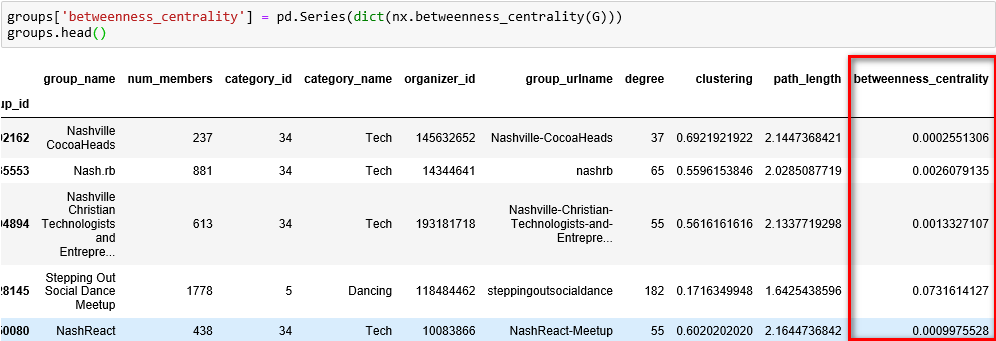


Ta có thể thấy được các nhóm như Tech, Book Clubs, Art & Culture, Fitness, Hoobies & Crafts và Fashion & Beauty là những danh mục rất phổ biến.Như đã đề cập khi xây dựng đồ thị thì những liên kết được tạo ra từ số thành viên chung giữa hai nhóm nên những thành viên thuộc những nhóm của các danh mục phổ biến này cũng sẽ có xu hướng tham gia vào các nhóm trong các danh mục phổ biến kể trên.

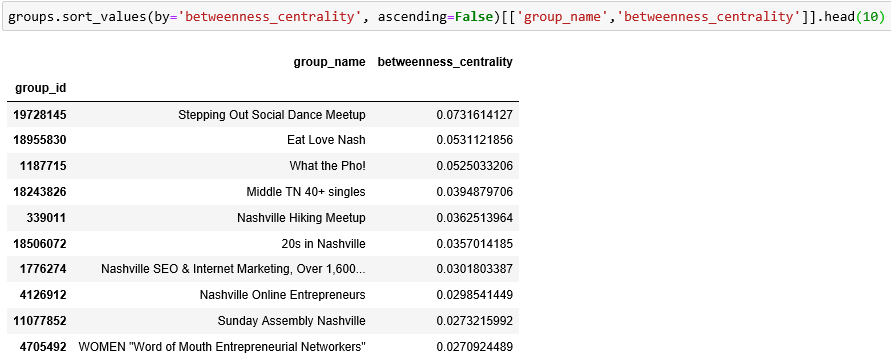
### Tìm nhóm trung tâm trong mạng

#### Tính Betweenness Centrality

Tính giá trị betweenness centrality trong mạng và thêm vào Dataframe



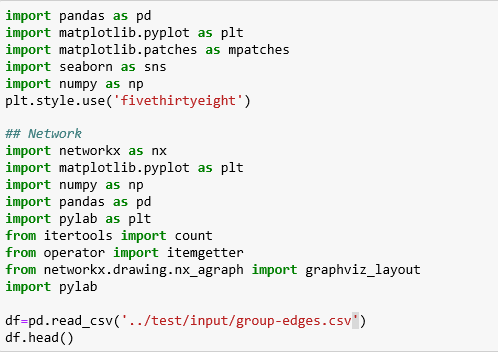
#### Top 10 nhóm trung tâm trong mạng



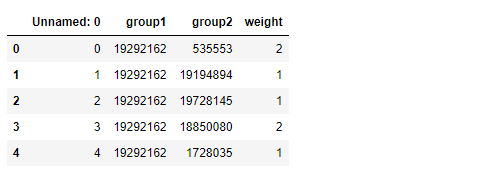
# Phân tích các nhóm thuộc danh mục ‘Tech’ trên Meetup ở Nashville

## Xây dựng và vẽ đồ thị

### Nhập thư viện và đọc dữ liệu



Kết quả :



