Instagram Network Analysis

Social Network Analysis
Coursera

Marc Lester S. Tan

Introduction

Social networking has become one of the most important topics in recent times because of the several social networking sites that have now become part of our day-to-day living. There's a huge amount of information that we can gather from analyzing these social network.

In this document, I am going to analyze the social network based on wife's Instagram account. We will try to deduce information such as important people in her network and find groups of people or communities within her network.

Data Gathering

Instagram exposes the user's data through API. You can retrieve the list of users that my wife is following as well as the list of users who follows her. Basically, we need to get the list of her close friends, in which in my definition, are those users that my wife is following and also follow her. This relationship means these people are most likely close to my wife or she personally knows her, probably in work, school or some community.

I decided to use node.js to retrieve data and generate the GML as it is very easy to prototype using a dynamic language such as Javascript.

Retrieving all the nodes and egdes

My wife follows 284 users as of this writing and 46 of them follows her back. Each of these 46 users follows another set of Instagram users.

Based on my calculation, there will be around 10000+ nodes, which are going to require us to perform 10000 API calls. Unfortunately, Instagram has a limit of 5000 API calls only per hour for each client so I had to divide the operation into several steps and caching the downloaded data locally before doing another set of API calls.

- Retrieve the list of users my wife follows and also retrieve the list of users they follow.
- Filter the list of users my wife follows so that only those who follow her back remains.
- Generate the nodes and edges for the remaining list of users. Each node represents a particular user and edge represents a user following another user

These steps have to be done in 1 hour separation to prevent reaching the API call limit. Before you start the script, make sure to replace the ACCESS_TOKEN value in constants.js. This can be taken from http://instagram.com/developer/ console.

Retrieve the list of users

We'll start by running our express-js application:

$ node app.js

Then start the process by opening http://localhost:3000?action=start in your browser.

This will start retrieving the people being followed by my wife as well as the people that they are following. It will take almost half an hour up to 1 hour to finish. After that, generate the huge json file by loading http://localhost:3000/ into your browser. Save this file locally as /Users/i071571/Desktop/nodes.json or change it as necessary depending on your OS.

Filter close friends

Next, run the script called filterCloseFriends.js by running:

$ node filterCloseFriends.js

This script assumes that you have the nodes.json present in your desktop folder. It will generate a new file called /Users/i071571/Desktop/closefriends.txt.

Parse nodes and generate GML

Last step is to parse and create our nodes and then generate the Graph ML that can be loaded into Gephi. Run the script by executing:

$ node parseNodes.js

This script assumes you have both nodes.json and closefriends.txt in your desktop and will generate a file called /Users/i071571/Desktop/output.graphml.

Once we have the GraphML file, we can now feed this into Gephi for further analysis.

A Detailed Look

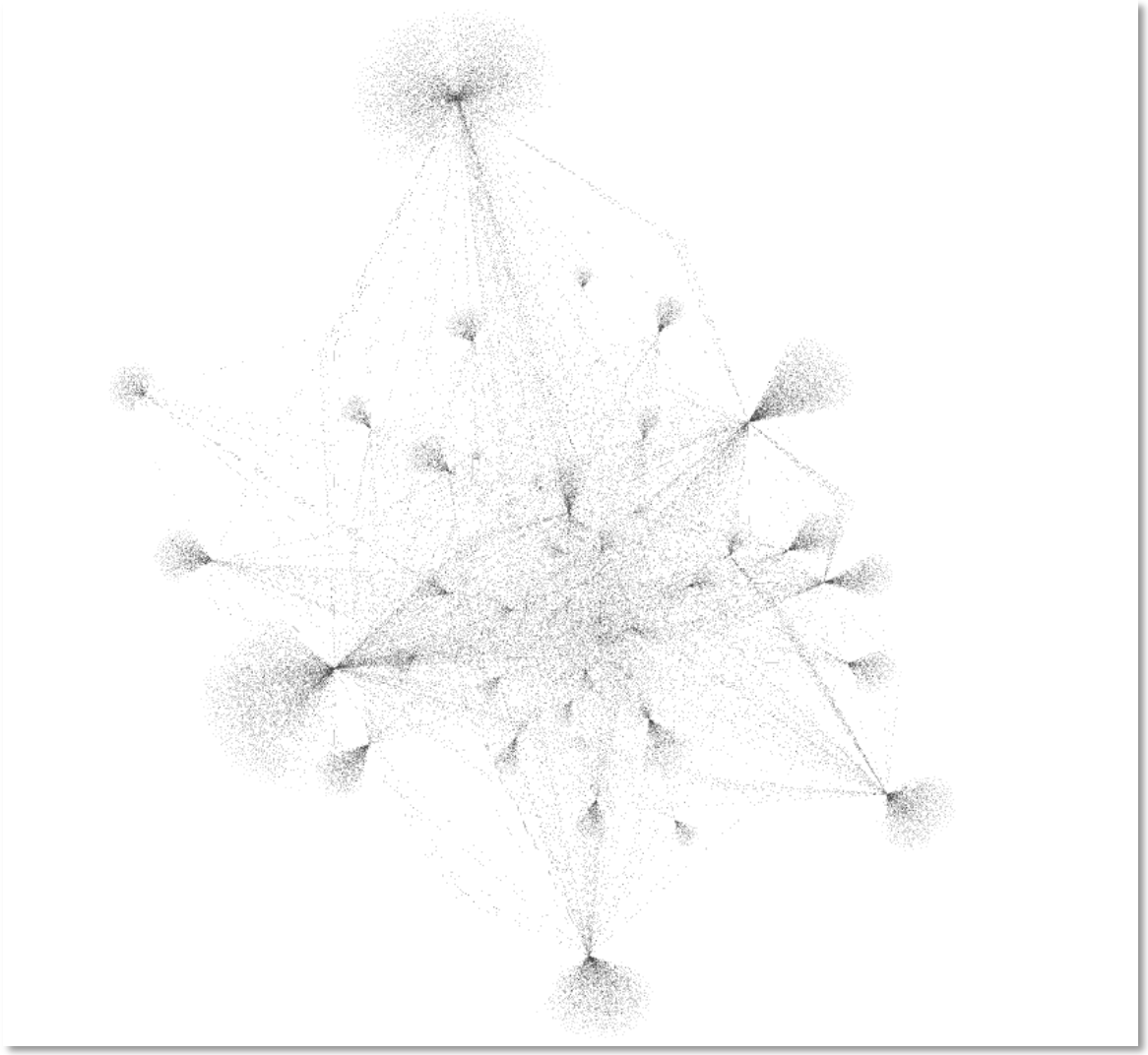
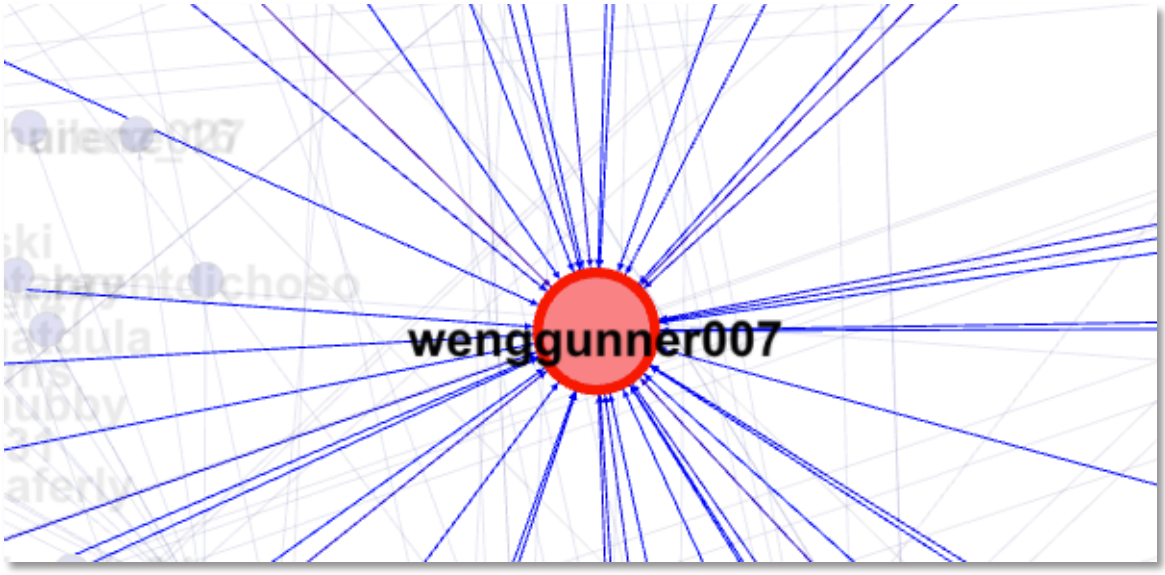


