Social Network Analysis of ISIS Related Tweets in Twitter

NATHANIEL SIMON, Marquette University
WILL HIBBARD, Marquette University
FAYIKA NOVA, Marquette University

1 ABSTRACT

The paper explores the ISIS related tweets and twitter environment and how its social network is constructed. Within the timeline from January, 2015 to May, 2016, our paper mainly focused on four pivotal sections of this tweeter dataset: Network Analysis, Timelines Analysis, Hashtag Analysis, and Topic modeling. Through this paper, we were able to see the most pivotal nodes within the network, prominent timelines of sharing tweets and related events, vital hashtags and its influences on nodes, and the most important topics that are discussed within such twitter network. Such in-depth exploration of ISIS related twitter network can inform readers about how events, hashtags, and other keywords can influence an online twitter community to be updated on the news related to ISIS terrorist organization.

ACM Reference Format:
Nathaniel Simon, Will Hibbard, and Fayika Nova. 2018. Social Network Analysis of ISIS Related Tweets in Twitter. 1, 1 (December 2018), 16 pages. https://doi.org/10.1145/1122445.1122456

2 INTRODUCTION

ISIS is a radical terrorist group started in 2004 as a local militant group and subsection of Al-Qaida. In 2013, the group absorbed an Al-Qaida backed military group and declared itself the Islamic State in Iraq (ISIS). Its goal is to create a society under Sharia law and mirror itself to 18th century Iraq (6). It gained prominence in 2014 after taking over several Iraqi cities and putting out a message calling for young Sunni men to come and fight with them (4). Their philosophy is to use violent means to achieve their goal of conquering territory in Iraq and Syria for their society (6).

The goal of this project is to analyze a set of ISIS related twitter accounts and derive patterns of activity from within them. Our group chose this set for its text and attribute components, and for its relevance to overall current politics. The goal of this project was to use network and textual analysis to derive patterns from our selected data set, and possibly make predictions about the network based on our data. The outcome of this will be several visual graphs and networks that illustrate our findings and will help us answer our research questions. We have several research questions that form the guiding framework for our project. The **first** question is to find central nodes in the data set and see who work as the bridges within this online community. Our **second** question is what times and events were the most important in the ISIS twitter community. This means deriving if there were spikes in online activities following different events during the data set's time frame. And **finally**, what hashtags, keywords and discussion topics were most prominent

Authors' addresses: Nathaniel Simon, Marquette University; Will Hibbard, Marquette University; Fayika Nova, Marquette University.

XXXX-XXXX/2018/12-ART $15.00
https://doi.org/10.1145/1122445.1122456

in the data set. Finding these will allow for contextualization of the data set and indications of which topics were most important to the group.

The other part of this project will be performing a lit review on our sources to inform our discussion and to provide the human context for our results.

3 LITERATURE REVIEW

While ISIS carries out wars in Syria and Iraq, their presence on the front lines of social media is how they threaten the rest of the world. ISIS has taken a modern-day approach to their military recruitment, using sites like Facebook and Twitter to spread propaganda globally (4). Thanks to their tech savvy outreach network, they have been able to recruit tens of thousands of members since 2015 (7). In this new age of social media information, data theft, and online anonymity, ISIS's online conscription has prompted authorities and researchers to consider new questions about how they deploy their tactics through the world wide web (4). Three such questions guide this report as we analyzed a data set of 17000+ tweets from the extremist group.

The first question regarding ISIS is about how their network within social media is structured. It is surprising to learn that ISIS has a complex and nuanced strategy to how their online presence operates. For starters, ISIS divides into online activities among six different types of members. In short order there are: reporters, who post ISIS related news; reconnectors, who retweet the names of alternate ISIS accounts if the primary accounts get suspended; intellectuals, who use academic fields like political theory and philosophy to justify the groups actions without directly supporting them; fanboys, who have pro-ISIS profile pictures or celebrate ISIS military victories; recruiters, who privately DM profiles they see as potential recruits; and mujahideen, who are front line fighters that depict ISIS life as glamorous and heroic (4). From these six classes, we can cast a wide net of behaviors that can catch more users than just those who directly and openly support ISIS. Users who discuss ISIS can also be broken up into simpler categories based on their language regarding the group. Researchers like Magdy and her team (2) used keyword searches to analyze what words and hashtags could be considered pro-ISIS. The result was that user profiles that were for ISIS used a different set of terms than twitter profiles that were consistently against or neutral to ISIS. For example, pro-ISIS tweets in their dataset consistently referred to ISIS as "The Islamic State" while consistently critical and neutral tweets simply referred to the group as ISIS (2). Additionally, it has been found that interactions with known ISIS tweets such as reposting, disliking, commenting, etc. can yield an intent as to whether or not the user interacting with that post is for or against ISIS (3). A consistent pattern of interactions with pro-ISIS posts, such as retweeting, can out seemingly neutral users as ISIS members or sympathizers (3). Twitter is a great site for nodal network analysis, as nodes (users) are linked through edges (@'s and retweets). An @ in Twitter is a direct mention of another user in the text of a tweet, and a retweet is when a user republishes a tweet on their profile while the network still accredits the original poster as the author of said tweet. Each tweet is time stamped, which makes it very easy to form a time progression in the set.(11)A unique feature that Twitter offers to nodal network analysis is the hash tags. Hashtags often act as a contextual word or phrase to summarize of link a tweet to the subject matter the tweet is referring to (10). This is tremendously helpful for keyword and thematic analysis. With these three angels of user roles, keyword corpus, and user interactions, a picture of how ISIS runs its social media campaign can be comprehensively determined.

