Twitter: Politic community started from @realDonaldTrump

GIÁO VIÊN HƯỚNG DẪN: TS. NGUYỄN KIÊM HIẾU

LỮ MẠNH HÙNG	20142078	NGUYỄN VĂN MẠNH	20142873
NGUYỄN MẠNH HÙNG	20142088	CAO VĂN THIỆN	20144262
NGUYỄN ĐỨC DUY	20140724	HÀ VĂN QUANG	20143578

Nội dung trình bày

- Giới thiệu bài toán
- Crawl và xử lý dữ liệu
- Visualize đồ thị
- Đánh giá kết quả

Giới thiệu bài toán

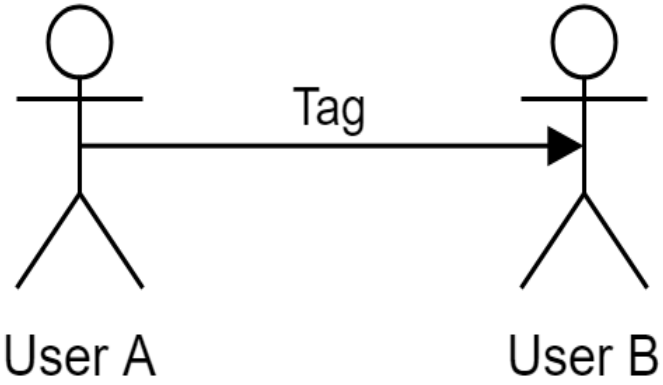
- Đề tài: "Phát hiện cộng đồng chính trị từ @realDonaldTrump“
- Mục tiêu của đề tài là phát hiện các cộng đồng chính trị trên Twitter.

➢ Nhìn nhận được các diễn biến chính trị đang diễn ra ở nước Mỹ nói riêng và trên thế giới nói chung

Crawl và xử lý dữ liệu

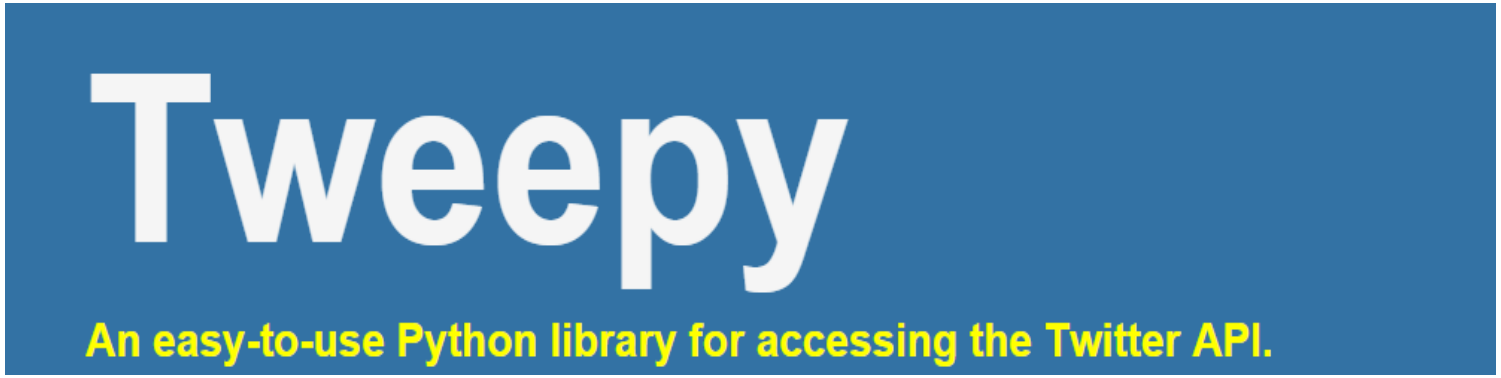
Chiến lược crawl dữ liệu

- Node bắt đầu: *"@realDonaldTrump"* mở rộng theo từng level
- Các tweet được crawl từ khi Donald Trump bắt đầu nhậm chức tổng thống: 11/9/2016 đến 1/5/2018
- Mở rộng bằng Tag nên đồ thị của nhóm là đồ thị có hướng (hình minh họa)



Crawl và xử lý dữ liệu
Thư viện sử dụng

- Tweepy: thư viện chạy trên Python cung cấp khả năng truy cập các Twitter APIs dễ dàng hơn
- Search API: cho phép thu thập dữ liệu dựa trên từ khóa tìm kiếm
- Streaming API: được phép truy xuất để lấy dữ liệu Twitter theo thời gian thực



Crawl và xử lý dữ liệu

Xử lý dữ liệu

- Xây dựng một module nhận diện quan điểm "political tweet" để lọc các tweet liên quan đến chính trị
- Tìm user được tag bắt đầu bằng ký tự "@" và xử lý các trường hợp user không tồn tại *"api.get_user(username)"*

Crawl và xử lý dữ liệu

Module lọc political tweets

- Tách từ *word_tokenize* sử dụng thư viện *nltk (Natural Language Toolkit)*
- Lọc ký tự, dấu câu và số chỉ giữ lại các từ
- Lọc stopword english
- Sử dụng từ điển chính trị *"Stanford Political Dictionary"* gồm 3778 từ.
- Mỗi từ của tweet nằm trong từ điển *score = score + 1* hoặc *score = score – 1*
- Kết quả score > 0, thì tweet đó liên quan đến chính trị và ngược lại.

	A	B	C	D	E	F	G	H
1	Word	Positive	Negative	Strong	Weak	Active	Passive	
2	abandon		2		3		3	
3	abandonment		2		3		3	
4	abdicat		1		3		3	
5	abet	2				3		
6	abhor		3				2	
7	abhorrenc		3					
8	abhorrent		3					
9	abid						3	
10	abiliti	1		3				
11	ability	1		2				
12	abl			3				
13	abnormal		3					
14	abolish		1	3		3		
15	abolition		2	3		2		
16	abominabl		3					
17	abominat		3				1	
18	abomination		3					