Figure 1 Instagram Network Graph

Degree Centrality

There are 8596 with 10586 edges. I've used the ForceAtlas 2 layout with Dissuade Hubs and Prevent Overlap both checked. I then run an Average Degree analysis that gives an average degree of 1.232. We can use this information to identify who is being followed a lot by my wife's Instagram friends.



Label	In-Degree
wenggunner007	45
annecurtissmith	27
iamangelicap	17
beaalonzo	17
xtina_ontherocks	17

Out of 45 close friends, 60% of them follow annecurtissmith followed by another actress in the Philippines. Actually, the top 7 most followed users are all actresses, as you would expect from a female Instagram user.

Closeness Centrality

To measure closeness centrality, we can use the Average Path Length algorithm. This will give us the average path length to be 2.851. Closeness centrality can be used to identify the people who can easily dissiminate information to other nodes. They are mostly sitting in the center of the network graph in which they have almost direct or short path to other nodes in the network.

Label	Closeness...
iamsirbong	2.994
lyn_aeon	2.991
corkmissy	2.987
elynormacaranas	2.986
tan_11	2.971
diane_mamaril	2.97
ronanfab	2.959
misshosey	2.958
liamgab	2.951
bongdelacruz	2.949

So if my wife wants to distribute a very important photo for relief information or emergency numbers for the Typhoon Haiyan, she can tell iamsirbong or lyn_aeon to re-gram (retweet in Twitter) it.

Betweenness Centrality

Betweenness centrality can help us find the people who can connect us to a much larger network. These individuals often times doesn't have shortest path to everyone else but the fact that they can connect us to a large network can help us either dissiminate information or connect to another individual.

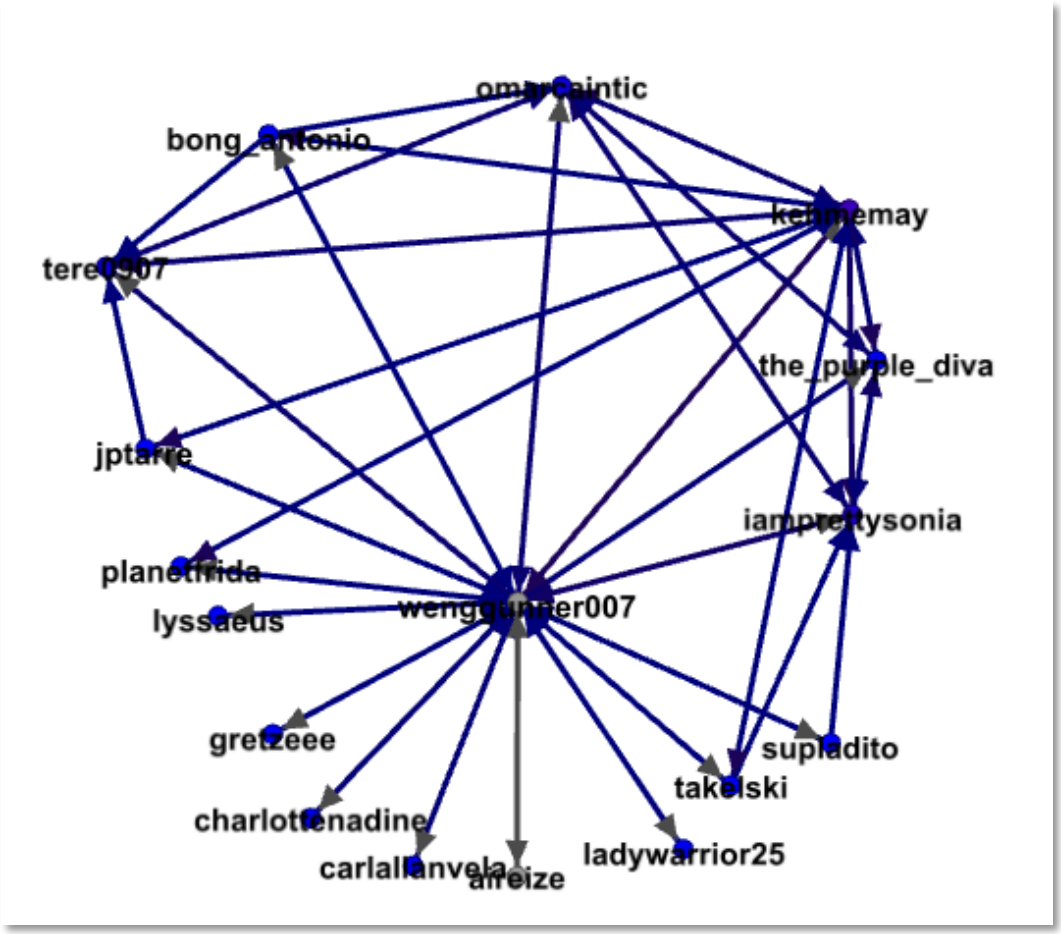
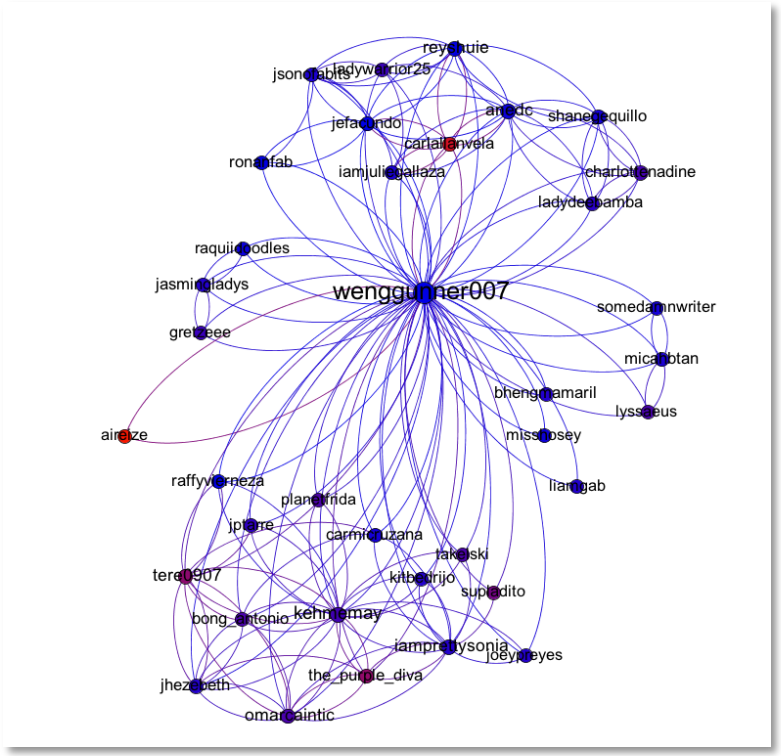