### Tạo đồ thị các nhóm thuộc danh mục ‘Tech’ trên Meetup ở Nashville

* Tạo đồ thị các nhóm trên Meetup



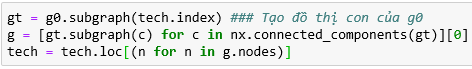
* Đọc dữ liệu thông tin các nhóm



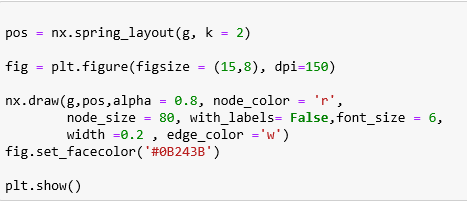
* Lấy ra các nhóm thuộc danh mục ‘Tech’

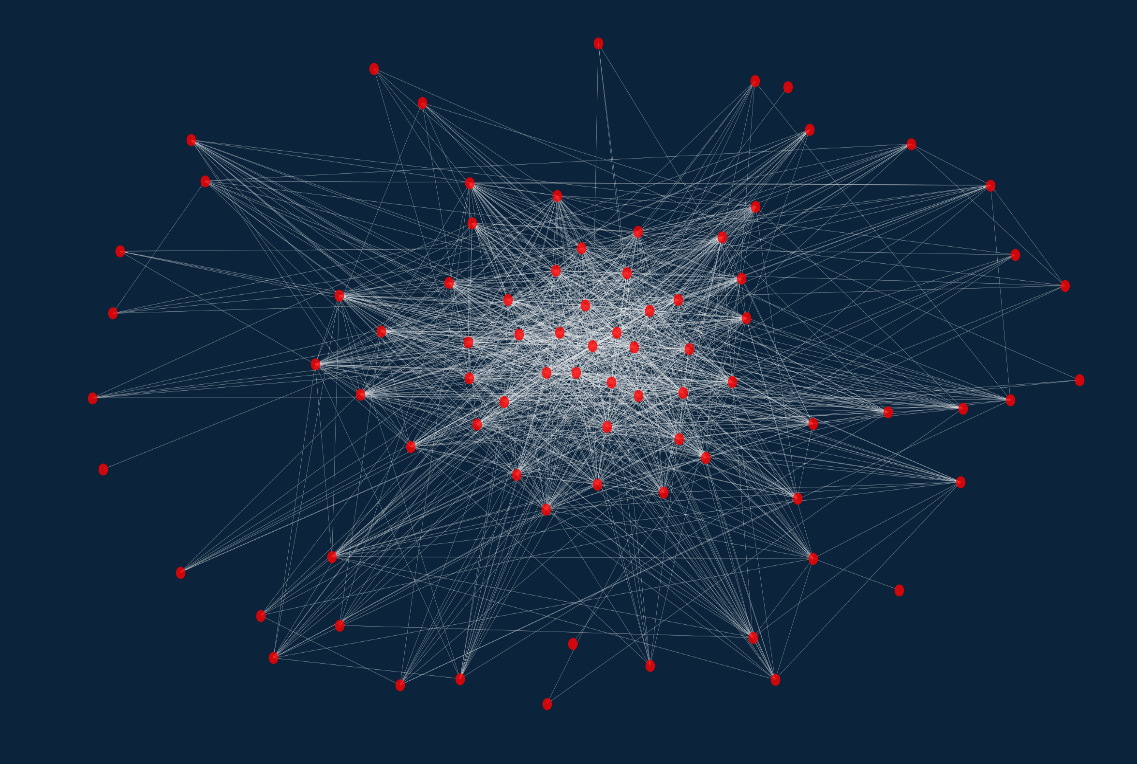


* Tạo đồ thị các nhóm thuộc danh mục ‘Tech’ từ đồ thị các nhóm trên Meetup



### Vẽ đồ thị các nhóm thuộc danh mục ‘Tech’

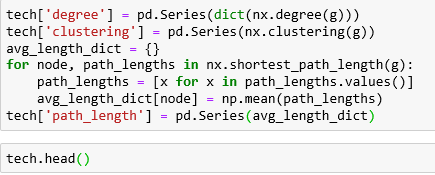




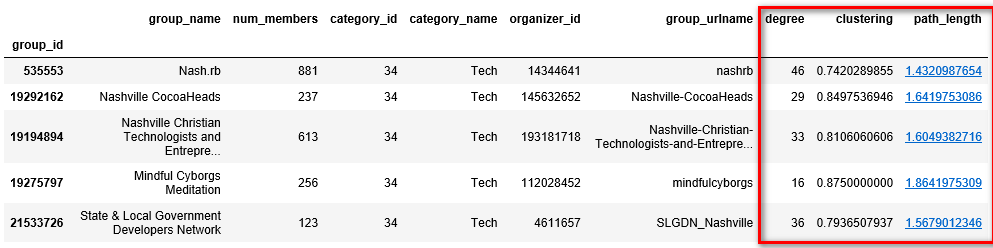
## Phân tích đồ thị

### Tính các đặc trưng của mạng

Tính các đặc trưng như Degree, Clustering,Path length và thêm các giá trị đó vào DataFrame



Kết quả:

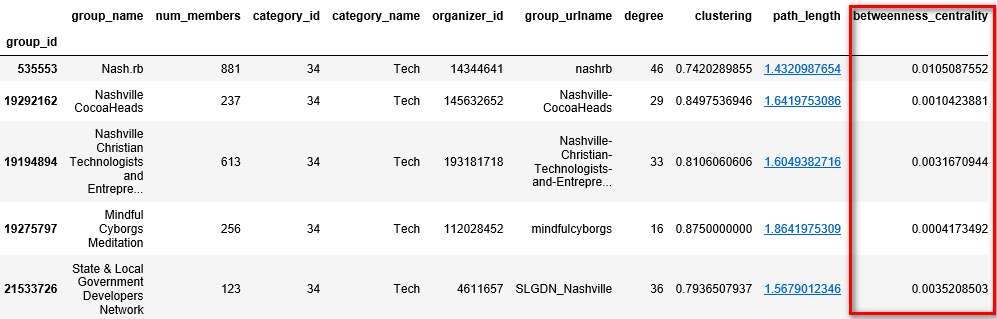