While it is easy to see that ISIS have gain support over time based on the metric of how many people are recruited online, it is also worth investigating what events in the past have caused a spike in user activity, and possibly what event could trigger a rise in ISIS related usage in the future. Magdy and her team (2) have used hashtag analysis in order to see which ones have been the most prominent on dates significant to ISIS. They used topic modeling and hashtag analysis to create a timeline of when the greatest activity of pro and anti-ISIS discussions took place from October to December in 2013 (2). They started their search by looking at dates of ISIS military victories and loses and searching for the most prominent hashtags on those days. They took the data gathered from this and put it into a figure depicting prominent pro and anti-ISIS hashtags based on the day they were used. In the long term, support for ISIS can be tracked using keyword topic searches and looking at how many members have joined since their launch (5). Researchers like Imran Awan (5) has investigated this by gathering statistics on people who join ISIS over time. While their work focuses more on the determining what types of people join ISIS, the dates and other time related information paints a good picture for ISIS's support metrics in 2015 (5). Research done by a Cambridge team led by Suleyman Ozeren supplements this fact by doing an in depth of the month by month break down(8) of ISIS related accounts created during that same year.

The final question guiding the inquiry into the ISIS social network is determining what the most prominent hashtags are in a given set of ISIS tweets. One the outset, this task would seem the simplest, as all one needs to do is look at the most prominent tweets within the data set and look up what they mean. This is not the case however, as there are several complications to this Occom's Razor type solution. One complication is that finding the most prevalent hashtags in a dataset falls under the category of topic modeling. Some researchers have used a Term-Frequency Inverse-Document score and a word2vec model to get a corpus of the most relevant words (1). Code often needs to be written and tailored to meet the goals of the project, though the base for this code is sometimes available online. Isolating the top-k most prominent hashtags then leads to another hurtle, that of hashtag hijacking. By no means exclusive to just ISIS, online hate groups can hijack a popular hashtag or threat by posting their own content and tagging the post with that hashtag. It is how hate groups break into and take over mainstream conversations. With this in mind, simply googling a hashtag's meaning may give a result unrelated to the topic the hate group was using it for. This brings us to the method of corpus analysis, and the corpus mentioned earlier from the Inverse-Document score and the word2vec model. A corpus is a collection of texts, and in the case of what was being searched for, a set of the top-k tweets featuring a prominent ISIS hashtag. Researchers can go through these tweets manually to derive the common themes and meaning. This textual analysis is often the final step of this process, are people are needed in order to understand the language and subtly that algorithms can't.

Looking at the three specific research inquiries on prominent nodes, timeline, hashtag and topics in our dataset, the road map for this paper was to determine how ISIS connects it online network between users. This may be accomplished with a topic modeling analysis and psych profile like other researchers have used in the past (1). The next step is to take the topic model and apply it to a time line to determine when ISIS has become prominent on Twitter. This may be accomplished with crossing hashtag analysis with temporal data on a timeline. That approach would be similar to Alfifi, Kaghazgaran, and Caverlee's method to track ISIS's influence across social media (3). After tracking how ISIS has grown over time, the last step is to try and determine which hashtags ISIS uses most frequently, as well as what those hashtags means. This step will be similar to Madgy, Walid, et. al.'s appraoch to mapping pro and anti-ISIS related hashtags in a given time frame (2).

4 METHODS

The dataset was selected by browsing Kaggle and looking at different datasets to see if they were both interesting and fit the criteria provided. The criteria were that the dataset needed to be relatively large but still manageable to process, had a text component to analyze, and have some way that the nodes were linked to each other. After discussing a few sets that did not live up to the criteria, the ISIS-tweets dataset was selected (9).

- **Name**- The Name of the twitter profile
- **Username** – The unique username for the profile, displayed below the name of the profile
- **Description** – The description box of the profile in which the user can put information about the account or the profile
- **Location** – The location of each users. Putting location is not compulsory for users
- **Followers** – A follower is someone who subscribes to an account's posts
- **Numberstatus** – It shows the total number of tweets an user has tweeted till that date
- **Time** – The time is the time stamp of when a tweet was created
- **Tweets** – The textual components of the tweets

It was relatively large, having 17000+ tweets large, with 100+ accounts reported to be run by "pro-ISIS" users. The time frame of the dataset was from January of 2015 to May of 2016. The tweets were written in English, Arabic, and French, though the Arabic and French tweets were cleaned for the conveniences. The dataset did have location information in some user descriptions, but this was not the case for every user.

The software used for programming code for processing the dataset was Anaconda Distribution and Jupyter Notebook (14). There were several different quantitative measures taken to analyze the dataset. The first method used was creating a centrality matrix. A centrality matrix is a matrix that looks at how many degrees or connections a node has. (12). The degree of centrality value assigned at the end of the calculation is a numeric representation of how important or vital a node is to the network as a whole. The next method used was visualizing the top-k nodes, in this case user profiles, of the network. By graphing the nodes in Jupyter and connecting them by using the @'s in the tweets as directionality for which way an edge was pointing. The thicker an edge was between nodes, the more of a connection or contact those two nodes had with each other. Doing this revealed the clusters of certain users within the network, as well as showing a visual representation of which users were acting as bridges of contact or information for other users. The betweenness centrality score is a numeric value used to determine bridge nodes, or nodes that act as brokers of information between clusters in the network (12). The betweenness centrality score is determined by determining how many nodes a piece of information has to travel between two nodes to take the shortest path. The numeric score is a representation for how many "shortest-paths" a given node is a part of or a crossroads to. Nodes with the highest score are often revealed to have much power in the network as information brokers between clusters.Eigenvector centrality is numerical value is assigned based on how much influence any given node has on the whole network. In other words, if one twitter user were deleted or they sent out a tweet, what effect would that activity surrounding the users have on the rest of the network (12). Closeness centrality is the measure of how far a node is from all other nodes. This is useful for determining the visualization of the network as the nodes with the highest Eigenvalues to each other are not necessarily the closest physically (12).