Figure 2 Nodes with more than 10000 Betweenness Centrality

Label	Betweenness Ce...
wenggunner007	324,946.949
aireize	42,408.459
carlallanvela	30,189.77
the_purple_diva	24,288.064
kehmemay	22,774.68
tere0907	20,699.002
supladito	19,268.645
omarcaintic	13,645.903
lyssaeus	13,396.29
charlottenadine	12,213.996

The first item is my wife's Instagram account. She has high betweenness because we filtered the users so that we will only retain those who also follow her back. This makes her at the center of the graph and directly connected to all other nodes which in turn are connected to the people they follow

Communities



We can use betweenness centrality to filter the nodes so that only those close friends are displayed. In the image above, we can identify groups or communities based on their position in the graph.

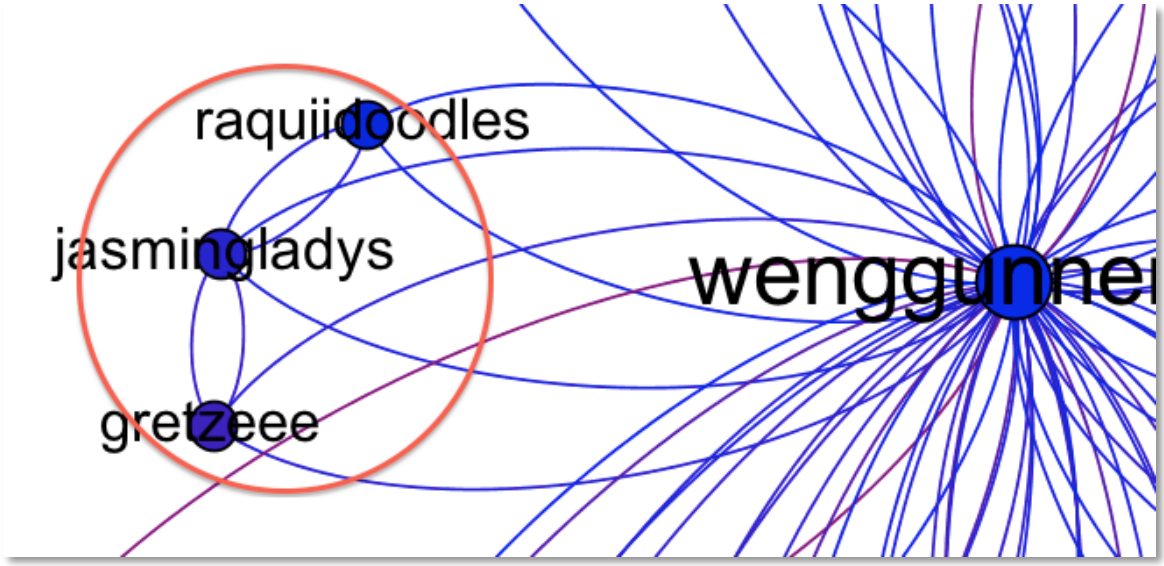
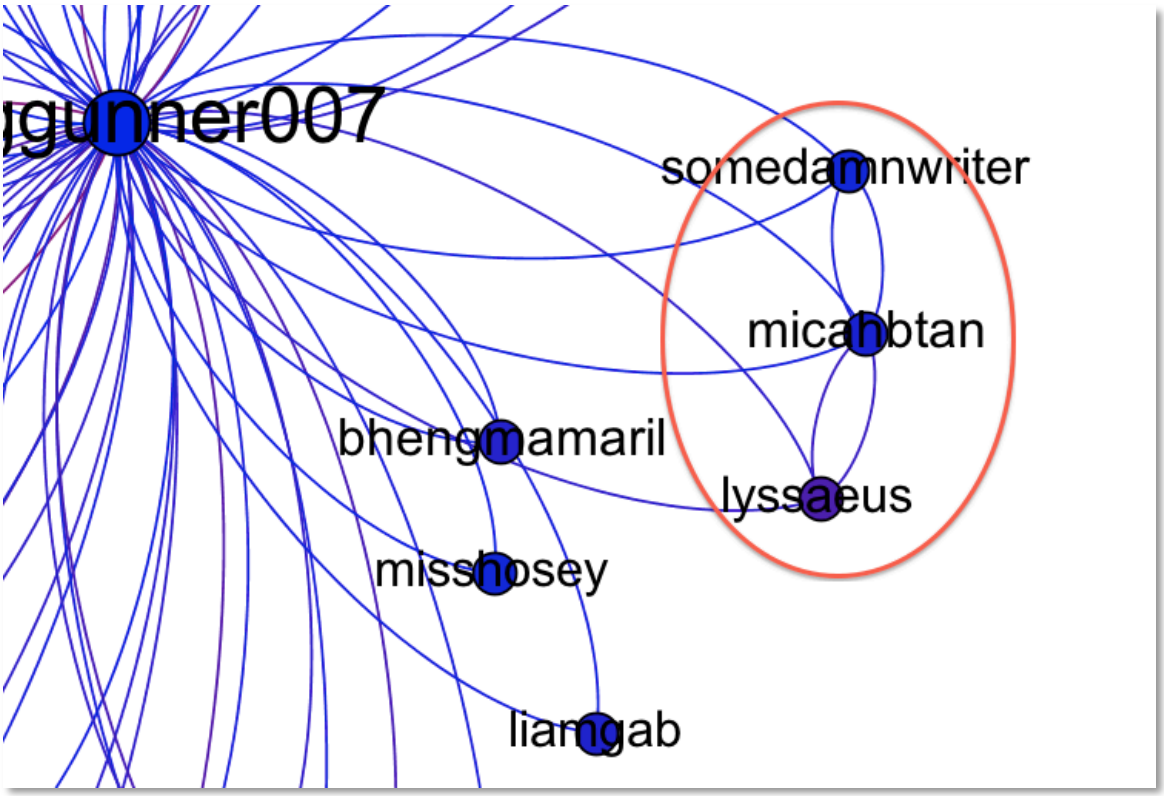
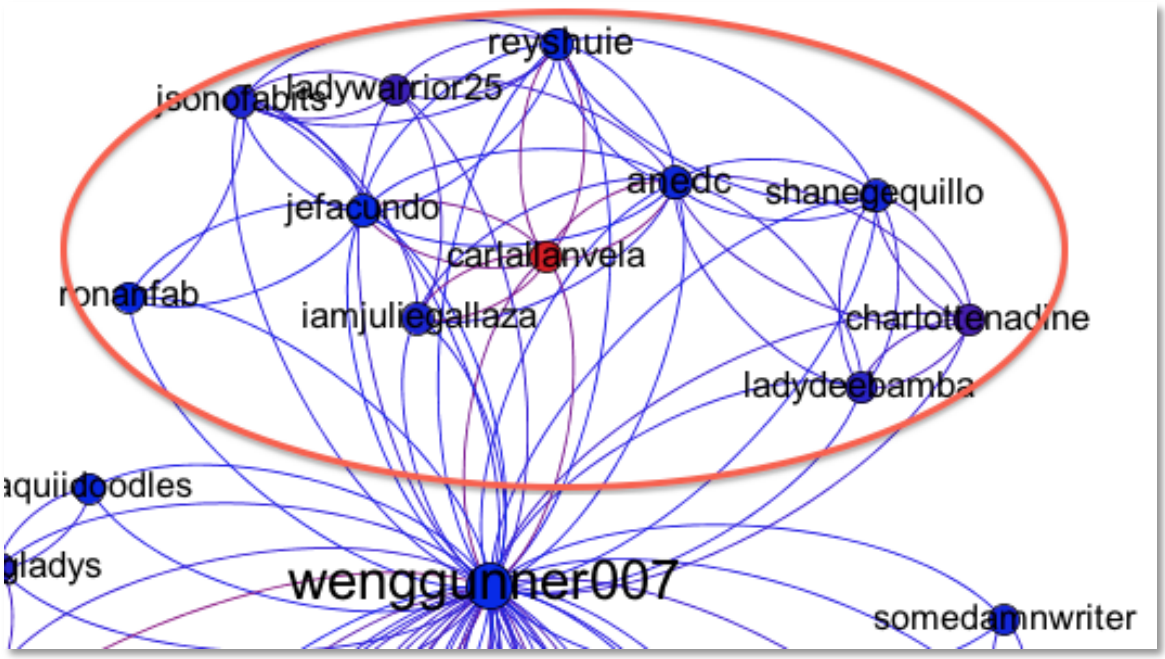