### Tìm nhóm trung tâm trong mạng

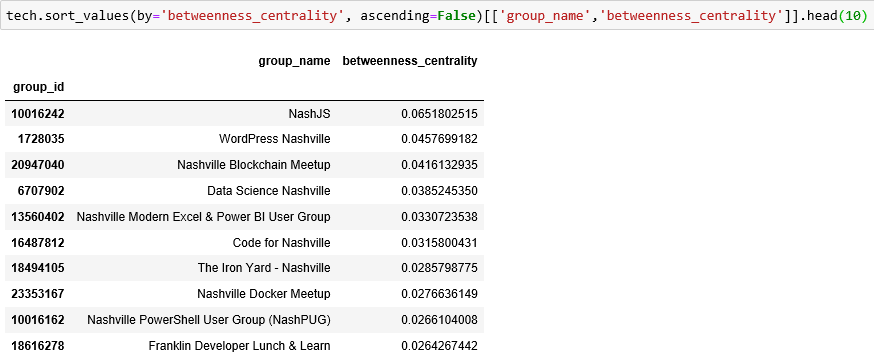
#### Tính Betweenness Centrality



Kết quả :



#### Top 10 nhóm trung tâm của mạng các nhóm ‘Tech’



# Ứng dụng-Xây dựng một hệ tư vấn đơn giản (Recommender System)

## Đặt vấn đề

Giả sử khi bạn mới bắt đầu sử dụng mạng xã hội Meetup và muốn tham gia vào một nhóm nào đó.Ví dụ cụ thể ở đây là bạn muốn tìm một nhóm thuộc danh mục ‘Tech’ và muốn tham gia vào nhóm ‘NashJS’ (nhóm trung tâm của danh mục).Và sau khi tham gia vào nhóm này rồi thì bạn cũng có nhu cầu tham gia vào những nhóm khác .Vậy bạn sẽ tham gia vào nhóm nào trong hàng trăm nhóm? Nhóm nào có tính tương đồng như nhóm mình đã tham gia?Hệ tư vấn sẽ giúp bạn làm điều đó.