Crawl và xử lý dữ liệu

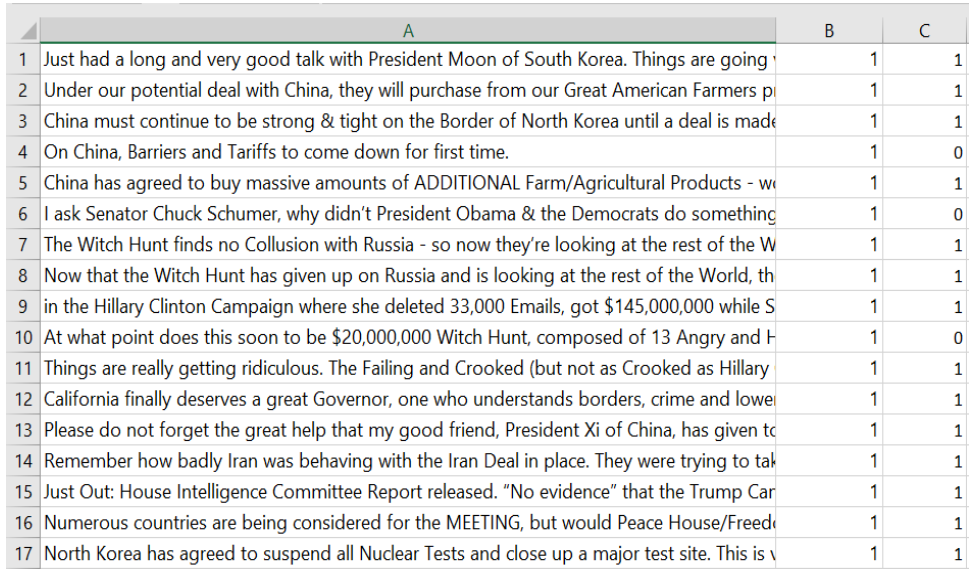
Module lọc political tweets

- Hiệu quả:

Nhóm tự xây dựng test data gồm 100 tweet

- Confusion matrix

n = 100	Pridicted: Yes	Predicted: No	
Actual: Yes	TP = 34	FN = 16	50
Actual: No	FP = 16	TN = 34	50
	50	50	



	A	B	C
1	Just had a long and very good talk with President Moon of South Korea. Things are going	1	1
2	Under our potential deal with China, they will purchase from our Great American Farmers p	1	1
3	China must continue to be strong & tight on the Border of North Korea until a deal is made	1	1
4	On China, Barriers and Tariffs to come down for first time.	1	0
5	China has agreed to buy massive amounts of ADDITIONAL Farm/Agricultural Products - we	1	1
6	I ask Senator Chuck Schumer, why didn't President Obama & the Democrats do something	1	0
7	The Witch Hunt finds no Collusion with Russia - so now they're looking at the rest of the W	1	1
8	Now that the Witch Hunt has given up on Russia and is looking at the rest of the World, th	1	1
9	in the Hillary Clinton Campaign where she deleted 33,000 Emails, got $145,000,000 while S	1	1
10	At what point does this soon to be $20,000,000 Witch Hunt, composed of 13 Angry and H	1	0
11	Things are really getting ridiculous. The Failing and Crooked (but not as Crooked as Hillary	1	1
12	California finally deserves a great Governor, one who understands borders, crime and lowe	1	1
13	Please do not forget the great help that my good friend, President Xi of China, has given to	1	1
14	Remember how badly Iran was behaving with the Iran Deal in place. They were trying to tak	1	1
15	Just Out: House Intelligence Committee Report released. "No evidence" that the Trump Car	1	1
16	Numerous countries are being considered for the MEETING, but would Peace House/Freed	1	1
17	North Korea has agreed to suspend all Nuclear Tests and close up a major test site. This is	1	1

Độ chính xác: P = 34/(34+16) = 0.68

Độ bao phủ: R = 34/(34+16) = 0.68

$$F = \frac{2 \times 0.68 \times 0.68}{0.68 + 0.68} = 0.68$$

Crawl và xử lý dữ liệu

Kết quả

- Đồ thị gồm có: 71981 nodes và 126929 cạnh

Trong đó:

- Level 1: 242 nodes
- Level 2: 34544 nodes

	A	B
1	Source	Target
2	realDonaldTrump	HowieCarrShow
3	realDonaldTrump	FoxNews
4	realDonaldTrump	charliekirk11
5	realDonaldTrump	RandPaul
6	realDonaldTrump	SenSchumer
7	realDonaldTrump	JackPosobiec
8	realDonaldTrump	DonnaWR8
9	realDonaldTrump	Netanyahu
10	realDonaldTrump	IvankaTrump
11	realDonaldTrump	WhiteHouse
12	realDonaldTrump	TuckerCarlson
13	realDonaldTrump	DHSgov
14	realDonaldTrump	shawgerald4
15	realDonaldTrump	Cubs
16	realDonaldTrump	GovChristie
17	realDonaldTrump	Comcast
18	realDonaldTrump	TBN
19	realDonaldTrump	ClemsonFB
20	realDonaldTrump	AmericanLegion
21	realDonaldTrump	FoxFriendsFirst
22	realDonaldTrump	AndyPuzder
23	realDonaldTrump	BradThor
24	realDonaldTrump	TurnbullMalcolm
25	realDonaldTrump	newtgingrich
26	realDonaldTrump	MittRomney
27	realDonaldTrump	GovMikeHuckabee
28	realDonaldTrump	USCG
29	realDonaldTrump	HouseGOP

Visualize đồ thị

Giải thuật modularity

- Phân chia bằng độ đo modularity của phân vùng
- Giá trị của modularity trong khoảng từ -1 đến 1, đo độ dày đặc của các liên kết trong cùng một cộng đồng so với các liên kết giữa các cộng đồng

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

  - A_{ij} là trọng số của cạnh nối giữa i và j
  - $k_i = \sum_j A_{ij}$ là tổng trọng số của các cạnh được nối với đỉnh i
  - c_i là cộng đồng mà đỉnh i thuộc vào
  - Hàm δ (u, v) có giá trị bằng 1 khi u = v và 0 nếu ngược lại
  - $m = \frac{1}{2} \sum_{ij} A_{ij}$

Visualize đồ thị
Giải thuật modularity

- Thuật toán bao gồm 2 pha chính thực hiện lặp lại:

Pha 1:

- Giả sử chúng ta bắt đầu với một mạng có N nút. Ban đầu chúng ta gán mỗi một cộng đồng khác nhau cho mỗi nút trong mạng
- Mỗi nút i ta xem xét các hàng xóm j của nó, tính toán giá trị nhận được của modularity nếu rời nút i từ cộng đồng của nó sang cộng đồng của nút j
- Nút i sau đó được đặt vào một cộng đồng mà làm cho giá trị nhận được này lớn nhất

Visualize đồ thị
Giải thuật modularity

- Thuật toán bao gồm 2 pha chính thực hiện lặp lại:

Pha 1:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

Trong đó:

- $\sum in$ là tổng các trọng số của các liên kết bên trong cộng đồng C
- $\sum tot$ là tổng các trọng số của các liên kết đi đến các nút trong C
- k_i là tổng trọng số các liên kết chạm đến đỉnh I
- $k_{i,\,in}$ là tổng các trọng số của các liên kết từ i đến các nút trong C
- m là tổng số các liên kết trong mạng

Visualize đồ thị
Giải thuật modularity

- Thuật toán bao gồm 2 pha chính thực hiện lặp lại:

Pha 2: thực hiện xây dựng một mạng mới mà các nút bây giờ là các cộng đồng đã được nhận diện ở pha trước

- Liên kết giữa các nút mới được cho bởi tổng trọng số của các liên kết giữa các nút tương ứng trong hai cộng đồng
- Các liên kết bên trong cộng đồng trở thành các liên kết tự lặp cho cộng đồng này trong mạng mới
- Một khi pha thứ hai được hoàn thành, lặp lại pha thứ nhất của thuật toán
- Thuật toán sẽ lặp đi lặp lại hai pha cho đến khi không còn sự thay đổi nào nữa, và một giá trị modularity tối ưu được nhận về

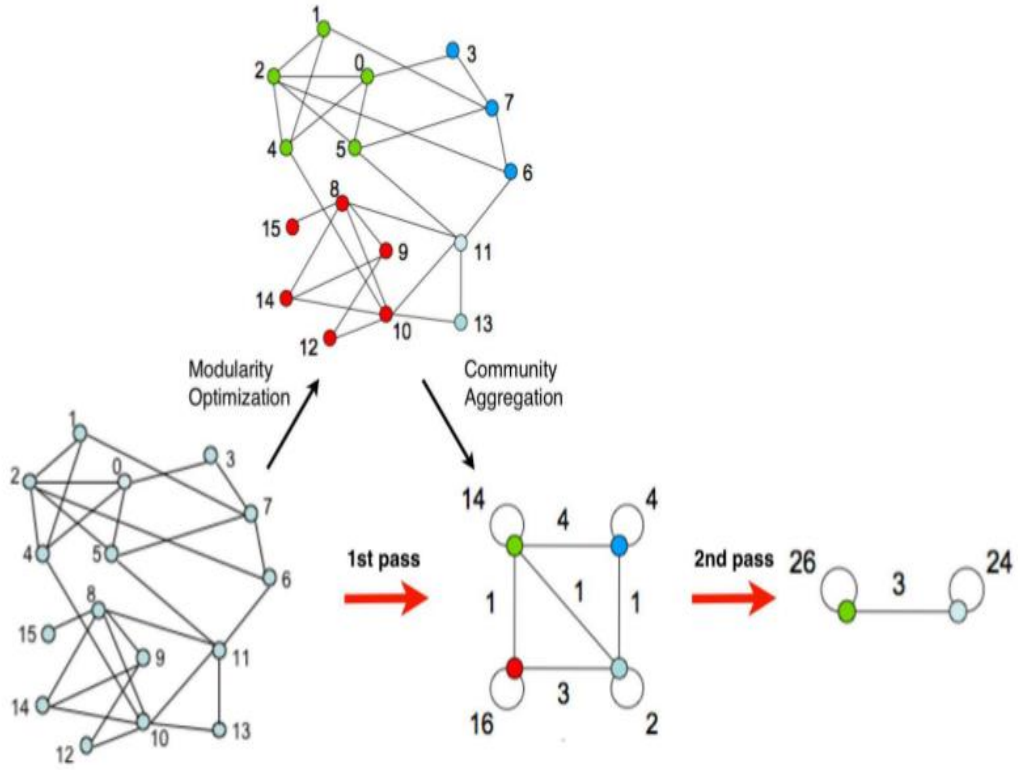
Visualize đồ thị
Giải thuật modularity

- Thuật toán bao gồm 2 pha chính thực hiện lặp lại:

Pha 2: thực hiện xây dựng một mạng mới mà các nút bây giờ là các cộng đồng đã được nhận diện ở pha trước

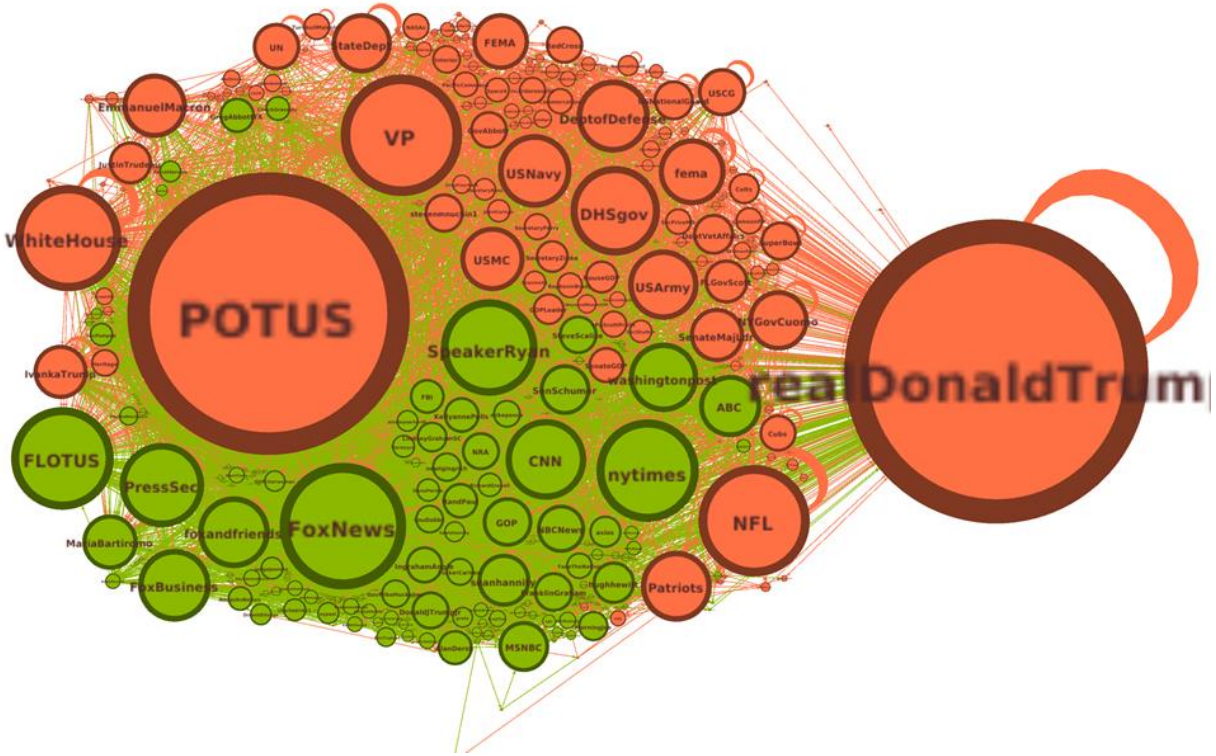
Tham số resolution:

- Cho phép dừng lại ở một trong các bước trung gian. Càng ở các bước đầu thì số cộng đồng càng lớn và kích thước cộng đồng nhỏ, trong khi càng về các bước sau thì ngược lại.