After all of these scores were taken among the top-n, in this case 20, users, the next step was to look at the users and try to determine their possible function in the network. The users appeared in the top 20 list due to the high volume of tweets they either authored or were mentioned in. Due to this, they can be viewed as significant to the network in some capacity. A nodes possible function was determined by its score of any of the above-mentioned values. If one node had consistently high scores throughout the listings, then they were determined to be significant to the network based on the score. For example, the node that scored the highest betweenness value was likely a powerful bridge of information or an information broker in the network. The node with the highest Eigenvalue may node have the most connections in the network, but it is probably an originator of

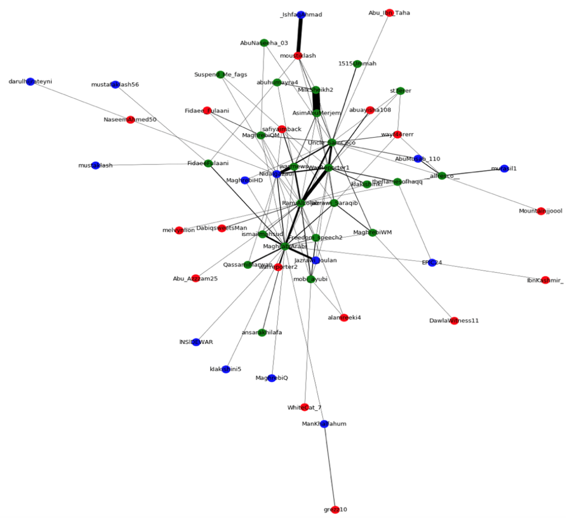

Fig. 1. Sender/ Receiver Network graph

content in the network. That node is then likely to be viewed as a ringleader in the context of the ISIS network.

5 CHALLENGES

There were obstacles to clean the dataset. The first and most apparent obstacle was that some of the tweets were corrupted when downloaded. There was no common factor to the corrupted nodes outside of a few common symbols, so most of the corrupted nodes had to be deleted manually. This shrunk the size of the dataset a relatively significant amount, this the decreased size actually solved another issue of slow processing due to the large file size. From then on, cleaning out data did increase the run time of process code, but that had the trade-off of removing tweets that could have been used to get a fuller picture of the ISIS network. Tweets with emojis and other symbols that caused an error in the code were removed, as well as tweets with links to images. Ultimately, the cleaning process left us with a smaller dataset than we had started with. This left us with, while still comprehensive, a smaller picture than the original dataset would have painted had we the code to fix it.

6 RESULTS

6.1 Network Graph

In order to thoroughly visualize all of our user in our social network we plotted them on a network graph. To do this to our dataset, we first needed to extract all the usernames from the set. We were then able to program a method to determine if they were a sender, receiver or both. We were then able to analyze the table and created a visual of weight of the lines to show the amount of conversations nodes had with one another. So if a line is thicker than another it implies that those two nodes converse more at each other. We were then also able to color code the sender, receiver and both. So Blue was nodes that were only receivers, red was nodes that were only senders and green associated with both senders and receivers. We are able to see the thickest line in the graph is between Milksheik2 and AsimAbuMerjem. RamiAlLoah and WarReporter1 also had thicker edges between them. This means they had many conversations directed at one another. And we can also see that they all are both senders and receivers (Fig. 1).

- **BLUE**: SENDER
- **RED**: RECIEVER
- **GREEN**: BOTH

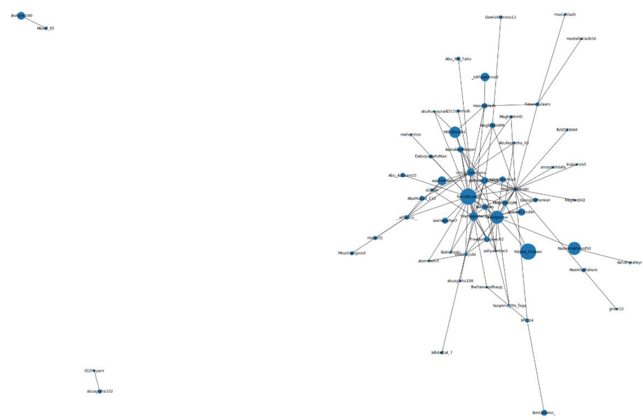

Fig. 2. Occurrence network graph

We were also able to visualize the occurrence of the nodes in the network. By occurance, we meant the nodes with highest tags by other nodes. To do this we had to create a counter to determine the amount of sends and receives in the network. We were then able to graph this and the nodes with the most occurrence had larger circles. Once we plotted this we were able to see that RamiAlLolah was one of the nodes with the highest occurrence in the dataset. This made sense since he is the node with one of the most tweets (18000+), followers (34000+) and thicker edges (Fig. 2).

6.2 Top Users

After conducting the above-mentioned tests, the results answer the first fundamental question of the paper: who are the central nodes, who are the bridges, and how did we find these. In first looking at degree of centrality, the top two accounts were RamiAlLolah and MagrabiArabi both having 19 degree and 0.3275 as degree centrality (Fig. 3).

```
In [15]: n_1 = float(len(node_to_degrees) - 1)
         degree_centrality = [ (n, k, k/n_1) for n, k in node_to_degrees ]

         ## node, k, dc
         sorted(degree_centrality, key = lambda x: x[2], reverse=True)

Out[15]: [('RamiAlLolah', 19, 0.3275862068965517),
          ('MaghrabiArabi', 19, 0.3275862068965517),
          ('Uncle_SamCoco', 16, 0.27586206896551724),
          ('WarReporter1', 14, 0.2413793103448276),
          ('Nidalgazaui', 14, 0.2413793103448276),
          ('Jazrawi_Saraqib', 8, 0.13793103448275862),
          ('__alfresco__', 6, 0.10344827586206896),
          ('warrnews', 6, 0.10344827586206896),
          ('mobi_ayubi', 6, 0.10344827586206896),
          ('MaghrebiQM', 6, 0.10344827586206896),
          ('moustiklash', 5, 0.08620689655172414),
          ('AsimAbuMerjem', 5, 0.08620689655172414),
          ('safiyaimback', 5, 0.08620689655172414),
          ('Freedom_speech2', 5, 0.08620689655172414),
          ('FidaeeFulaani', 4, 0.06896551724137931),
```