## Ý tưởng

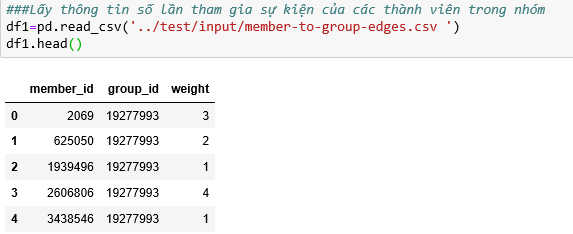
Mạng xã hội Meetup được tạo ra nhằm mục đích có thể tạo ra một mạng lưới các nhóm để những người dùng có thể tham gia vào các nhóm thuộc sở thích,lĩnh vực học tập,nghiên cứu,…và để tham gia vào các sự kiện trong nhóm để gặp những người có cùng sở thích,kiến thức,kinh nghiệm trong nhóm để học tập,chia sẽ,giao lưu,..Như vậy thì ta có thể sử dụng thuộc tính số lần tham gia sự kiện của các thành viên trong nhóm và trung bình số lần tham dự sự kiện trên một thành viên của nhóm để tìm ra sự tương đồng này.Hiểu một cách đơn giản là bạn sẽ tìm ra những nhóm có nhiều thành viên,nhiều sự kiện được tổ chức và số lần tham gia sự kiện của các thành trong nhóm có cao hay không?Điều này giúp bạn có thể thỏa mãn được nhu cầu tìm ra được nhóm có nhiều sự kiện được tổ chức,nhóm hoạt động sôi nổi,trong nhóm có nhiều thành viên tích cực thông qua số lần tham gia sự kiện của thanh viên đó,…

## Thực hiện ý tưởng

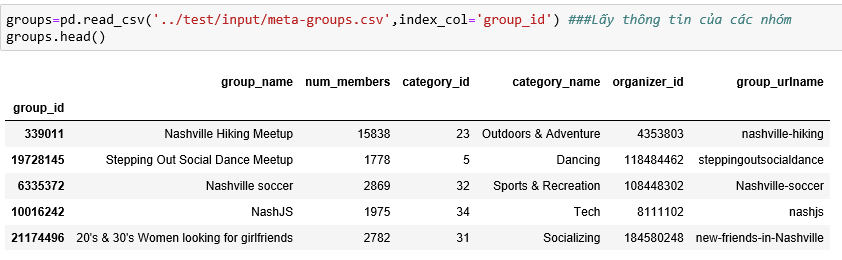
### Import thư viên và lấy những thông tin cần thiết từ dataset



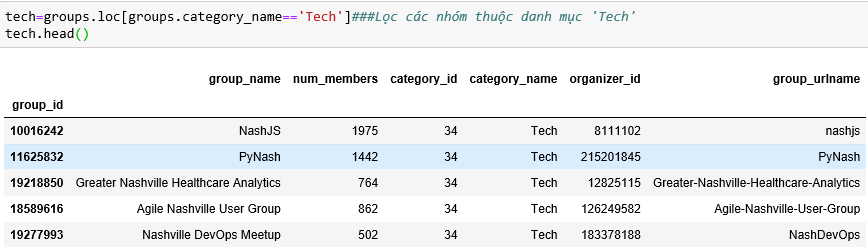
* Lấy số lần tham gia sự kiện của các thanh viên trong nhóm



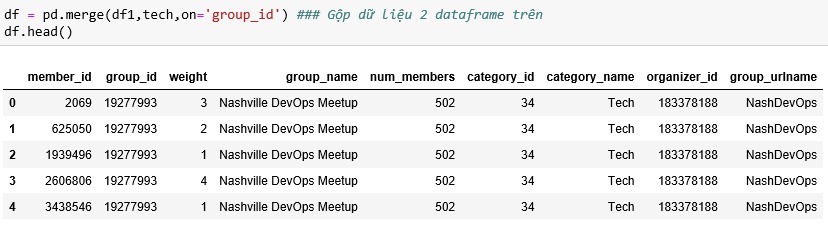
* Lấy thông tin của nhóm



* Lọc các nhóm thuộc danh mục ‘Tech’

