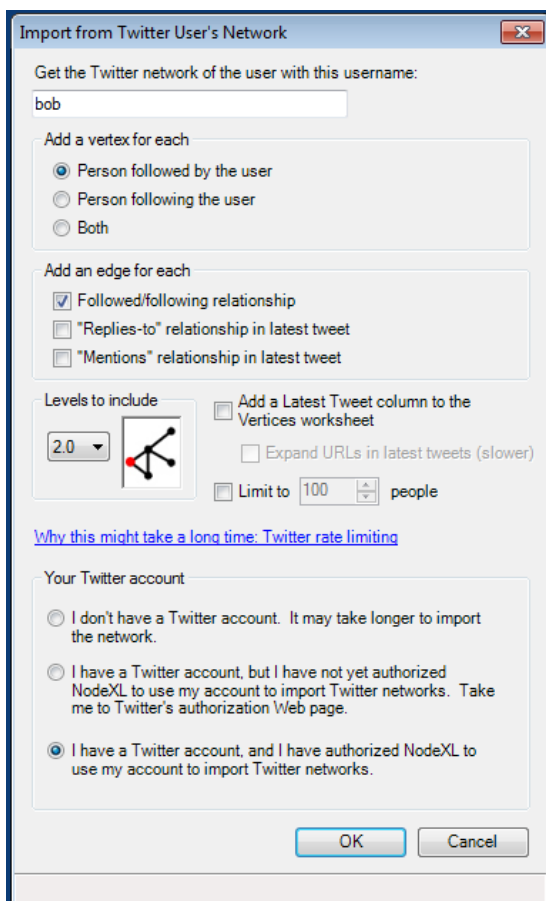


## EMPIRICAL NETWORK ANALYSIS PROJECT

We extract the social media data of a Twitter account, creating a directed graph of the relationships of friends and friends of friends (FOF). This graph is then analyzed for community structure. The question we want to answer is whether this structure can be used to reflect the interests of the user. This structure could be used as a recommendation application where the interests of your friends and FOFs could be used to present you with music, videos, news, etc. that you might find relevant and interesting. A personalized electronic newspaper, [The Tweeted Times](#), uses this approach.

### Data

We use NodeXL<sup>1</sup>, a tool to explore social network structure using Excel, to extract my Twitter user graph. The settings I used are:



Import from Twitter User's Network

Get the Twitter network of the user with this username:  
bob

Add a vertex for each

- ☒ Person followed by the user
- ☐ Person following the user
- ☐ Both

Add an edge for each

- ☒ Followed/following relationship
- ☐ "Replies-to" relationship in latest tweet
- ☐ "Mentions" relationship in latest tweet

Levels to include  
2.0

☐ Add a Latest Tweet column to the Vertices worksheet

☐ Expand URLs in latest tweets (slower)

☐ Limit to 100 people

[Why this might take a long time: Twitter rate limiting](#)

Your Twitter account

- ☐ I don't have a Twitter account. It may take longer to import the network.
- ☐ I have a Twitter account, but I have not yet authorized NodeXL to use my account to import Twitter networks. Take me to Twitter's authorization Web page.
- ☒ I have a Twitter account, and I have authorized NodeXL to use my account to import Twitter networks.

OK Cancel

<sup>1</sup> NodeXL <http://nodexl.codeplex.com/>

The main thing here is that I chose to import up to level 2.0 of my user network to pick up friends and FOFs. Since Twitter now imposes rate limiting for API requests, NodeXL will automatically pause to keep the job running within the Twitter guidelines. This process took many hours to complete and is prone to errors. In addition, I set the maximum people to 300 which certainly covered all my followers but some of the people I follow certainly exceeded that parameter. It is easy to see that when you begin to crawl your FOFs, the graph will grow exponentially so you need some judicious choices when you plan this extraction.