Figure 3 Friends from AIG

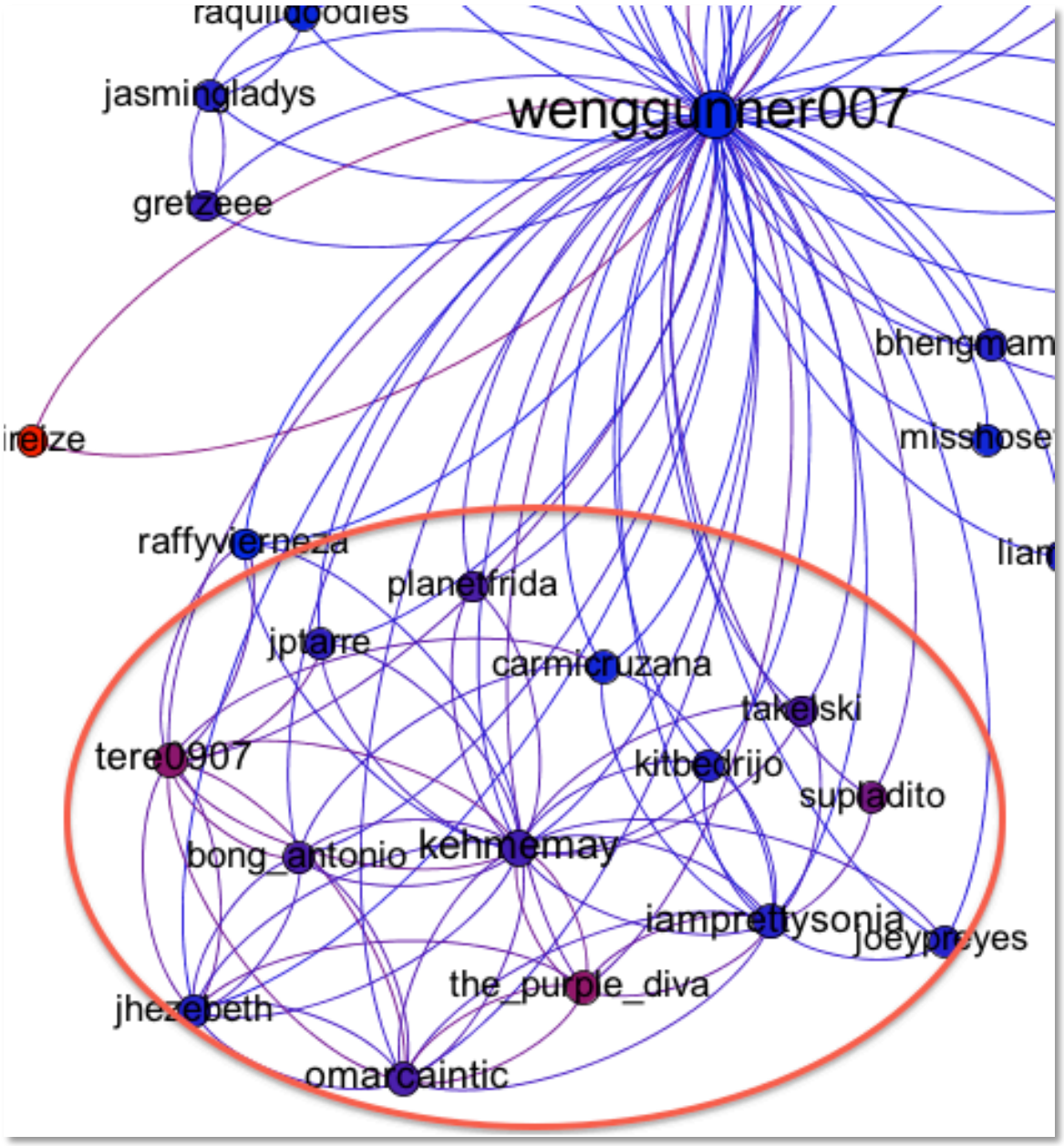
The image above shows 3 people from AIG, my wife's previous company.



In this image, it shows my cousin, sister and brother's girlfriend all positioned near to each other.

