Finding the Influencers in an Anorexia Twitter Network

Nicholas Branco
1 May 2018

1 Executive Summary

Twitter is home to many communities, many of which are good healthy communities, but some are not. In this report I take a look at two Twitter hashtag networks dealing with anorexia. The hashtags are #ProAna and #ThinSpo. Through running some metrics on these two networks, it becomes apparent who the influencers of the two networks are.

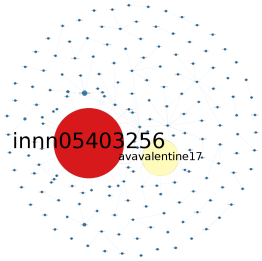


Figure 1.1: The #ProAna network by Betweenness Centrality.

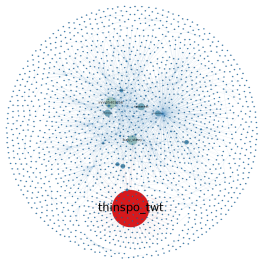


Figure 1.2: The #ThinSpo network by Betweenness Centrality.

From my analysis, I have found the following accounts to be prevalent members of the two networks:

innn05403256

thin_dreams5

thinspo_twt

wowsothin

I believe the most important node to be innn05403256 as it is very important in the #ProAna network and fairly important in the #ThinSpo network.

2 Introduction

Whether we like it or not, Instagram models and Youtubers are influencers in our society. That essentially means that they have an impact on our lives. In a world where social media is prevalent wherever you go (how many people ask you to like their Facebook page and follow their Twitter account?), people who are important on social media become important in the lives of those who are connected to them directly or indirectly. As someone who spends his free time listening to Youtube channels and podcasts, I've noticed that the communities surrounding these people tend to be pretty positive. When a Youtuber or someone close to one is in trouble, the fan base pitches in to help. I've subscribed to multiple Youtubers who have attempted suicide and yet are still around because of the support they have received. It can truly be wonderful.

And yet, it isn't always. Just as there are people rallying around those with depression to help them get through it, there are those who don't see things that way. If you've spent any amount of time on the internet, you would know that it's not all sunshine and rainbows. People are messed up, and they tend to group together with others who think similarly. This project looks at two of those networks: Twitter networks made up of people who are anorexic.

The purpose of this is to find out who in these networks are influencers. Identifying the influencers would be step one in being able to take the network apart. By removing influencers or changing their behavior the network changes. These networks are made up of people who encourage each other to be anorexic, to continue to keep losing weight until they're too weak to move. Being able to dissolve the network would make it easier for the users to find help with their anorexia. The community (or more likely one of the few) that they were apart of would no longer be a viable way to find "encouragement" to stay the way they are.

A word of warning: there will be user accounts mentioned in this report. Looking through these accounts or the hashtags I am studying is a very depressing experience. I would advise you not to look these people up on Twitter.

3 Methodology

The first step in any analysis is to get data. There are many ways we can go about getting Twitter data through Python or R, but those won't really help us here. What we want is a network, preferably one we don't have to build ourselves as that could take a long time. Dr. Jason Shulman of Stockton University introduced me to a program called NodeXL which could scrape Twitter for me and build the network I want to look at. We put in the hashtags #ProAna and #ThinSpo as those are prominent tags for anorexia posts. NodeXL built the two networks (and ran some metrics as well) and output them to a GraphML file.

From there I examined both networks simultaneously using the Python package Networkx. I looked at the following five metrics to examine which user

accounts (nodes) are the most important:

1. Betweenness Centrality
2. In-Degree
3. Out-Degree
4. Page Rank
5. Katz Centrality

I also used the program Gephi to help visualize the networks as it is vastly superior (in my opinion) to Networkx in drawing graphs. I have the two networks displayed below.