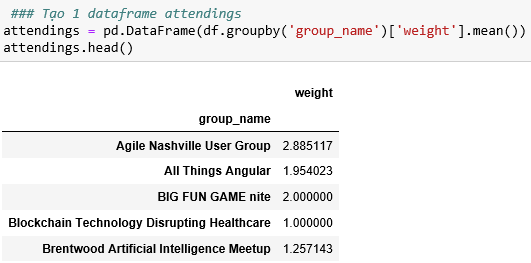


### Gộp dữ liệu hai bảng trên



### Lấy giá trị trung bình số lần tham gia sự kiện

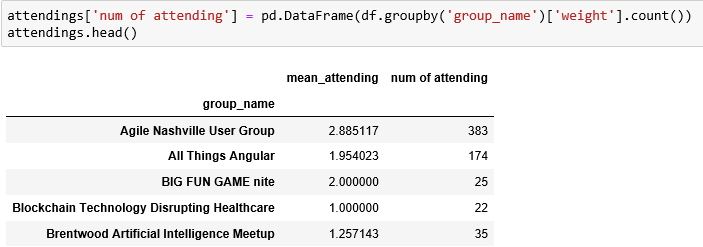
Lấy giá trị trung bình số lần tham gia sự kiện của các thành viên theo tên nhóm và tạo một dataframe **attendings**:



Giá trị ‘*weight’* bây giờ là giá trị trung bình số lần tham gia sự kiện của các thành viên trong nhóm.

### Lấy giá trị tổng số lần tham gia sự kiện

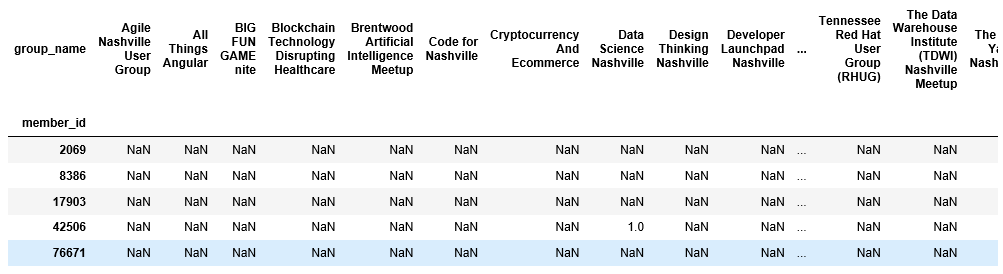
Đếm tổng số lần tham gia sự kiện của các thành viên theo tên nhóm và thêm giá trị này vào bảng **attendings** ở trên:



### Thực hiện đề xuất nhóm tương tự

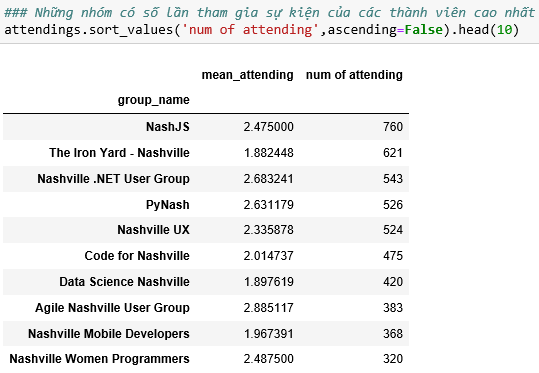
Trước tiên chúng ta cần tạo một ma trận *group\_matrix* trong đó *member\_id* sẽ nằm trên một trục và trục còn lại là thông tin tên nhóm.Mỗi ô sẽ là số lần mà thành viên đó tham dự sự kiện trong nhóm trên.



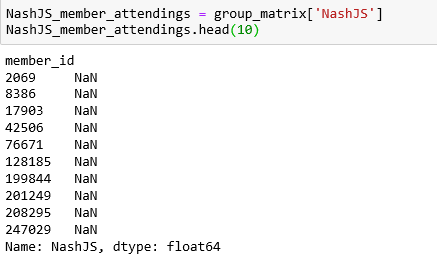


Ta có thể thấy có rất nhiều giá trị NaN là bởi vì không phải tất cả các người dùng đều thuộc các thành viên của các nhóm trên.