Fig. 3. Degree Centrality

The degree of centrality assumes that the node with the most connections is the most popular node in the network, which means both RamiAlLohah and MaghrabiArabi are the most popular nodes in the network. Even though both of these nodes had the highest degree centrality, we wanted to check if they had the same betweenness centrality or not. As it showed, MagrabiArabi had the highest betweenness centrality within the network even though both MaghrabiArabi and RamiAlLohah had same degree centrality (Fig. 4). Betweenness value suggests which node has

```
Name: MaghrabiArabi | Betweenness Centrality: 0.337684532648235 | Degree: 19
Name: RamiAlLolah | Betweenness Centrality: 0.2529413266709093 | Degree: 19
Name: Uncle_SamCoco | Betweenness Centrality: 0.22275477093988902 | Degree: 16
Name: Nidalgazaui | Betweenness Centrality: 0.1455415576558952 | Degree: 14
Name: WarReporter1 | Betweenness Centrality: 0.11256020144404898 | Degree: 14
Name: FidaeeFulaani | Betweenness Centrality: 0.06906634402097198 | Degree: 4
Name: __alfresco__ | Betweenness Centrality: 0.06879005245611418 | Degree: 6
Name: moustiklash | Betweenness Centrality: 0.05030752167775762 | Degree: 5
Name: EPlC24 | Betweenness Centrality: 0.03856624319419238 | Degree: 3
Name: mobi_ayubi | Betweenness Centrality: 0.035597614726471355 | Degree: 6
Name: Jazrawi_Saraqib | Betweenness Centrality: 0.03540015137111326 | Degree: 8
Name: MaghrebiWM | Betweenness Centrality: 0.03259873822487253 | Degree: 4
Name: NaseemAhmed50 | Betweenness Centrality: 0.032062915910465825 | Degree: 2
Name: ManKhalfahum | Betweenness Centrality: 0.032062915910465825 | Degree: 2
Name: MaghrebiQM | Betweenness Centrality: 0.02042462478034166 | Degree: 6
Name: warrnews | Betweenness Centrality: 0.008406072652896609 | Degree: 6
Name: Freedom_speech2 | Betweenness Centrality: 0.006827413360988678 | Degree: 5
Name: MilkSheikh2 | Betweenness Centrality: 0.006563823351482151 | Degree: 4
Name: theflamesofhaqq | Betweenness Centrality: 0.00504133897963299 | Degree: 2
Name: AsimAbuMerjem | Betweenness Centrality: 0.004612366862820585 | Degree: 5
```

Fig. 4. Betweenness Centrality

the shorted path between two nodes in a network. That being said, nodes with high betweenness scores are assumed to be bridges of information between different parts of the network because they are on so many shortest paths. Therefore, in this dataset, MaghrabiArabi works as the bridge between nodes. Interestingly, for the Eigenvector centrality, RamiAlLohah was the top node having the highest centrality of 0.3981. Eigenvector centrality score is based on how much influence one node has over the entire network, with weight being assigned based on the greatest affect among neighboring nodes. In that sense, RamiAlLohah had the highest influence within this dattaset over other nodes even though MaghrabiArabi was the bridges between them (Fig. 5). Finally, the last score assigned is the closeness centrality, which means that a node with a high closeness value physically most central to all other nodes in the graph. Once again, the highest scores are MagrabiArabi followed by RamiAlLolah, indicating that they are the most physically central to all other nodes in the network (Fig. 6). Expanding the top-n list to the top 5, the next top three users after RamiAlLolah and MagrabiArabi have been varying placements of UncleSamCoco, Nidalgazaui, and WarReporter1. Looking at the descriptions of accounts that placed consistently in each scoring list, the accounts are either on the ground activists or news sources. A caveat to mention is that the news source account tweets have only been ISIS related because this is an ISIS related tweet data set, so it is unsure whether the new profiles specialize in ISIS related stories or not.

6.3 Timeline Analysis

In our dataset, each tweet is marked with a unique timestamp. These timestamps are associated to the specific time a twitter user sends out a tweet into the twitter universe. Timestamps of tweets can show many different relations including twitter traffic. We wanted to look at how timestamps affected many different relations in the dataset. To do so, we needed to make a visual charts to see activity of our Twitter dataset users. First, we plotted a table to see the twitter activity over time. We named the title "Isis related tweets over time". While the x value is the dates 11-1-2015 to 4-30-2016 which was the entirety of the time line in our dataset. The Y values were the number of tweets during those days also referred to as the twitter activity. We can see that the first half of the table has more less spikes and has less traffic while the second half of the table has higher and more consistent peaks showing that there was more traffic. (Fig. 7)

```
Top 20 nodes by Eigenvector centrality:
('RamiAlLolah', 0.3981788586596127)
('MaghrabiArabi', 0.3458132586796163)
('Nidalgazaui', 0.34121851819674337)
('Uncle_SamCoco', 0.31350398783165534)
('WarReporter1', 0.31009343768250436)
('Jazrawi_Saraqib', 0.21917096989559448)
('warrnews', 0.21599108528911026)
('Freedom_speech2', 0.19777793684027065)
('MaghrebiQM', 0.1949280594210907)
('mobi_ayubi', 0.15875417971660566)
('AsimAbuMerjem', 0.15570177960633486)
('safiyaimback', 0.14616420964377327)
('ismailmahsud', 0.13811004631769974)
('MaghrebiWM', 0.11364306639584368)
('__alfresco__', 0.11308498353690169)
('moustiklash', 0.09858112029170368)
('wayf44rerr', 0.09380399901609851)
('st3erer', 0.09130160235737904)
('klakishinki', 0.09013857540348404)
('QassamiMarwan', 0.08743565224898271)
```