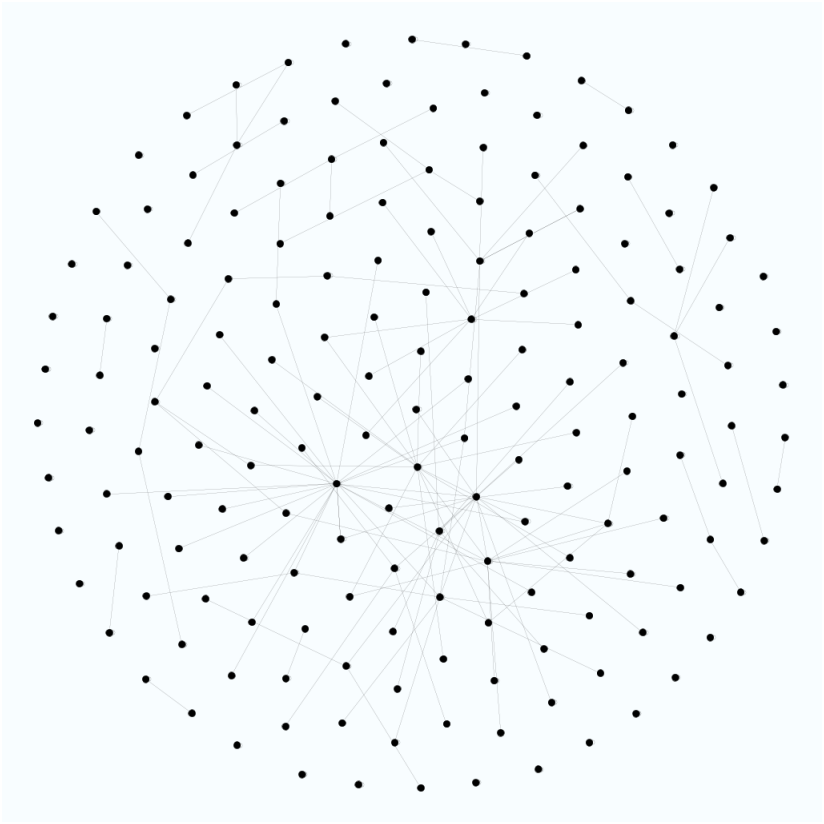


Figure 3.1: The #ProAna network.

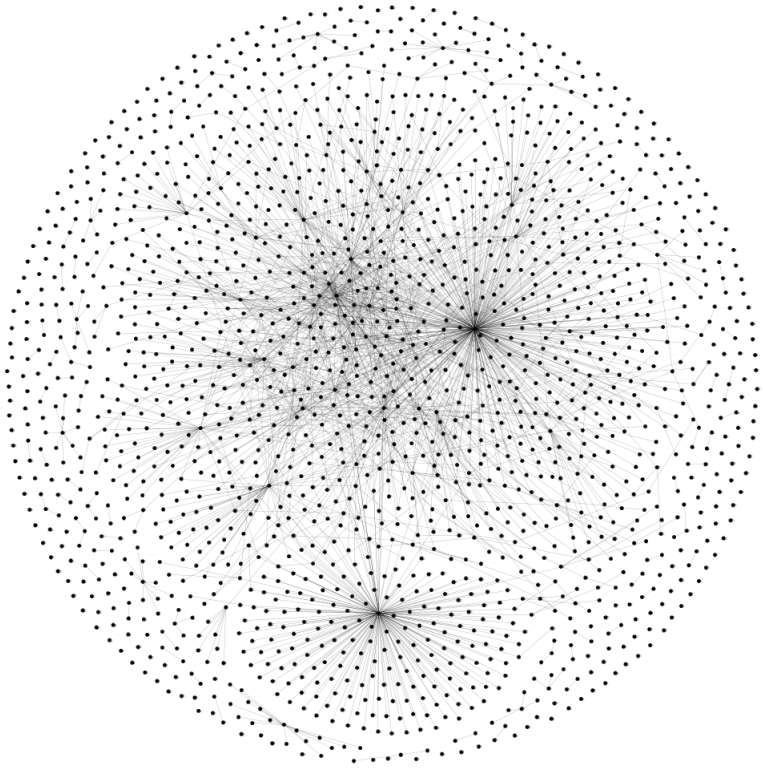


Figure 3.2: The #ThinSpo network.

We can see that the #ThinSpo network is a lot larger than the #ProAna network.