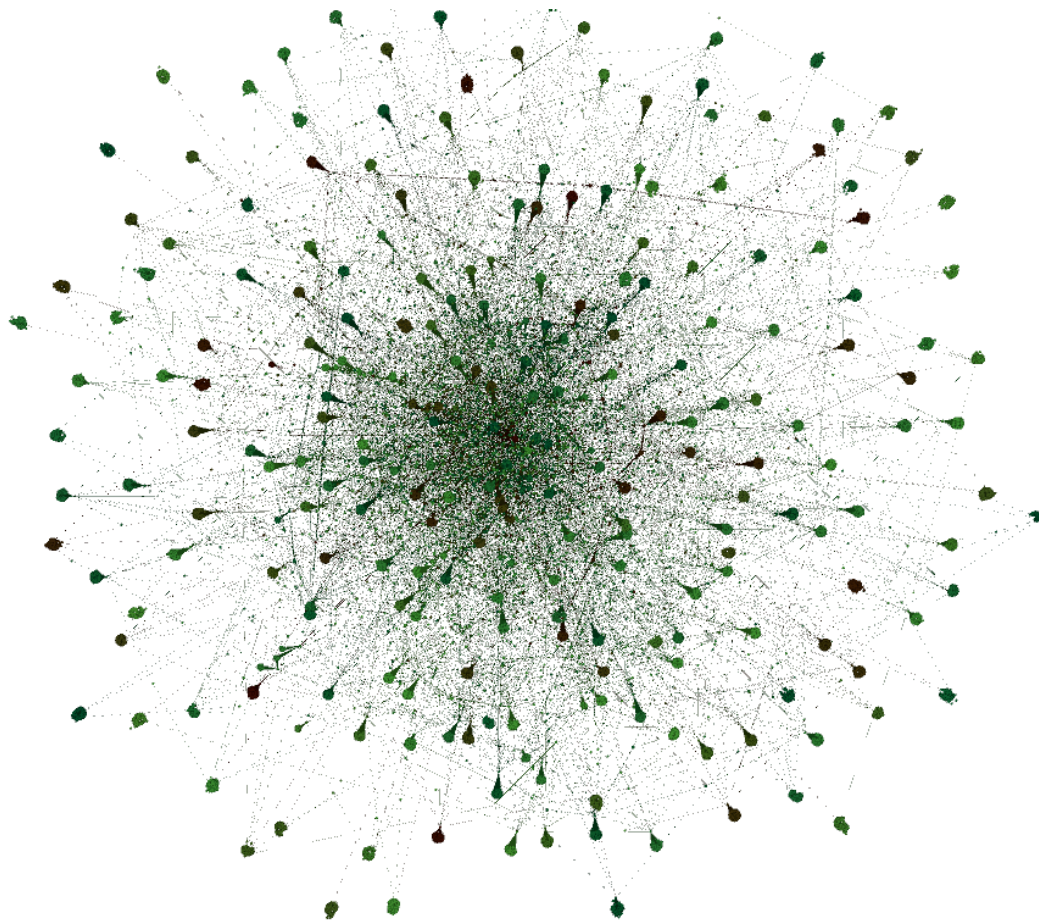
Finally, the excel spreadsheet was scrubbed of URL, location, and most identifying data. In my case, 3 private users were removed and all the rest were left in since they were public. The last step is to export the data to a GraphML file to be used for analysis with Gephi.

## Analysis

We use Gephi to analyze the data. The result is a directed graph with 65,628 nodes and 71,082 edges. We ran the following metrics:

Metrics	Result	Description
Average Weighted Degree	1.083	
Modularity	0.92	142 communities

Modularity was used to color the 142 communities. I applied a ForcedAtlas2 layout and the result shows a clear radial structure around the primary user and a “ice-cream cone” effect of some of communities discovered.





Zooming in gives a better picture of this detail which will be explained in the next section