Fig. 5. Eigenvector Centrality

```
Top 20 nodes by Closeness centrality:
('MaghrabiArabi', 0.5292196007259528)
('RamiAlLolah', 0.5237068965517241)
('Uncle_SamCoco', 0.5130190007037297)
('Nidalgazaui', 0.49290060851926976)
('WarReporter1', 0.46124644099968365)
('MaghrebiQM', 0.441016333938294)
('Jazrawi_Saraqib', 0.4371814092953523)
('warrnews', 0.4297082228116711)
('Freedom_speech2', 0.4260666277030976)
('ismailmahsud', 0.3990147783251231)
('MaghrebiWM', 0.3990147783251231)
('__alfresco__', 0.386737400530504)
('moustiklash', 0.38378520663332455)
('MaghrebiHD', 0.3780140005185377)
('mobi_ayubi', 0.37519300051466803)
('AsimAbuMerjem', 0.37519300051466803)
('FidaeeFulaani', 0.36697709539390 89)
('Jazrawi_Joulan', 0.3669770953939089)
('QassamiMarwan', 0.3643178410794603)
('safiyaimback', 0.36169684941701813)
```

Fig. 6. Closeness Centrality

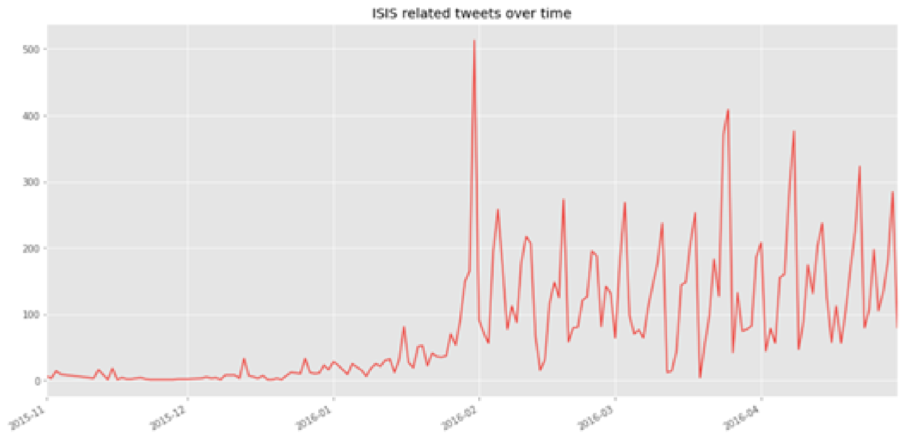

Fig. 7. ISIS related tweets over time

6.3.1 April 2016 in Depth. We wanted to take a look to see if specific dates corresponded to specific peaks in the time line. From Fig. 7, we could see the highest peak of tweeting was around the time February, 2016. However, as we observed that in the month of April, the frequency of higher peaks increased, we wanted to take a closer look at the twitter activity in that specific month. To do this, we took the tweets from 4-1-2016 to 4-30-2016 and set them all into a list called april-tweets. We needed to create another visualization using the specific tweets from the month of April we extracted. To do so we have to create a second chart called "Total number of ISIS related tweets over time". To create this table the x axis has to be april-tweets list we referenced earlier, and the y is the count of the number of tweets on that specific day. Next, we made a graph with the label "A", "B", "C", "D" Which were specific dates we were able to gather to see that there was rise in the peaks on those days. Below you can see the chart Displaying the specific events that occurred compared to the twitter activity on the particular day.(Fig. 8)

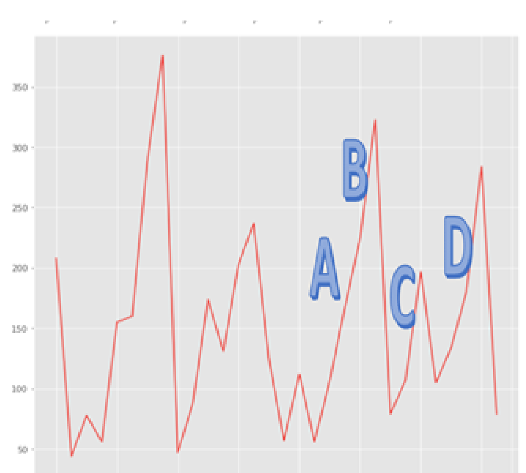

Fig. 8. ISIS twitter traffic in the month of April

- **A**: truck bombing in Kabul -> April 19, 2016
- **B**: Obama visits Saudi Arabia -> April 21, 2016
- **C**: Us Cyberattack on ISIS -> April 24, 2016
- **D**: truck bombing in Kabul -> April 27, 2016

6.4 Hashtag Analysis

Looking through the dataset we can see that there are many different hashtags being used. And observing them in many different manners can find a plethora of connections in the dataset. As we referenced earlier Hashtags often act as a contextual word or phrase to summarize of link a tweet to the subject matter the tweet is referring to. So, we wanted to analyze the important and pivotal hashtags in the dataset to find any links in data. To do this we had to scrape the dataset. To do that we have to fist go into the tweet section of the and create a function that looks specifically for the hashtag character. This function then would extract the word in which the hashtag character was linked to. After extracting the specific hashtags, we had to create a hashtag counter which counted the frequency of unique hashtags. We then added it into a chart making the x axis the top 10 unique hashtags and the y axis as the number of times those specific hashtags were used. We can see in the graph below that the top hashtag is ISIS and the 10th most common hashtag is Breaking.(Fig. 9)