4 Data

For the #ProAna network, we have the following statistics:

Username	Betweenness Centrality	In-Degree	Out-Degree	Page Rank	Katz Centrality
neelessie	3.522118906734291e-05	3	2	0.007545242891350206	0.1032319561362469
ana___winter	0.0	7	1	0.03239965502207746	0.1147911241935483
skinnyribs2017	0.0	9	1	0.04499346786616223	0.1291400013910617
minutrodivita	7.044237813468582e-05	1	2	0.0016929786814907618	0.07102700112021892
thin_dreams5	0.0	15	1	0.06356392504800376	0.17865251901805992
aesthicc_png	0.00017610594533671456	7	2	0.02031074380686926	0.11608088075579671
xyvvvx	3.522118906734291e-05	2	2	0.002944562550730447	0.07891888999008768
clanais_83	0.0	0	14	0.0009151229118655222	0.06457000101838084
avavalentine17	0.0021484925331079177	8	6	0.005544016054873571	0.12276271682968085
wildforhim3	0.0	9	1	0.0671062201479085	0.1513712700085873
innn05403256	0.004226542688081149	20	6	0.01861254608950138	0.21882022848007432

Of course, I visualized this data using Gephi as well. One of the names got cut off for the three graphs. That username was 'innn05403256'.

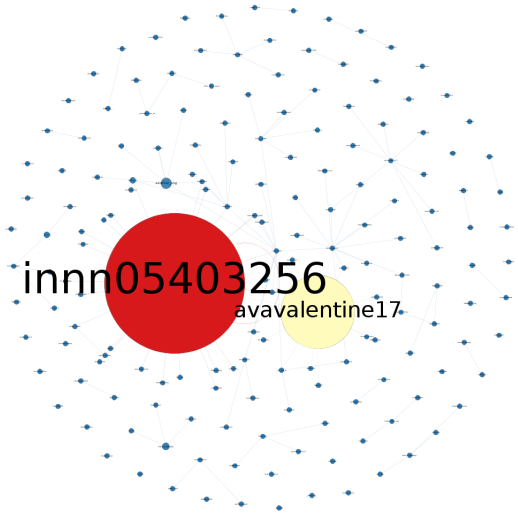


Figure 4.1: The #ProAna network by Betweenness Centrality.

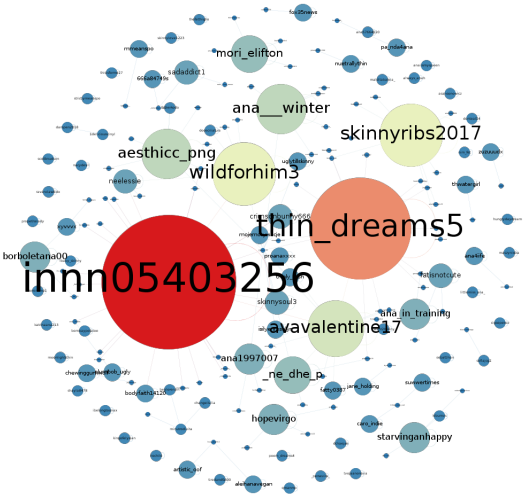


Figure 4.2: The #ProAna network by in-degree.

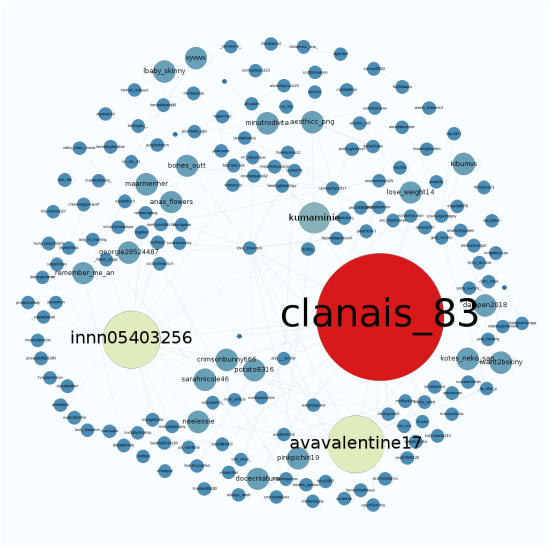


Figure 4.3: The #ProAna network by out-degree.