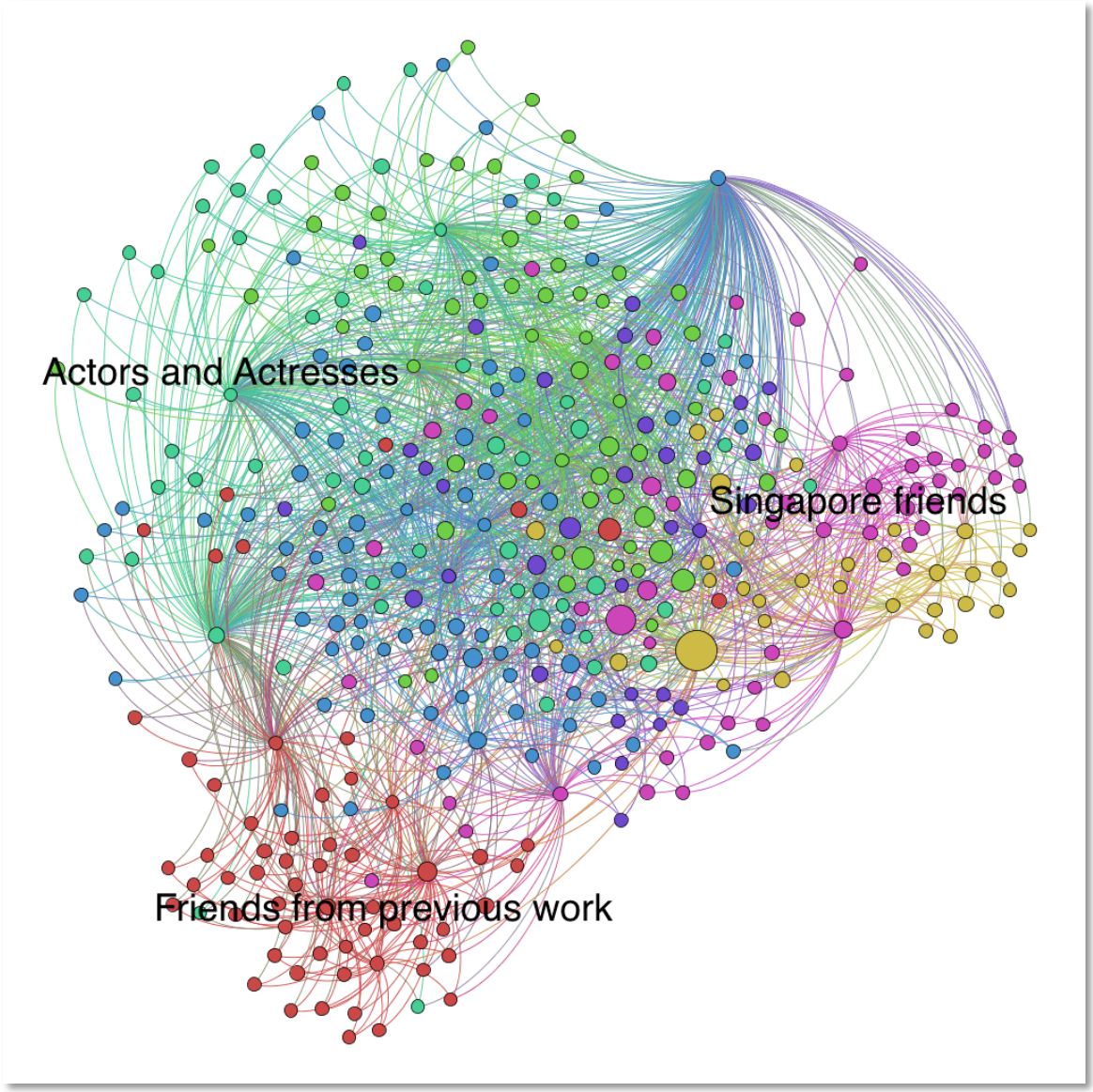


These people are mostly from previous company in the Philippines but some of them are from my wife's new company in Singapore, which happened to have attended the same school as the others in the group.



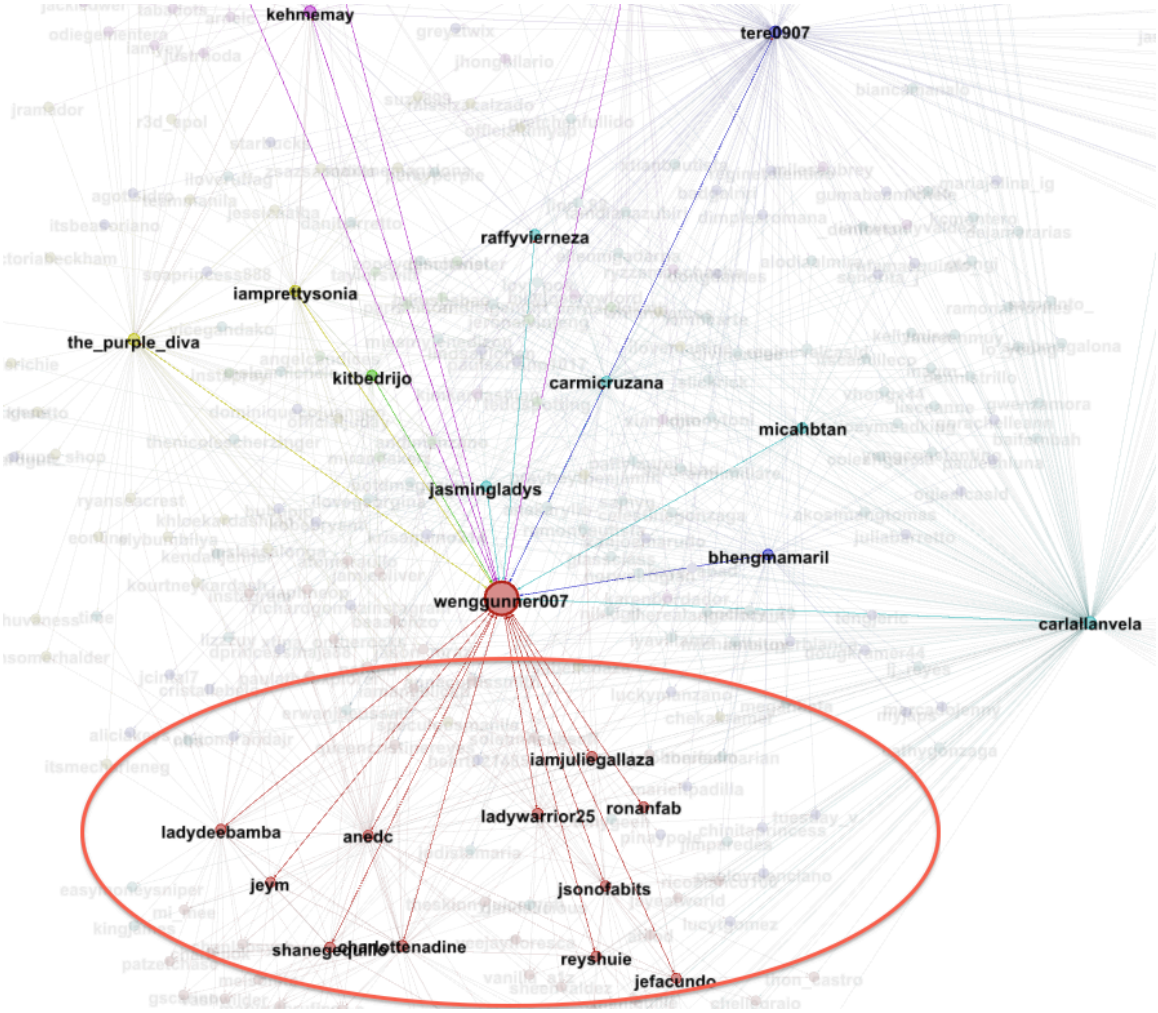
Lastly, this one shows the group of friends where my wife started her first job.

Now, we'll try to identify the much larger communities using the Modularity algorithm. I have filtered the nodes to show only those who have at least three in-degrees to reduce the clutter and make sure that the nodes are someone whom people really know and not just simply following. Here's an overall graph that shows communities represented by different colors.



Finding Relevant People

Let's try to take a look at the bigger picture.