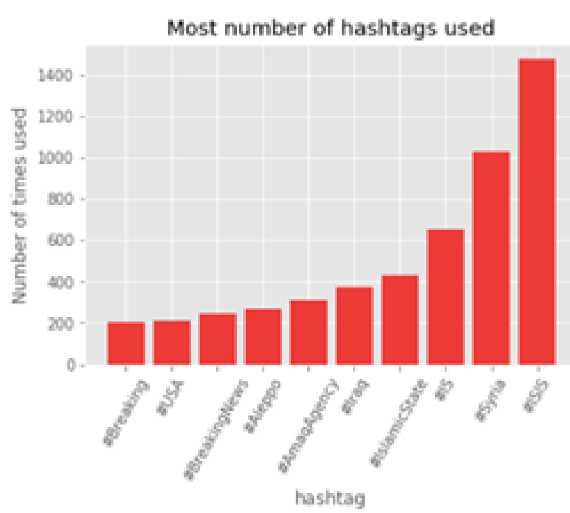

Fig. 9. Graph of the 10 most common hashtags

6.4.1 Frequency. With the frequency of specific hashtags, we found earlier we were able to also the list of the top 20 unique hashtags and see the number of how many times it was used throughout the dataset. We can see in the table below that that the top hashtag is ISIS with it being used 1497 times and the 20th most common hashtag is mosul which was used 152 times (Fig. 10). To understand the importance of these hashtags, doing a thematic analysis on our top 5 nodes from this dataset, we observed that they all used all the top 5 hashtags in their tweets. Next, we wanted to look for the top most common words in the dataset for a better and closer looks on the data. To do this, we looked through all the tweets and created a counter of the single words and placed in on a list. With this we were able to create a top 20 words table. We can see in the table below rt is the top word showing and it was mentioned 5874 times and the 20th most common word was

```
[('#isis', 1497),
 ('#syria', 979),
 ('#', 681),
 ('#is', 643),
 ('#islamicstate', 448),
 ('#iraq', 398),
 ('#aleppo', 378),
 ('#amaqagency', 322),
 ('#breaking', 320),
 ('#iraqi', 259),
 ('#breakingnews', 247),
 ('#turkey', 222),
 ('#syri', 216),
 ('#usa', 207),
 ('#palmyra', 204),
 ('#ypg', 193),
 ('#syrian', 191),
 ('#russian', 183),
 ('#assad', 157),
 ('#mosul', 152)]
```

Fig. 10. Frequency of the top 20 hashtagged words in the dataset

people which was mentioned 418 times. We can also see that retweets on tweets were a large part of the dataset due to the fact that rt was the most common word used in the dataset (Fig. 10)

```
Out[25]: [('rt', 5874),
          ('isis', 2306),
          ('amp', 1355),
          ('killed', 1287),
          ('syria', 1272),
          ('army', 902),
          ('state', 692),
          ('islamic', 686),
          ('assad', 591),
          ('aleppo', 586),
          ('ramiallolah', 577),
          ('near', 575),
          ('allah', 569),
          ('iraq', 562),
          ('breaking', 555),
          ('iraqi', 549),
          ('syrian', 543),
          ('soldiers', 528),
          ('today', 511),
          ('attack', 500),
          ('forces', 484),
          ('al', 477),
          ('city', 459),
          ('islamicstate', 456),
          ('people', 418)]
```

Fig. 11. Frequency of top 20 words in the dataset

6.5 Topic Modeling

Now that we have talked about the most important hashtags that we observed our nodes mostly shared through analyzing the tweets in the dataset, we dug a little deeper to understand what primary topics the nodes talk or discuss about within this network. Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic []. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. In this paper, we used LDA to model topics in our dataset.

6.5.1 Data Pre-processing. Before we model our topics, we pre-processed our dataset to divide them into tokens, remove punctuation, convert them into first person words and so on. We performed the following steps to pre-process our dataset:

- **Tokenization**: Split the text into sentences and the sentences into words
- Lowercase the words and remove punctuation
- Words that have fewer than 3 characters are removed
- All **stopwords** are removed
- Words are **lemmatized**— words in third person are changed to first person and verbs in past and future tenses are changed into present
- Words are **stemmed** — words are reduced to their root form

The two main inputs to the LDA topic model are the dictionary(id2word) and the corpus (Fig. 12). After we pre-processed the data, we created corpus and dictionary from the bag of words containing the number of times a word appears in the training set. Once we have those, we have everything required to train the LDA model.

```
dictionary.filter_extremes(no_below=15, no_above=0.5, keep_n=100000)
```

```
bow_corpus = [dictionary.doc2bow(doc) for doc in processed_docs]
bow_corpus[4310]
```

Fig. 12. Dictionary and Corpus For LDA

6.5.2 The Topics. We trained our LDA model using gensim.models.LdaMulticore and saved it to ‘lda_model’. Our LDA model is built with top 3 different topics where each topic is a combination of keywords and each keyword contributes a certain weightage to the topic. The keywords for each topic and the weightage(importance) of each keyword are shown next (Fig. 13).

```
Topic: 0
Words: 0.023*"kill" + 0.018*"break" + 0.016*"armi" + 0.016*"scotsmaninfidel" + 0.015*"isi" + 0.015*"soldier" + 0.0
14*"iraqi" + 0.013*"forc" + 0.013*"amaqag" + 0.012*"spicylatt"
Topic: 1
Words: 0.038*"isi" + 0.022*"syria" + 0.020*"islam" + 0.016*"state" + 0.015*"iraq" + 0.013*"ramiallolah" + 0.012*"a
ssad" + 0.011*"report" + 0.009*"allah" + 0.009*"armi"
Topic: 2
Words: 0.019*"islamicst" + 0.019*"kill" + 0.014*"isi" + 0.011*"muslim" + 0.010*"amaqag" + 0.009*"bomb" + 0.009*"al
lah" + 0.008*"support" + 0.008*"syria" + 0.008*"break"
```