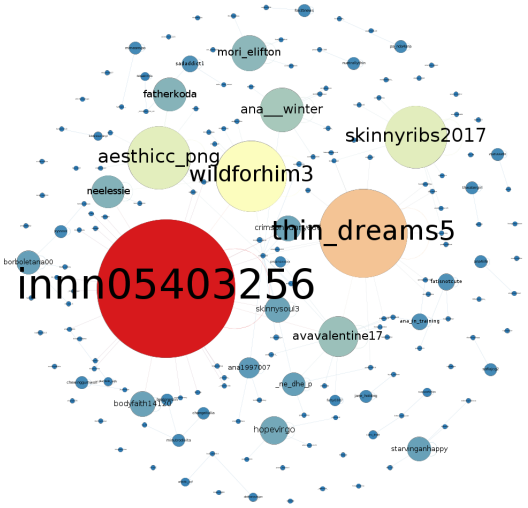


Figure 4.4: The #ProAna network by page rank.

Notice that in the Page Rank graph, innn05403256 is larger compared to thin_dreams5 even though the table shows that thin_dreams5 has a higher rank. This is due to the slight differences in algorithms for page rank between Gephi and Networkx.

For the #ThinSpo network:

Username	Betweenness Centrality	In-Degree	Out-Degree	Page Rank	Katz Centrality
eatingdisxrder	0.00011256951629823774	56	6	0.0021738605715538976	0.10569611783975785
miinxup	0.0	55	1	0.021495330542171568	0.11580392247007133
anniexlia	0.0	38	1	0.017275780270885357	0.07669191768516982
fatbluewhale	0.00025194618930618354	5	3	0.001467343232016007	0.034565036998764516
thinspo_twt	0.001573456847564647	19	207	0.0011563068232999024	0.048063992136482575
beautybuterflyy	0.0	31	1	0.016571726860017493	0.07485649196580835
sadeyedshadow	0.0	76	1	0.018000755924895116	0.14746493279031322
lbaby_skinny	0.00036718615033412403	63	4	0.004166427173239081	0.12585868214974555
chicx_thinspo	1.602266837821632e-05	3	25	0.0002162372975473688	0.018481621011818238
wowsothin	0.0	367	1	0.1470304719994481	0.6234995006509706
bluerainn	0.00023238004619047703	27	6	0.0008937544299082944	0.05947983841755417
avavalentine17	7.045865709907946e-05	8	27	0.0002972560574415999	0.025767790967350473
innn05403256	0.0004254223873241794	14	25	0.0004933076167373437	0.04271878134467371

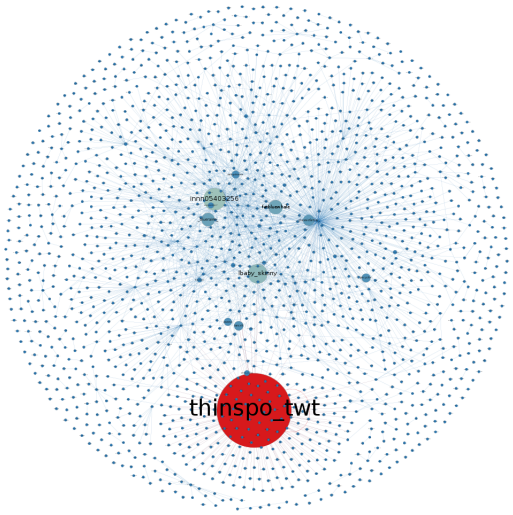


Figure 4.5: The #ThinSpo network by Betweenness Centrality.

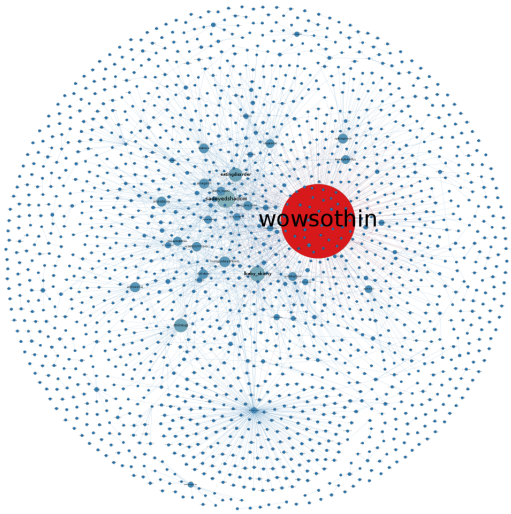


Figure 4.6: The #ThinSpo network by in-degree.