Visualize đồ thị

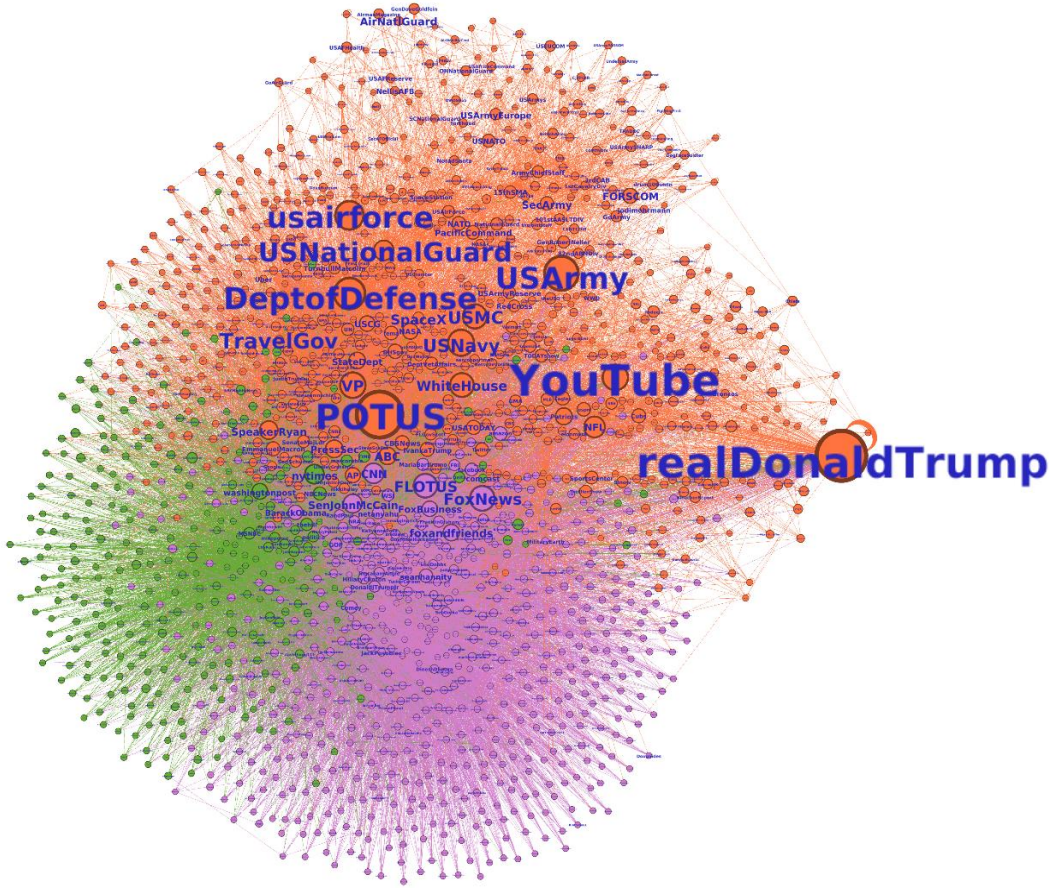
- Level 1: 366 nodes và 8564 cạnh



Level 1: resolution = 1.5 có 2 cộng đồng	
Tên cộng đồng	Mô tả
Đơn vị chính trị (màu cam)	*POTUS:* trang đại biểu cho tổng thống Trump *VP*: phó tổng thống Mỹ: Mike Pence *WhiteHouse*: Nhà Trắng *DHSgov*: cục an ninh nội địa Mỹ *DeptofDefense*: bộ quốc phòng Mỹ
Đơn vị truyền thông (xanh lá)	*SpeakerRyan*: người phát ngôn của Nhà Trắng *PressSec*: thư ký báo chí của nhà trắng *FLOTUS:* đệ nhất phu nhân nước Mỹ *FoxNews:* kênh tin tức truyền hình cáp số 1 nước Mỹ *CNN:* mạng truyền hình cáp tại Mỹ *nytimes:* nhật báo NewYork