The circled part represents my wife's friends at her previous job in Makati Philippines but if you'll look closely:

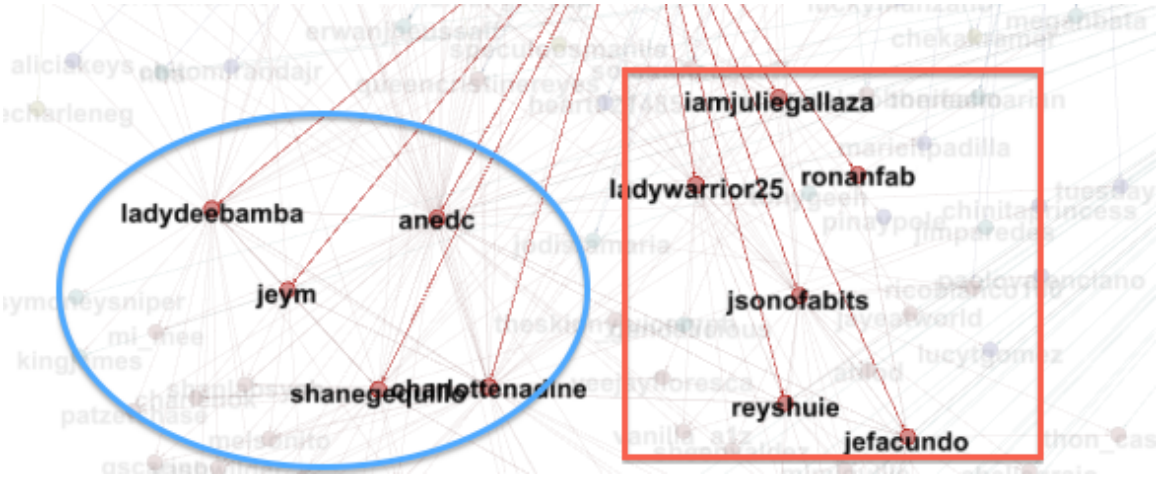


Figure 4 Friends in Makati

The ones inside the blue circles are a mix of those who went to the same company as the ones inside the red square and some Singapore friends. The reason why they are there as well is because they went to the same university as those other guys in the blue circle.

Now let's pick iamjuliegallaza and analyze her neighbors.

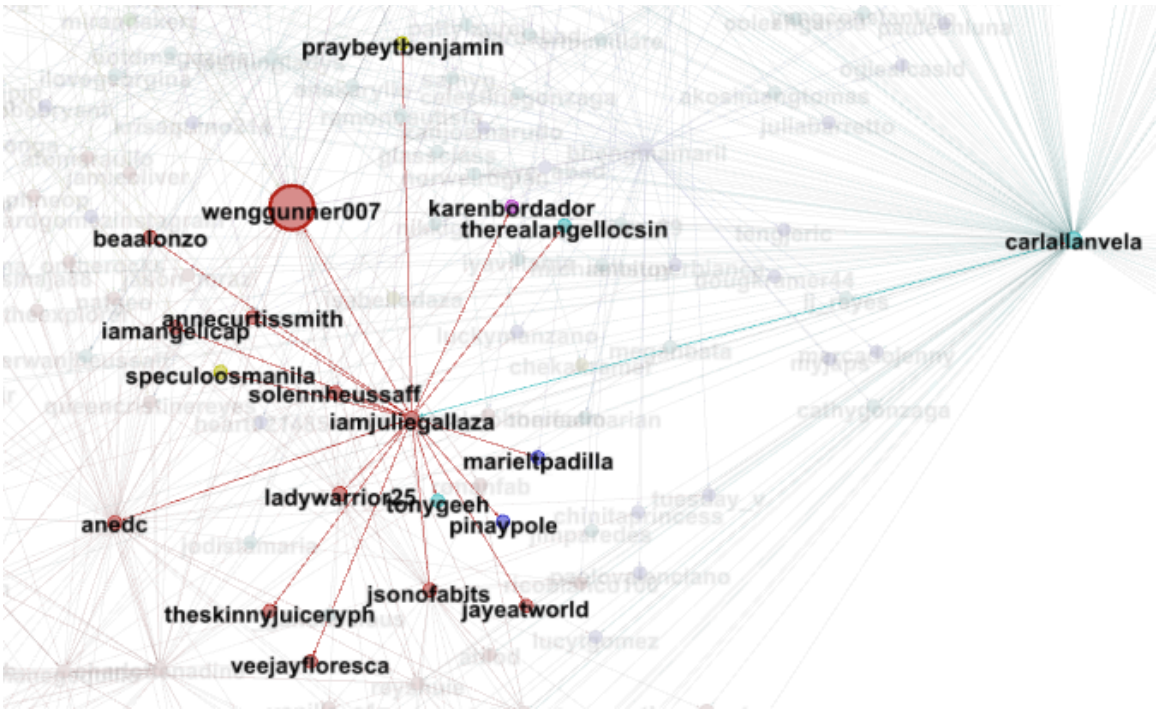


Figure 5 iamjuliegallaza's network

There are some people who are not directly connected to my wife but connected to at least one or two to her close friends.

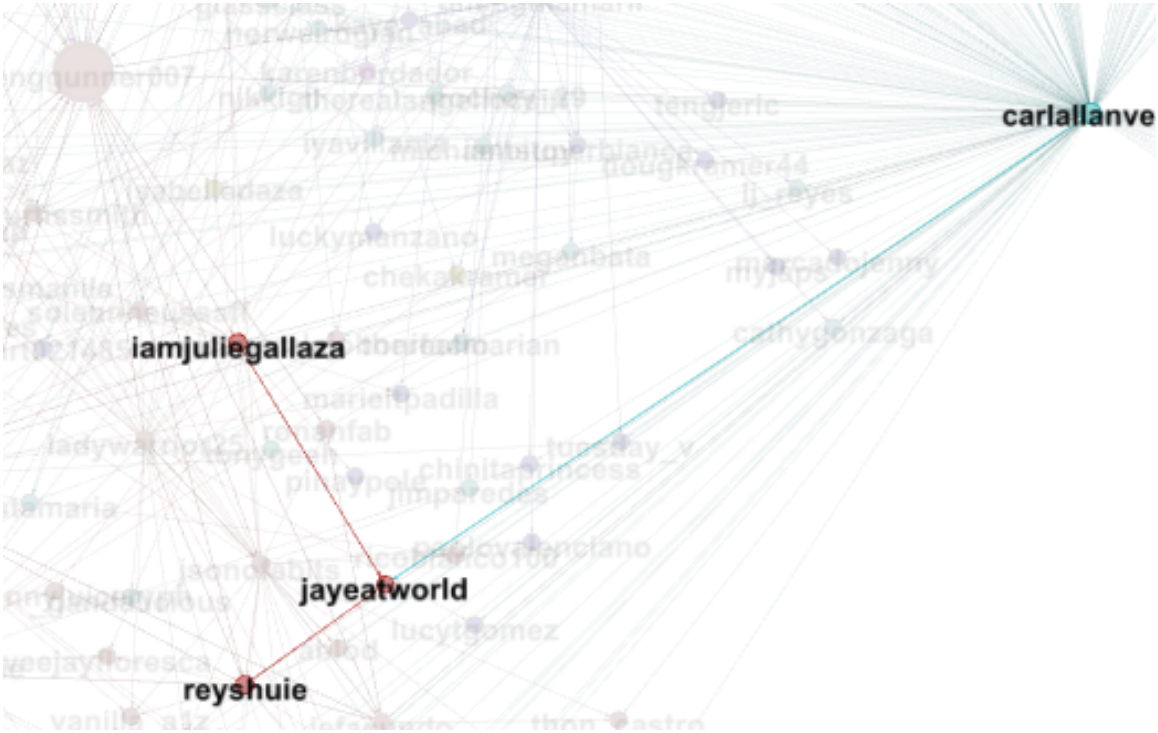


Figure 6 jayeatworld's network

jayeatworld is connected to iamjuliegallaza, reshuie and carlallanvela which are all connected to my wife. All of these three users came from my wife's previous company so there's a chance that my wife also knows jayeatworld probably from the same company.