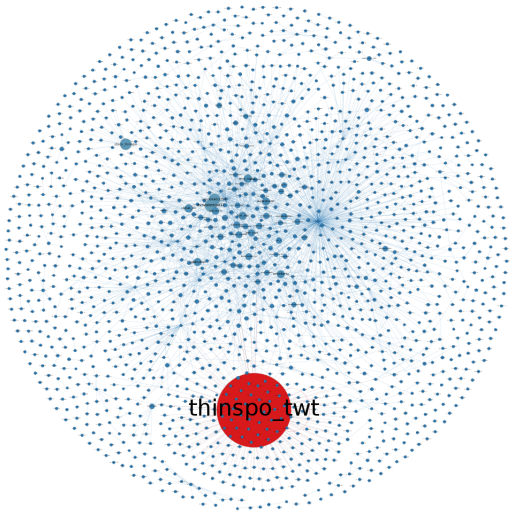


Figure 4.7: The #ThinSpo network by out-degree.

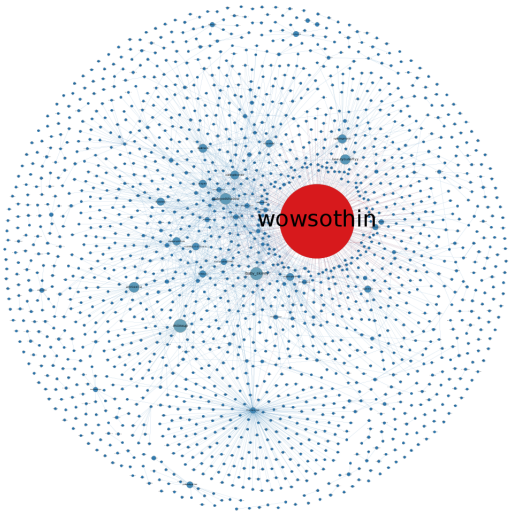


Figure 4.8: The #ThinSpo network by page rank.

The reason there is no visualization for Katz Centrality is because Gephi does not have that metric, which is unfortunate since it is fairly important to networks.

5 Conclusion

By Looking at the table, we can tell that using only one metric is not sufficient enough to find the important nodes. Some nodes with a betweenness centrality of 0.0 are important using other metrics. The influencers for the #ProAna Network are as follows:

innn05403256 — betweenness centrality

innn05403256 — in-degree

clanais_83 — out-degree

thin_dreams5 — page rank

innn05403256 — katz centrality

The influencers for the #ThinSpo network are as follows:

thinspo_twt — betweenness centrality

wowsothin — in-degree

thinspo_twt — out-degree

wowsothin — page rank

wowsothin — katz centrality

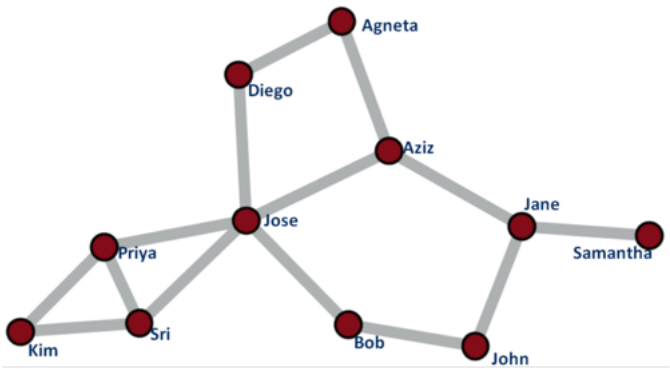
That is a bit unexpected. The #ProAna network has a diversity of users who are important in different ways. The only repeat on that list is 'innn05403256' who also appeared in the ThinSpo table. On the other hand, the #ThinSpo list has only two users spread over the five metrics we looked at.