Visualize đồ thị

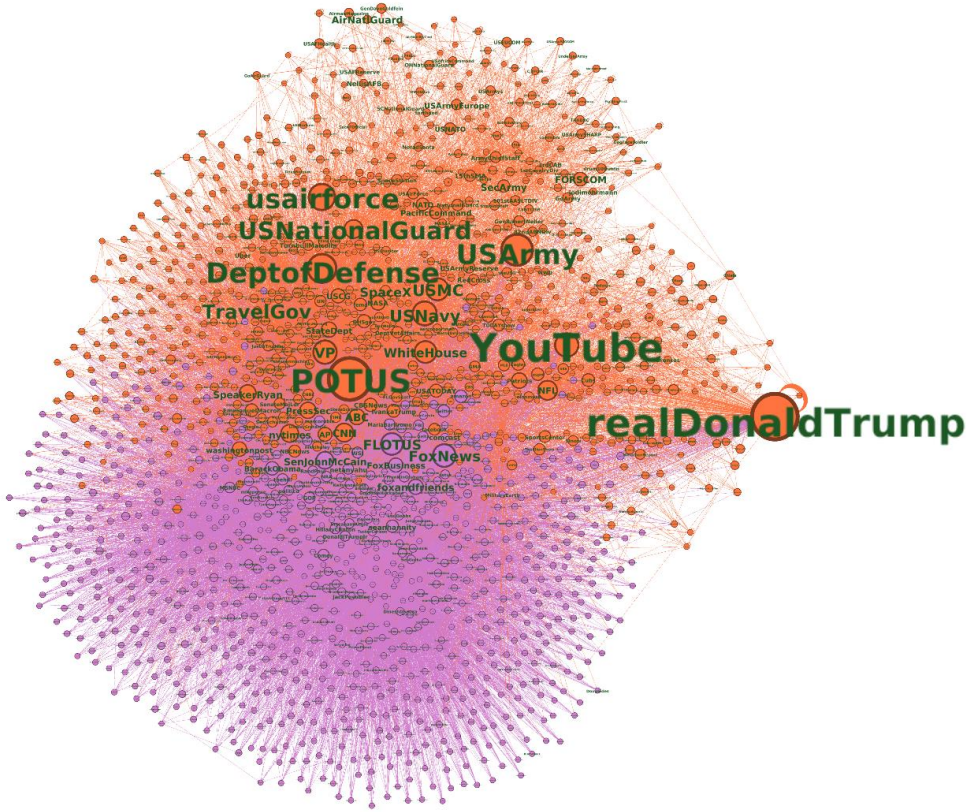
- Level 2: 54370 nodes và 107895 cạnh



Level 2: resolution = 1.0 có 3 cộng đồng	
Tên cộng đồng	Mô tả
Đơn vị chính trị (màu cam)	*POTUS:* trang đại biểu cho tổng thống Trump *DeptofDefense*: bộ quốc phòng Mỹ *USArmy*: quân đội Mỹ*VP*: phó tổng thống Mỹ: Mike Pence *WhiteHouse*: Nhà trắng
Đơn vị truyền thông – thiên hướng giải trí (màu tím)	*FLOTUS:* đệ nhất phu nhân nước Mỹ *FoxNews:* kênh tin tức truyền hình cáp số 1 nước Mỹ *CNN:* mạng truyền hình cáp tại Mỹ *FoxBusiness:* tạp chí kinh doanh của Fox*seanhannity:* kênh radio nổi tiếng
Đơn vị truyền thông – thiên hướng chính luận (xanh lá)	*nytimes:* nhật báo NewYork *washingtonpost:* nhật báo lâu đời nhất nước Mỹ *politico:* trang chuyên về các tin tức chính trị *thehill:* trang tin tức chính trị nổi tiếng

Visualize đồ thị

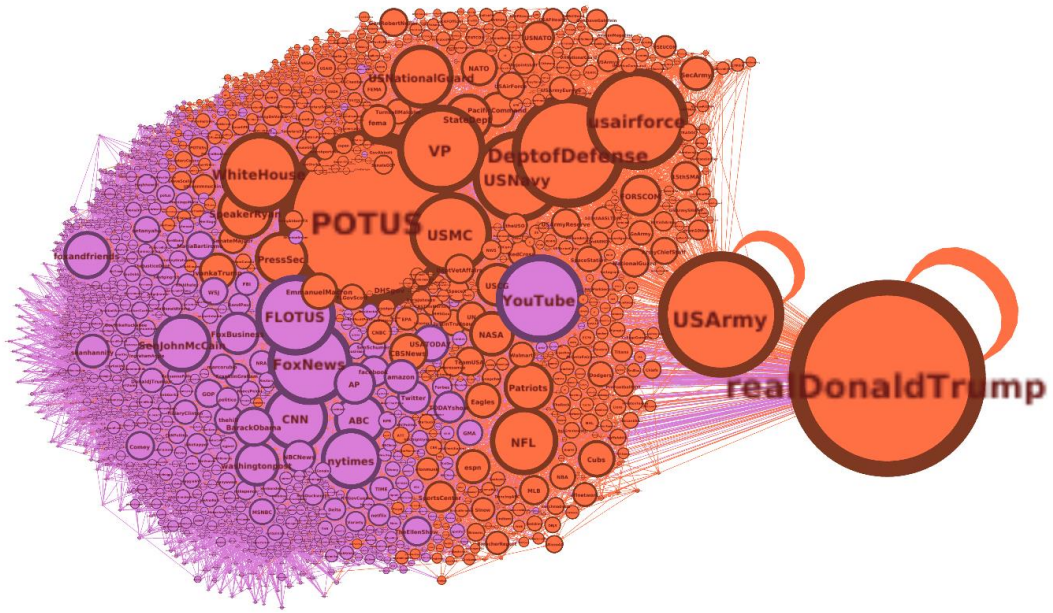
- Level 2: 54370 nodes và 107895 cạnh



Level 2: resolution = 2.0 có 2 cộng đồng	
Tên cộng đồng	Mô tả
Đơn vị chính trị (màu cam)	*POTUS:* trang đại biểu cho tổng thống Trump *DeptofDefense*: bộ quốc phòng Mỹ *usairforce*: không lực Mỹ *USNavy*: hải quân Mỹ *VP*: phó tổng thống Mỹ: Mike Pence *WhiteHouse*: Nhà trắng *USMC*: thủy quân lục chiến Mỹ
Đơn vị truyền thông (xanh lá)	*FLOTUS* (đệ nhất phu nhân nước Mỹ) *FoxNews* (kênh tin tức truyền hình cáp số 1 nước Mỹ) *FoxBusiness:* tạp chí kinh doanh của Fox *foxandfriends:* chương trình tin tức buổi sáng nước Mỹ *seanhannity:* kênh radio nổi tiếng