Fig. 13. Top 3 Topics From Our Dataset

But how do we interpret these topics? **Topic 0** is a represented as 0.023*”Kill” + 0.018* “break” + 0.016* “scotsmaninfidel” + 0.015* “isi” +0.015*”soldier”. +0.014* “Iraqi” +0.013*”forc” +0.013* “amaqag” + 0.012* “spicylatt”. It means the top 10 keywords that contribute to this topic are: ‘kill’, ‘break’, ‘scotmaninfidel’ etc. and the weight of ‘kill on topic 0 is 0.023. The weights reflect how important a keyword is to that topic. Looking at these keywords, it can be assumed that the topic is something related to the breaking news of ISIS, about killing someone, Iraqi soldiers, forces and

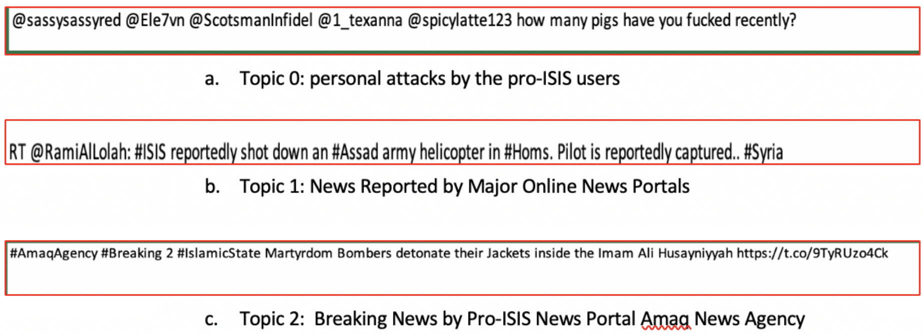

Fig. 14. Top 3 Topics Within Our ISIS Related Dataset: a) personal attacks by the pro-ISIS users, b) News Reported by Major Online News Portals, c) Breaking News by Pro-ISIS News Portal Amaq News Agency

some names like scottsmanfidel, spicylatte. For better understanding on the keywords and the topics, we went back to our initial dataset and manually searched for each of the keywords and the tweets that included the keywords. From the qualitative analysis, we saw the names scottsmanfidel, spicylatte are the shortened usernames of @scottsmanfidel, @spicylatte123 who have been tagged multiple times by other users in discussions of supporting ISIS. These users were outside of our dataset, which means we do not have information on them or what they tweeted related to ISIS. But based on the tweets where they are tagged, the majority of the tweets were concentrated towards bullying the users who were anti-ISIS and wanted to make statements against ISIS attacks.

Looking at the keywords and the texts itself, we can assume that the first topic in our dataset can be Personal Attacks by Pro-ISIS users (Fig. 14(a)). Similarly, looking at **Topic 1** and keywords generated by LDA, we can say that the topic is related to ISIS, Syria, Reporting News, Newsportals Rami Al Lolah and so on. Similarly, like previous one, going back to the dataset, we did see similar keywords in tweets mostly reporting news on ISIS attacks on Syria or Iraq states in the name of Islam and Allah (Fig. 14(b)). Our **Topic 2** is around keywords Islamic states, killing, ISIS, Amaq News Agency, bomb, breaking news and so on. Based on the qualitative analysis, we could interpret the topic as Breaking News on ISIS attack By Amaq News Agency (a pro-ISIS news portal online run/supported by the ISIS organization) that are vastly shared by the users in twitter (Fig. 14(c)).

7 IMPLICATIONS

7.1 Prominent Users

The results of this research imply an interesting trend about our ISIS related dataset. The top five nodes of the four scoring categories are either activists or news sources. It falls under the ISIS actors mentioned by (4). The fact that they are central nodes to the network means that they could be the main information suppliers. The activist accounts appear to give live coverage of the latest ISIS activities, while the news sources appear to give extensive coverage on the terrorist group. This suggests that much of the activity surrounding ISIS online could be people who are interested in becoming more informed about ISIS as well as communicating and discussing about these topics. The data makes it clear that events relating to ISIS cause increased activity in the ISIS related social network. Following the instance mentioned previously on how discussion in the ISIS network being based around discussion of learning more about ISIS, it is a logical next step that ISIS related events in the news spur people on to become informed on the matter. This is backed up by Figure 4, in which a back-to-back sequence of ISIS related events sparked an influx of tweets to be written on those dates. There are four specific events in this figure which are shown to be the cause of the increased activity. The four events are a truck bombing happening in Kabul, Barack Obama visiting Syria, a US cyber team attacking ISIS, and a suicide bombing taking place in Turkey. The event that spurred the most activity was Obama visiting Syria, while the event that sparked the lowest activity was the car bombing in Kabul. This trend in the data suggests that the events that spark discussion in this network are political based events, such as those involving policy and world leaders. It is worth noting that the news accounts fall under the categorization of reporter according to Zaman, but we decided that they were not pro-ISIS as all of their coverage and messaging was neutral (4).

7.2 Timeline

We can't see too much activity from November 2015 to January 2016 due to the fact that the spikes were not as high. This implies that there wasn't that much news to report on ISIS at that specific time. We can also see that the 4 labeled events: Truck Bomb in Kabul, Obama goes to Saudi Arabia, US Cyberattacks ISIS, Suicide Bombing in Turkey are all in locations where the spike starts to rise in the graph. This implies that important events like those cause larger traffic on those specific days. We see that different news worthy events cause large spikes in tweets. This implies when news occurs, we can see that many people tweet about it or have opinions about those specific topics. We can imply that larger ISIS events cause large twitter activity on ISIS twitter. We can see our largest spike in the data is toward the end of January 2016 and right before the start of February 2016. When observing the events in January we see in the Wilson center time line (13) there were two prominent events. The first one is when a suicide bomber with links to ISIS kills 10 people and injured 15 others - many of them German tourists - in Istanbul's Sultanahmet Square. The second event is when ISIS claimed responsibility for an attack in Jakarta, Indonesia, that killed at least two people and injured 19 others. These two events which were around the same time shows that with multiple prominent events in the span of a few days we can see large spikes and conversations carry for multiple days. We can also imply that large spikes in specific dates can show that ISIS activity during those days were highly shared and retweeted. This shows a relationship on how large events can show a large amount of retweets and information spreading.