It would be easy to call it there and say that these seven users are the most important in their respective networks, but we can't just yet. Not all metrics are equal. Just because 'clanais_83' has a large out-degree doesn't mean that that user is just as important as, say 'thin_dreams5' who has a high page rank.

Degree, whether it's in-degree or out-degree, is a simple way to check a node's popularity. In-degree in this network means that the account gets mentioned, or one of the account's tweets is retweeted or replied to. Out-degree is the amount of times that an account mentions other accounts, retweets a tweet or replies to a tweet.

Certainly, the nodes with the most connections are important in their own respect, but that hardly makes them the most important node. A node can have two connections and still be extremely important, such as being able to bridge two communities. That's where betweenness centrality comes in. You can take out the node with the most connections, and the network may not dissolve, as there are thousands of other nodes. But, if you take out the one node that bridges two halves of a network, you suddenly have two smaller networks. This is a simplified example, of course, as it is highly unlikely that there is only one node that bridges two communities.

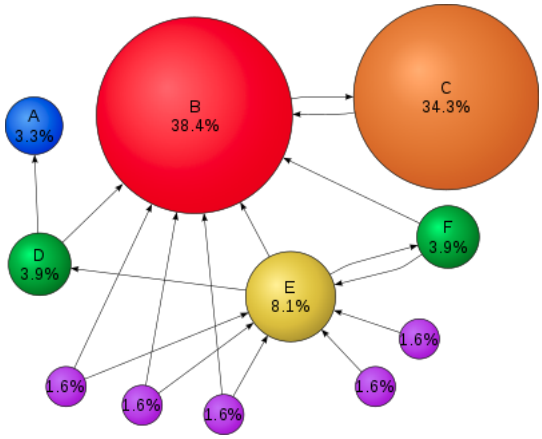
Even further, a node can have thousands of connections, and those connections can be relatively unimportant compared to the connection a node has with only one neighbor. This is where page rank is important.



In the above image from Wikipedia, each node is sized according to page rank, which is displayed under the letter of the node in a percentage. Notice that node E has more connections than node C. One would expect node E to have a higher page rank than C, and yet it doesn't. Think of node B as Google. If

Google doesn't link to your webpage like node B doesn't link to node E, people have a lower chance of seeing it no matter how many other websites it links to. Node B does however link to C and vice versa. Having such an important connection makes node C that much more important than node E.

The last metric that we looked at was Katz Centrality. It is similar to betweenness centrality in that it measures centrality (importance) of nodes. Where betweenness centrality measures importance using the shortest path through every node, Katz centrality measures importance by assigning weights to each of its connections.



Again, an image from Wikipedia. Using their example, the connection between John and Bob is strong since they are directly connected. John is also strongy connected to Jose through Bob, though his connection to Jose is weaker. John and Kim are weakly connected, though still connected. The Katz centrality algorithm assigns each of these a weight (determined by a constant that then drops off exponentially as the distance between two nodes is increased) and computes the node's relative importance.

With all of that in mind now, we can safely say that degree isn't really the best way to find an important node. If one wishes to target a network and dissolve it, then one should look for the nodes with high betweenness centrality, page rank, and Katz centrality.

In the end, all of this talk of importance doesn't change much. We now have the most important nodes of the #ProAna Network being:

innn05403256

thin_dreams5

with only one node being removed, and the most important nodes of the #ThinSpo network being:

thinspo_twt

wowsothin

Of note is that innn05403256 also appears to be somewhat important in the #ThinSpo network as well.

A final thought: the #ThinSpo network seems to be make up of people sharing pictures to encourage people to keep with their anorexia, to keep getting thinner and thinner. One would expect an account that solely posts like this to be very important to the network and, lo and behold, the user wowsothin posts pictures hourly with the hashtag ThinSpo. The ProAna network on the other hand, while smaller, is made up more of real users talking with one another and encouraging one another rather than accounts that don't talk much to each other. This may be why the #ThinSpo network is dominated by two accounts and the #ProAna network is a bit more diverse.