Visualize đồ thị

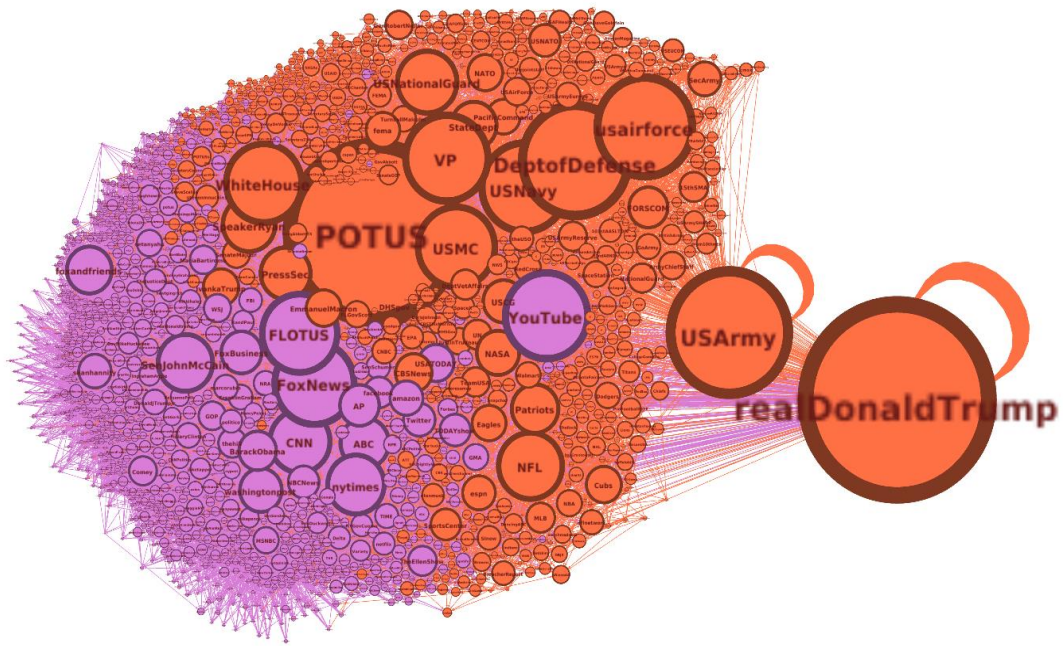
- All: 71981 nodes và 126929 cạnh



All graph: resolution = 1.75 có 2 cộng đồng	
Tên cộng đồng	Mô tả
Đơn vị chính trị (màu cam)	*POTUS:* trang đại biểu cho tổng thống Trump *DeptofDefense*: bộ quốc phòng Mỹ *usairforce*: không lực Mỹ *USNavy*: hải quân Mỹ *VP*: phó tổng thống Mỹ: Mike Pence *WhiteHouse*: Nhà trắng *USMC*: thủy quân lục chiến Mỹ *USNationalGuard*: cảnh sát quốc gia Mỹ
Đơn vị truyền thông (màu tím)	*SpeakerRyan*: người phát ngôn của Nhà Trắng *PressSec*: thư ký báo chí của nhà trắng *FLOTUS:* đệ nhất phu nhân nước Mỹ *FoxNews:* kênh tin tức truyền hình cáp số 1 nước Mỹ *CNN:* mạng truyền hình cáp tại Mỹ *nytimes:* nhật báo NewYork

Visualize đồ thị

Nhận xét đồ thị: Cộng đồng chính trị



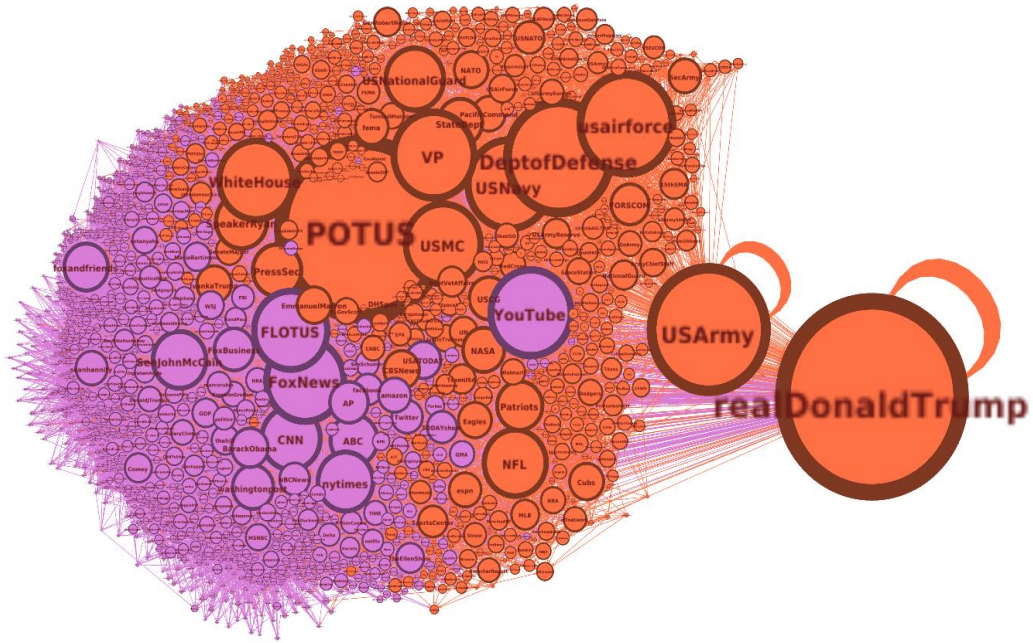
Từ *@realDonaldTrump* ta nhận thấy 2 cộng đồng lớn rõ rệt là cộng đồng hoạt động chính trị và cộng đồng hoạt động truyền thông.

<u>Cộng đồng chính trị</u>

- Node lớn nhất trong cộng đồng này là: POTUS - là trang phát ngôn chính thức của tổng thống chứ không phải là trang cá nhân.
- Các node VP (phó tổng thống Mỹ), WhiteHouse (Nhà trắng).

Visualize đồ thị

Nhận xét đồ thị: Cộng đồng chính trị



- Các node lớn chủ yếu trong cộng đồng này liên quan tới chính trị, đặc biệt là quân sự

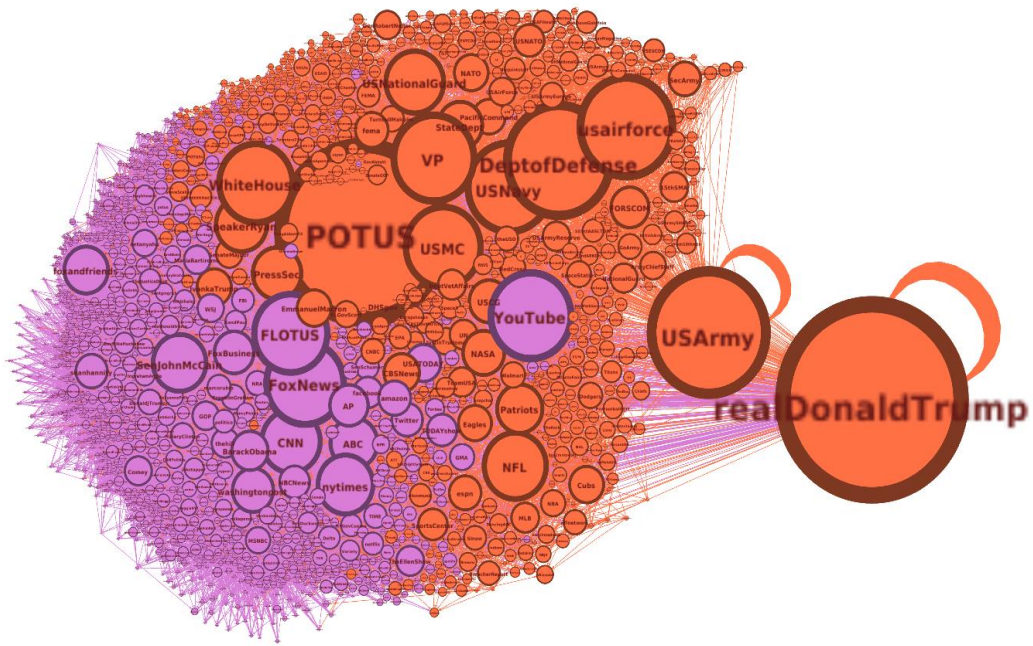
➢ Vấn đề chính trị quan tâm nhất của cộng đồng này là về vấn đề quân sự, và đặc biệt là sự can thiệp của Mỹ ở nước ngoài.