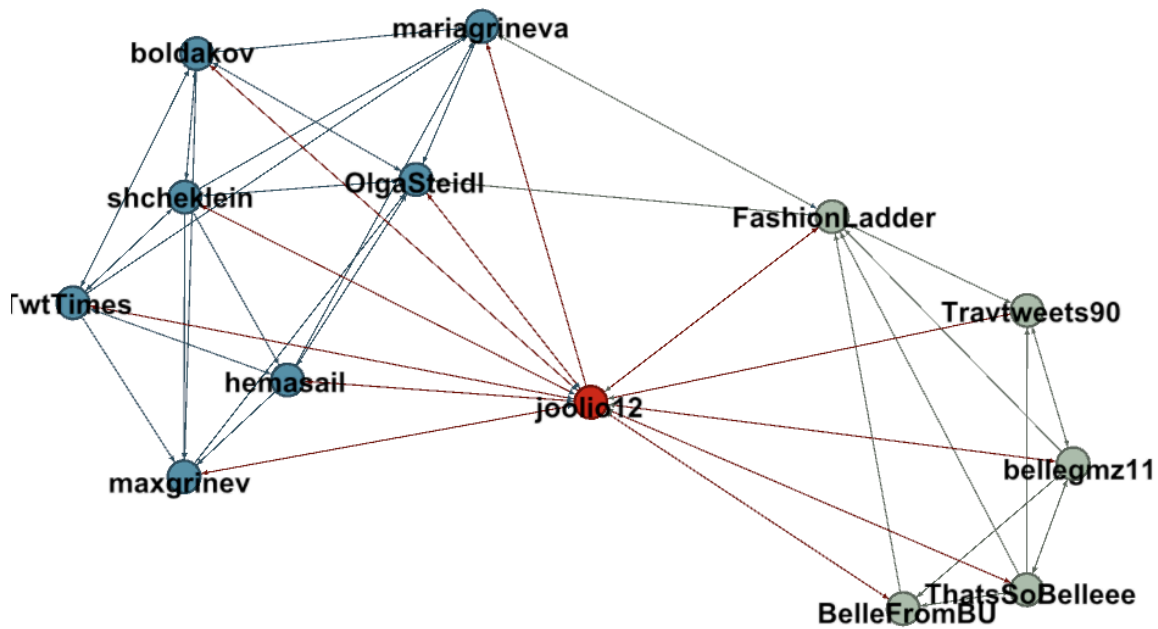
## Results

I used Gephi filters to understand some of the structures displayed. The snow-cones are hubs of FOFs. In other words, if I have a friend that has many friends that I don't follow directly, the edges will aggregate around the friend but not around me. I will get one edge from my friend hence the snow-cone. In many cases I found that it was unlikely that this community might offer any interest to me, especially if it was a friend that is not really that close. In Twitter a friend might be a follower or a person I follow but in some cases, this is one way only (no follow back). This model might be improved by excluding these relationships since it might be more meaningful to keep only those relationships that are mutual. I don't fully agree since there are some people I follow that may not follow back, like the actress Anna Silk, but I still find very interesting.

The next step was to use the K-Core filter to identify communities. K-Core was defined in class as a maximal subnetwork in which each vertex has at least degree  $k$ . In my graph, there

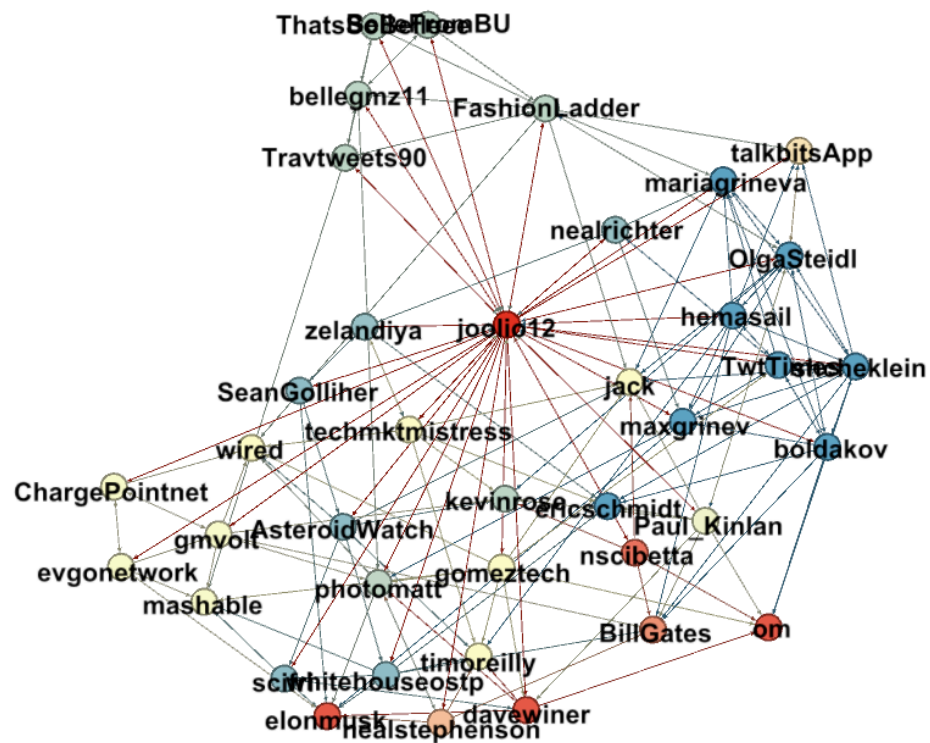
were no communities with a k-core  $> 7$ . I decided to look at the k-core = 7, which intuitively is the largest k-core in my graph and was extremely surprised to find this result

### **K-Core = 7**



This community is striking because it represents two very important sets of users, close family on the right and a strong work relationship on the left. In this case using k-core 7 and modularity would be a strong way to provide recommendation since these two groups are some of my most important and relevant followers.

**K-Core = 6**



In this case, some additional nodes begin to appear but the communities defined by the modularity classes has already lost some of the relevance by the introduction of some of the FOFs. Bill Gates, for example, is followed by many of the members of my work community but it clearly is not as relevant.

## Conclusion

K-core and modularity analysis can reveal significant community information on a users Twitter graph of friends and friends of friends. The effectiveness of k-core analysis in my data opens the question of whether using the maximal k-core of a Twitter graph can reveal the most important communities for a given user. A further area of study.

This paper and the Gephi data are available at [Github](#)

I want to thank Coursera and professor Lada Adamic for this great course.