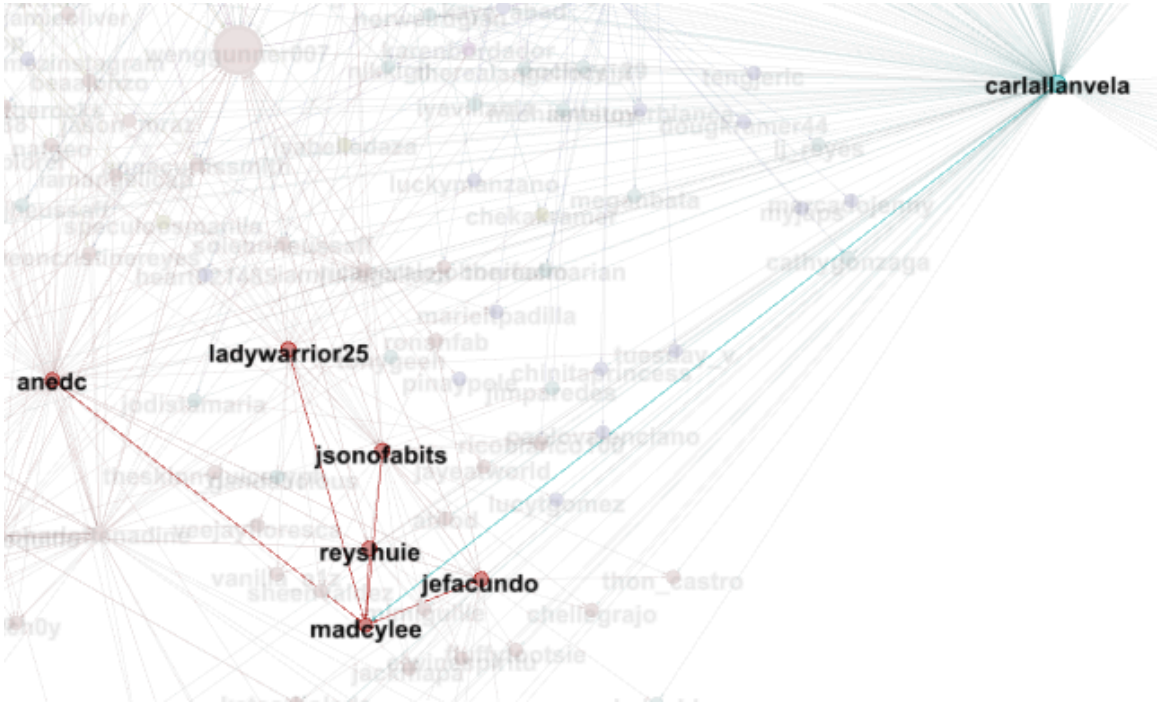


Figure 7 madcylee's network

Here, all the nodes that madcylee is connected to are all close friends of my wife giving her a higher chance that she knows my wife and vice versa.

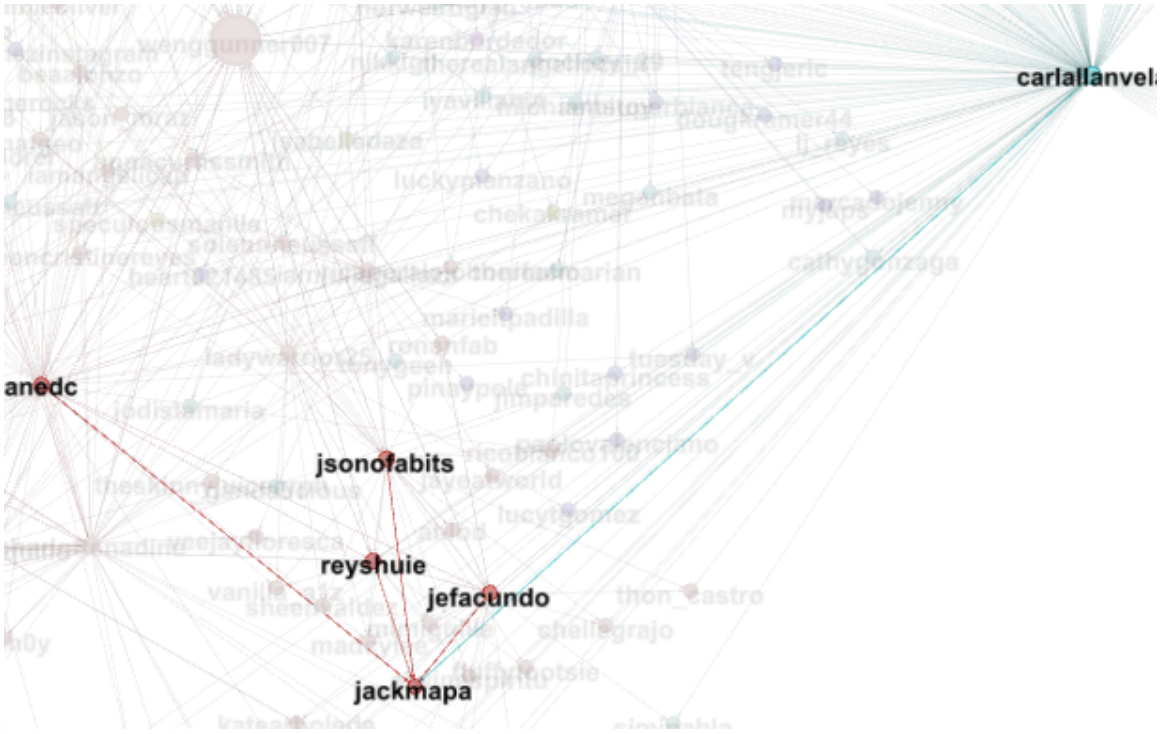


Figure 8 Jackmapa's network

jackmapa is also connected to all close friends of my wife.