- Các node xa và nhỏ hơn như Hiệp ước quân sự Bắc Đại Tây Dương (NATO), Quân đội Mỹ đóng tại Nato (USNATO), đóng tại châu Âu (USArmyEurope)

➢ Chính sách xoay trục quân sự của Mỹ sang các điểm nóng mới nổi của thế giới như Trung Đông hay Triều Tiên

Visualize đồ thị

Nhận xét đồ thị: Cộng đồng chính trị

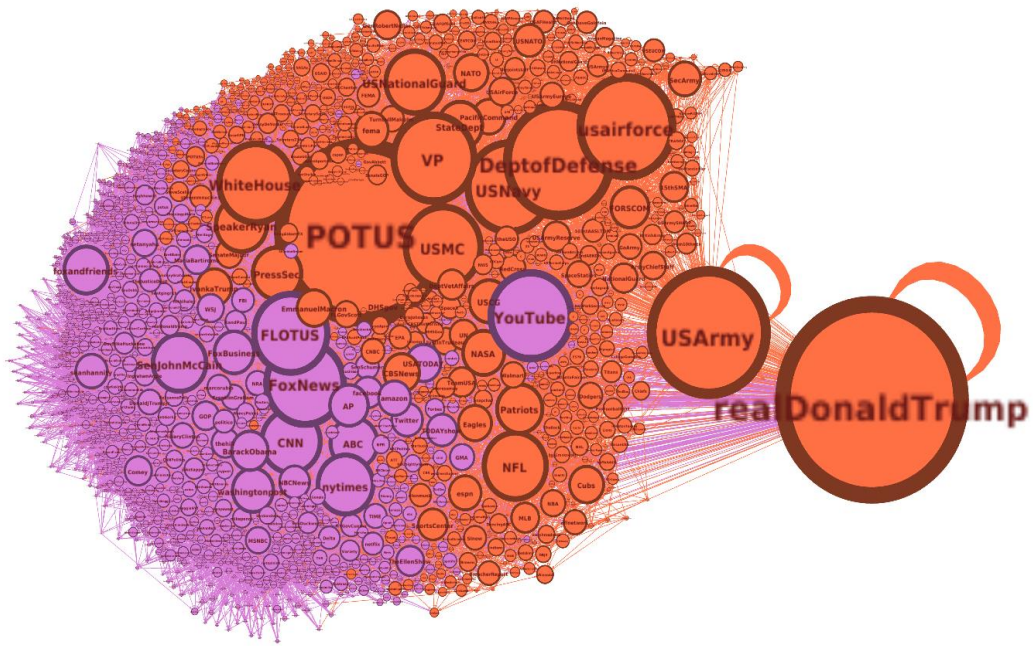


- 1 node khá lớn là Patriots (hệ thống tên lửa chống tên lửa đạn đạo), Mỹ cũng khá chủ động trong bảo đảm an ninh từ xa
- Một số node nhỏ và rất nhỏ khác như: SenateGOP (đại diện của các Thượng nghị sĩ đảng Cộng hòa) hay GovAbbott (đại diện văn phòng thống đốc bang Texas)

➢ Cho thấy các diễn biến chính trị nội bộ Mỹ vẫn diễn ra, nhưng không quá "sôi động" trên mạng xã hội Twitter, trong khi các vấn đề quốc tế lại cực kỳ rầm rộ.

Visualize đồ thị

Nhận xét đồ thị: Cộng đồng Truyền thông

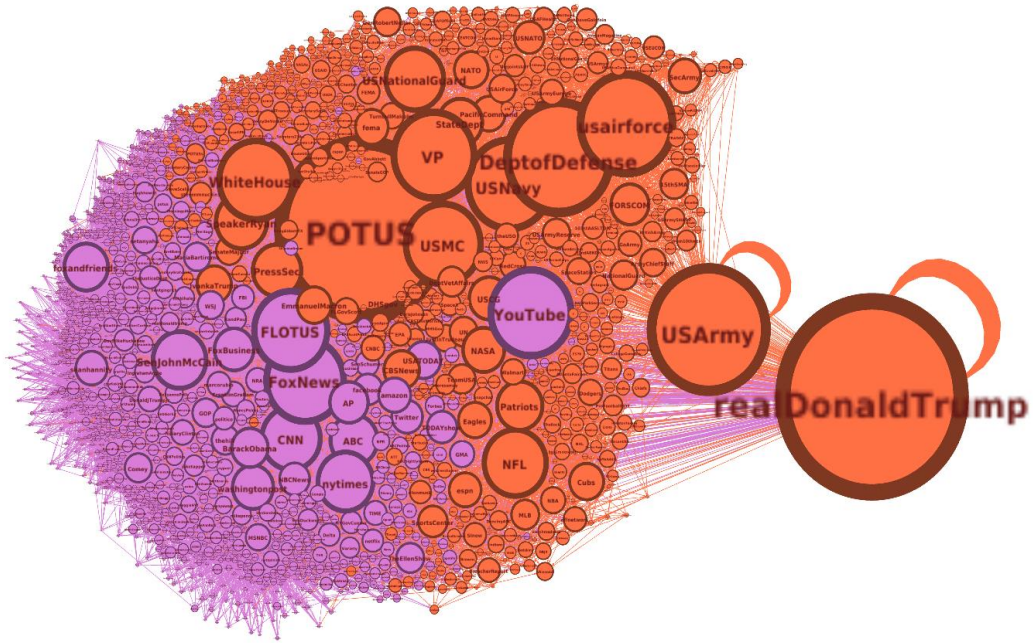


<u>Cộng đồng truyền thông</u> (báo chí, truyền hình, tin tức)

- Sở dĩ có sự xuất hiện của cộng đồng này là do họ là những bên đưa các tin về chính luận, quan điểm đường lối chính trị của Mỹ tới thế giới
- Các hãng thông tấn, tin tức nổi tiếng là các node lớn nhất trong cộng đồng này như: FoxNews (kênh tin tức truyền hình cáp số 1 nước Mỹ), CNN (mạng truyền hình cáp tại Mỹ), washingtonpost,…
- Một node khá lớn và đặc biệt là FLOTUS (đệ nhất phu nhân nước Mỹ). Đệ nhất phu nhân là người thường xuyên tiếp xúc báo chí, cung cấp các góc nhìn đa chiều từ vị thế là vợ của tổng thống