* Những nhóm có số lần tham gia sự kiện của các thành viên cao nhất:



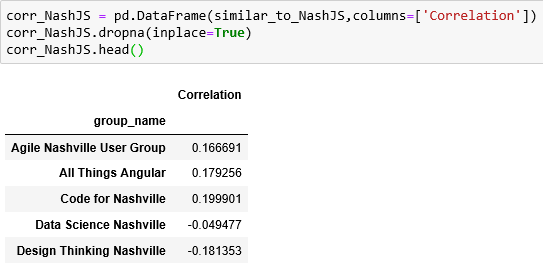
* Chọn nhóm ‘NashJS’ và lấy thông tin tham dự sự kiện của các thành viên:



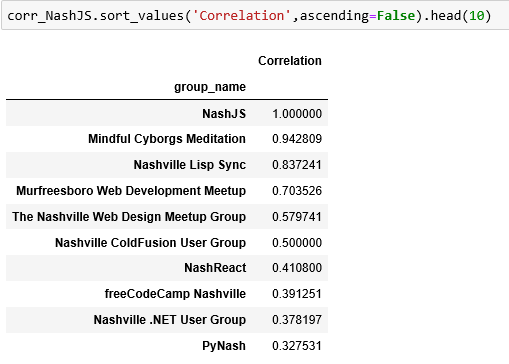
* Sử dụng phương thức *corrwith()* để lấy được những tương quan với nhóm ‘NashJS’



* Xóa các giá trị NaN và sử dụng DataFrame để hiển thị sự tương quan:



* Sắp xếp các giá trị tương đồng:



Ta có thể thấy bên trên là 9 nhóm tương đồng với ‘NashJS’ được đề xuất với tỉ lệ tương đồng từ 32% đến 94%.Tương tự ta có thể làm điều này với các nhóm còn lại.

# Kết luận

## Ưu điểm

Việc phân tích mạng xã hội mang đến những kiến thức mới,cũng như được hiểu thể về một mạng xã hội mới giúp ích cho mọi người.Việc thực hiện những phân tích mạng xã hội Meetup giúp chúng ta có thể hiểu kết cấu của mạng xã hội này và có thể rèn luyện và thực hành các kỹ năng phân tích một mạng xã hội với NetworkX.

Dataset **Nashville Meetup NetworkX** là một dataset cơ bản và được xây dựng sẵn các danh sách các liên kết trong mạng vì thể rất tiện lợi trong việc phân tích.

## Nhược điểm

Những kết quả đạt được chỉ nằm ở mức cơ bản nên cũng chưa có những đánh giá xác thực và chính xác nhất về mạng xã hội Meetup.

## Hướng phát triển đồ án

Do đồ án tập trung vào phân tích đối tượng *Group* trên Meetup nên chưa có cái nhìn tống quát về hai đối tượng *Member* và *Event* nên việc cần làm:

* Tiếp tục phân tích hai đối tượng Member và Event trong dataset.
* Sử dụng các thuật toán mới để khám phá cộng đồng
* Dự đoán những liên kết mới trong mạng.

# Bảng phân công công việc

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Tìm dataset | Mô tả dữ liệu | Phân tích mạng giữa các nhóm trong dataset | Phân tích mạng giữa các nhóm trong danh mục ‘Tech’ | Ứng dụng-xây dựng hệ tư vấn đơn giản | Kết luận | Viết báo cáo,slide,  hướng dẫn demo | Thuyết trình |
| **Hữu** | **X** | **X** | **X** | **X** | **X** | **X** | **X** | **X** |

# Tài liệu tham khảo

* <https://towardsdatascience.com/build-your-own-recommender-system-within-5-minutes-30dd40388fbf>
* <https://networkx.github.io/documentation/networkx-1.9.1/index.html>
* <https://www.kaggle.com/stkbailey/nashville-meetup>
* <https://www.kaggle.com/sirpunch/meetups-data-from-meetupcom>