We use Latent Dirichlet Allocation (LDA) model in processing tweets. LDA is a model that classifies observations into unobserved groups. Each document is viewed as containing several topics and the number of topics is determined manually beforehand. Each topic of a document will receive a probability that sums up to one for each document. The probability of each topic is determined by the words of the document. Meanwhile, each topic is also presented as a distribution of words.

We set the number of topics to 20 in our codes and use the LDA model to estimate the topic distribution of each document. After we get the distribution, we calculate the average percentage of each topic by user. That is, we sum up the percentage of each topic for all the tweets of each user, then divide them by the number of tweets. After this step, we get a percentage for topic for each user.

We decide to build links among users for our networks by how similar their topics are. Since we already have the average percentage of each topic by user, we calculate the squared Euclidean distance of topic percentage for each user pair. The formula for a similarity score is

It is obvious from the equation that when two users have a low score, their tweets are in general more similar. Since we do not have a very clear threshold for links, we decide to define the users of the bottom 5 percentile score with a strong link, the users of the bottom 5-10 percentile score with a weak link, and the users of the bottom 10 -15 percentile score with a weakest link. These links are represented in our networks using different color, width and opaqueness, with black, wide and opaque link representing the strongest link (or users with most similar tweets).

(Describe the final network)

Other than this, we also use the same model mentioned above to construct network for different time periods, as we expect there to be differences in topics over different time span, depending on the situation of ISIS and how the war is going. We separate our database into three periods, the first period is (describe three networks)