Visualize đồ thị

Nhận xét đồ thị



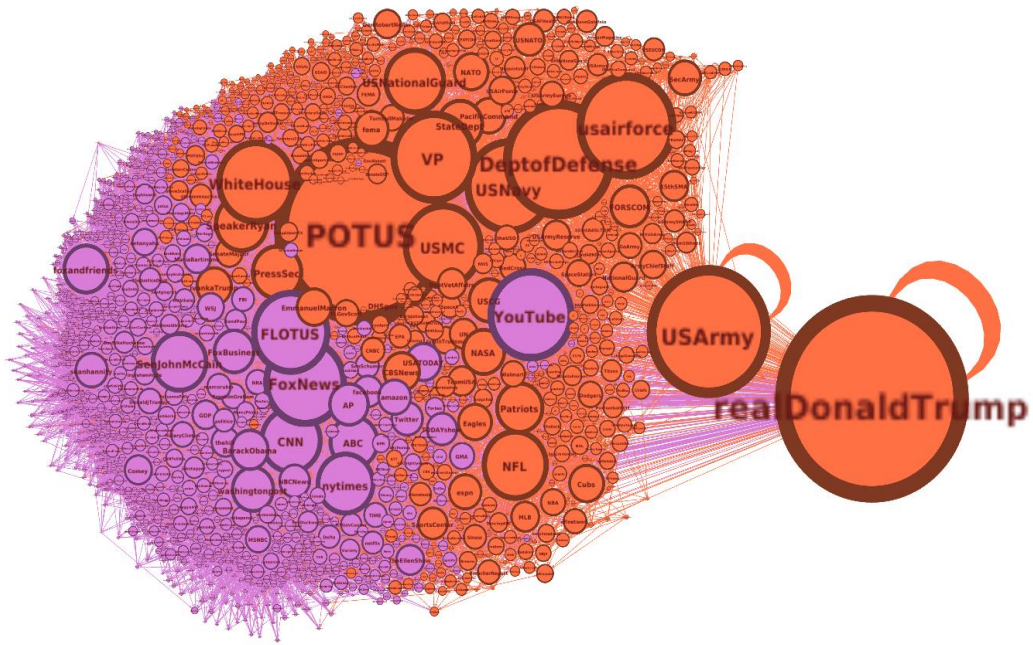
Phần giao thoa giữa 2 cộng đồng

- Giao giữa 2 miền là các node như:
- SpeakerRyan: người phát ngôn của Nhà Trắng
- PressSec: thư ký báo chí của nhà trắng
- HHSGov: Bộ Y tế và Dịch vụ Nhân sinh Hoa Kỳ
- IvankaTrump: trợ lý cho Tổng thống Hoa Kỳ Donald Trump, con gái của Donald Trump
- FLOTUS: đệ nhất phu nhân nước Mỹ

➢ Các node giao thoa là các node có liên quan tới cả 2 cộng đồng (vừa liên quan chính trị, vừa liên quan tới truyền thông)

Visualize đồ thị

Nhận xét đồ thị



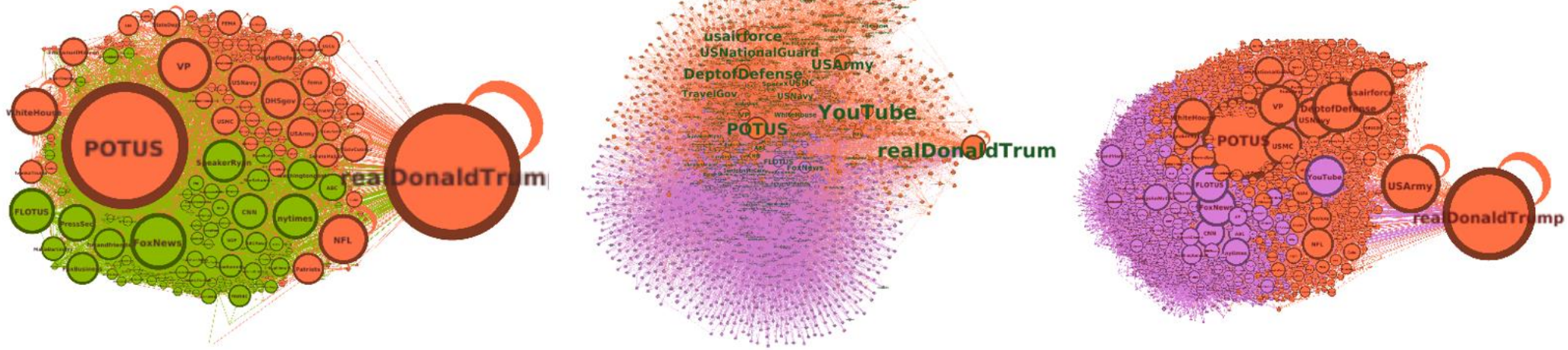
Các nodes không mong muốn

- Do sự mở rộng của đồ thị nên có lẫn 1 số tiểu cộng đồng về thể thao (espn, cubs, NBA, MLB, SportsCenter) và giải trí (TheEllenShow, netflix, Spotify, Snapchat, Disney, THR, ...)
- Đương nhiên giải trí, thể thao thì liên quan tới truyền thông, nhưng đó không phải là kết quả mong muốn của nhóm trong BTL này.

Visualize đồ thị

Lý do chọn độ lớn đồ thị

- Từ kết quả visualize và nhận xét đồ thị ta thấy 2 cộng đồng chính là ***cộng đồng hoạt động chính trị*** và ***cộng đồng hoạt động truyền thông***
- Hai cộng đồng xuất hiện rõ ràng và rất sớm ngay từ level 1 của đồ thị xuyên suốt đến các level sau.



Visualize đồ thị

Lý do chọn độ lớn đồ thị

- Tại level 2 xuất hiện một số node không mong muốn liên quan đến lĩnh vực thể thao và giải trí thay vì chính trị (espn, cubs, NBA, MLB, SportsCenter, TheEllenShow, netflix, Spotify, Snapchat, Disney, THR, ...)
- Nếu mở rộng thêm các level sau thì số node ngoài lề sẽ còn tăng lên nhiều (mặc dù đã có giải pháp sử dụng modun lọc tweet chính trị).

➢ Ta hoàn toàn có thể dừng tại level 2 vì:

- Ở tầng này đã thể hiện đầy đủ sự phân bố các cộng đồng cho ta cái nhìn tổng quan nhất về tình hình chính trị Mỹ.
- Đồng thời các node ngoài lề mới chỉ bắt đầu xuất hiện.

Thanks for watching!