Let's take a look from another angle:

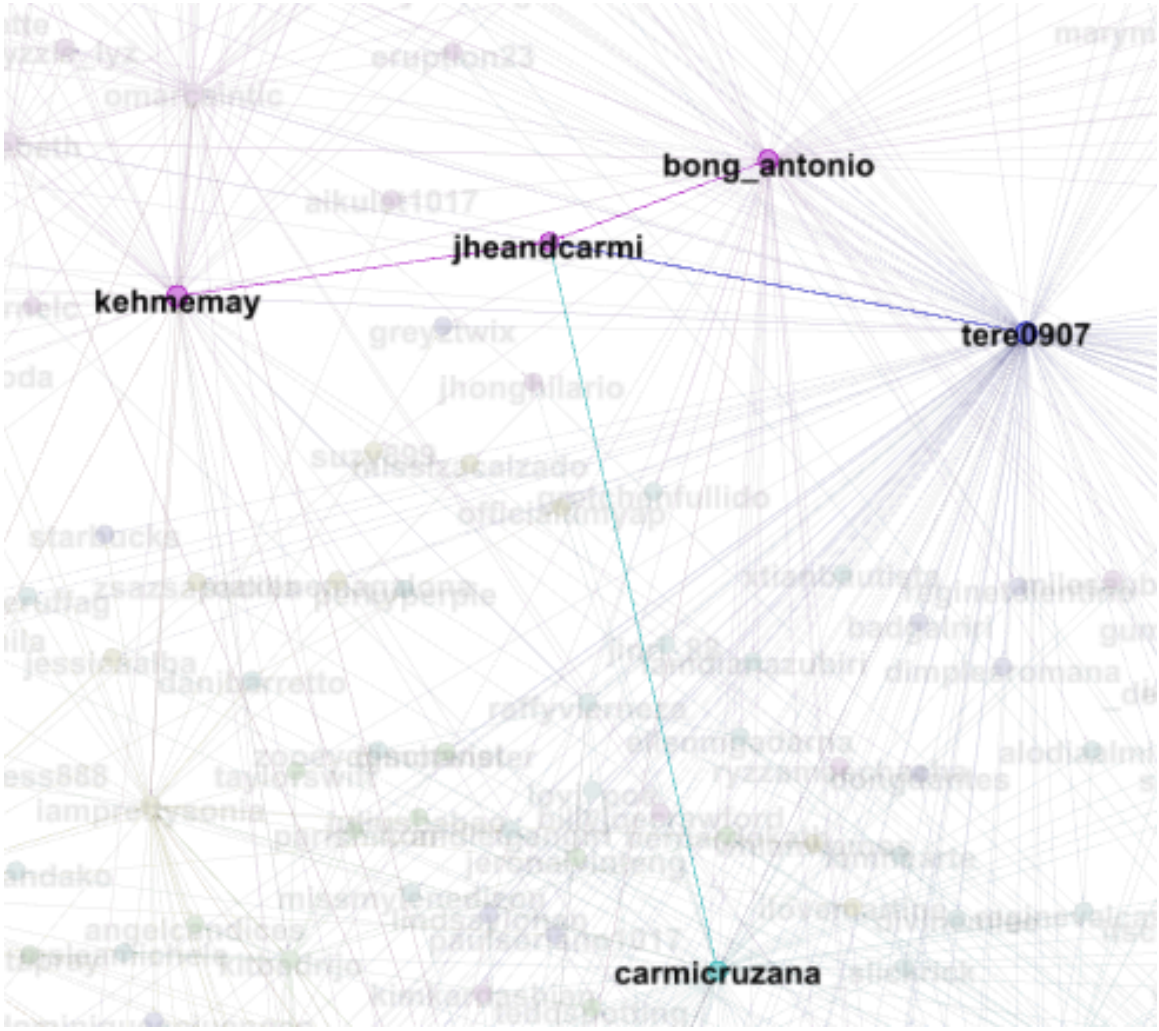


Figure 9 Jheandcarmi's network

jheandcarmi is connected also to close friends specially to kehmemay who is connected to several close friends.

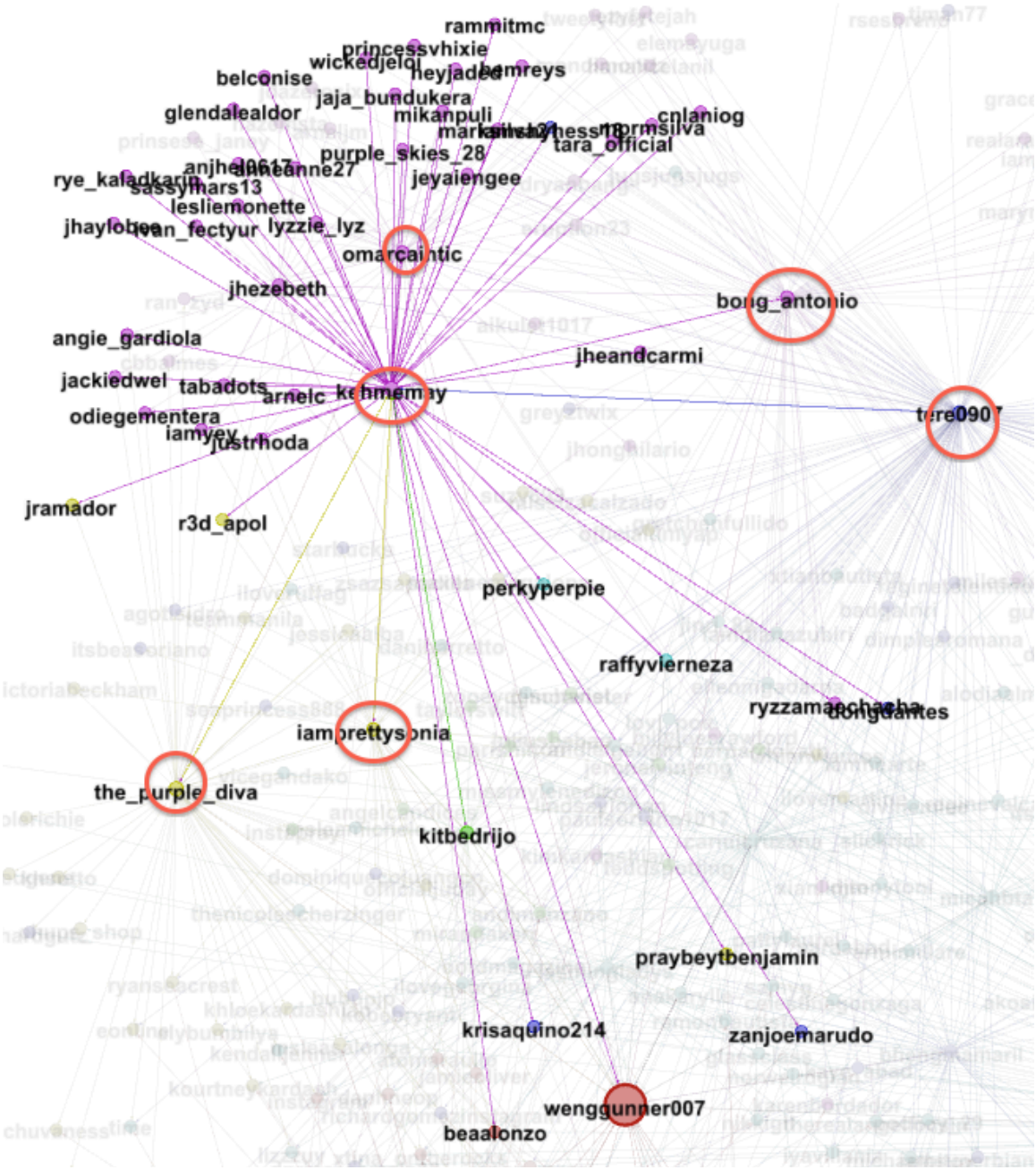


Figure 10 Kehmemay connected to several close friends

Conclusion

There are so many things that you can discover using Social network analysis. You can find groups or communities which are not clearly visible if you are just reading your Instagram stream or by just looking at the tabular data.

I think social network analysis is also useful not only in social networking sites but it can also be used for decision making and prediction when used in business with large amount of data.