7.3 Hashtags

We can see that isis is the most commonly used hashtag since it is used 1497 times. This shows that the topic of ISIS is of course the overall main theme of the dataset. Looking more into the hashtags and particularly when we analyze the top 20 hashtags, we can see that many of them are locations of countries. This could imply that there is heavy ISIS terroristic activity in those particular countries or regions. This shows the affect that ISIS has on other countries. For example, Russian was tweeted 183 times. This shows the prominence of ISIS in Russia. We can see in an ISIS timeline conducted by Wilson center says(9) "July 31: In a nine-minute YouTube video, the Islamic State urges its members to carry out attacks in Russia. "Listen Putin, we will come to Russia and will kill you at your homes ... Oh Brothers, carry out jihad and kill and fight them," says one masked man." Then later we see that they again reference Russia by stating (13), "Oct. 31 : Sinai Province, Egypt's ISIS affiliate, claims responsibility for bombing a Russian passenger plane over the Sinai Peninsula, killing all 224 on board." This shows that even when Russia isn't the primary country of ISIS they are still affected, and hashtags are able to help show importance of prominent events associated to tweets. This shows the relations between timelines and hashtags. It is also interesting to see that breaking is the tenth most tweeted hashtag in this dataset. This could imply decent amount of our users are not fully affiliated with ISIS but are more news reporters or are trying to spread their news through twitter. We did a thematic analysis with the top 10 users and found that they used all the top 20 hashtags in their tweets. This implies that the top users are very important due to the fact that they were able to spread the top hashtags. We can also see that Many of the top words are similar to the top tweets. With the word counter showing us that the top word is rt meaning retweet. This can imply that there was a spreading of information from nodes and that many other nodes received information from each other and spread it. This shows how fast news can spread around in a community and in this particular case ISIS twitter.

7.4 Topic Modeling

Our topic modeling helped us to build constructive understanding on what our nodes discuss in the twitter network envrionment and what are the prominent words that are associated with the topics. Whereas the hashtags and keywrods provided us information on what goes viral or which keywords were mostly popular, topic modeling backed us up with more defined topic containing those words. The top 3 topics we found were either related to war reports or personal attacks by pro-ISIS users. Also, as every single topic contains words about religion and war, this implies that every tweet in our dataset was related to ISIS's religious values of establishing Sharia law or some sort of terrorist attacks in different part of Syria and Iraq. Whereas news portal like RamiAlLohah, WarReporter1 were the nodes present in our dataset that frequently reported war related news by ISIS, one major news portal that came out of our topic modeling was Amaq News agency and their reported war news. Interestingly, RamiAlLohah and WarReporter1 were deemed as non-biased news agency, but Amaq News Agency was considered as pro-ISIS news portal, which were frequently criticized by the users in the dataset. Along with primarily talking or spreading news on the war, decent amount of the weights in the topic modeling application associated to specific users who were bullied by pro-ISIS users in our dataset. This implies that there are nodes in our dataset who are outwardly pro-ISIS fanboys supporting ISIS attacks and spread hatred towards those who criticise ISIS actions and attacks. Earlier we were able to see that ISIS online presence is seen in 6 different groups: reporters, re-connectors, intellectuals, fanboys, recruiters, and mujahideen (4). We see many of these users when we observing our tweets through timeline analysis, hashtag analysis and topic modeling. We can say these six groups of ISIS twitter can be separated in 3 outlooks on twitter which are for, against and neutral. The groups that are for are easier to decipher. This group would most likely contain: mujahideen, recruiters, fanboys, and re-connectors. The neutral and opposing groups are the most reporters, and the intellectuals. Some of these group members can be in multiple groups and there are some small grey lines. This topic of for, against, and neutral can also be seen when we did topic modeling. We can see that some information taken from the topic modeling were topics that were clearly for or against ISIS while some were basic information (neutral).

Overall we saw many different connections the dataset displayed. But some of the major takeaways from observing ISIS twitter trends would be the most influential people or nodes are those who are either activists or news portals that happen to be reporting on breaking news on ISIS. This shows that people want to know what is going on around them online through using social media platform like Twitter.

8 CONCLUSION

Nowadays people need to know news of an event or attack at the exact second and twitter is a large network of up to date news. The users in our dataset tried to be informed about such attacks by ISIS and one large way to do so was to follow or communicate with those who are informed on these particular issues such as news reporters or war activists. Our results suggested that, not only people follow the users who are mostly updated on ISIS related issues, but also shared or tweeted the most during any big attack by ISIS. This in turn created a large social network with many different nodes spreading information frequently. Through our social network and text analysis, we were able to understand the overall ecosystem and environment around the ISIS news on twitter.

9 REFERENCES

1. Nouh, Miriam, et al. “Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter.” Arxiv.org, Arxiv.org, 15 May 2019, https://arxiv.org/pdf/1905.08067.pdf.

2. Magdy, Walid, et al. “FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support.” First Monday, vol. 21, no. 2, 2016. Arvix.org, doi:10.5210/fm.v21i2.6372.

3. Alfifi, Majid, et al. “Measuring the Impact of ISIS Social Media Strategy.” Snap.stanford.edu, Texas AM University, 1 Jan. 2018, https://snap.stanford.edu/mis2/files/MIS2$_{p}aper_{2}3.pd$

4. Zaman, Khuram. “ISIS Has a Twitter Strategy and It Is Terrifying [Infographic].” Medium, Fifth Tribe Stories, 13 May 2016, https://medium.com/fifth-tribe-stories/isis-has-a-twitter-strategy-and-it-is-terrifying-7cc059ccf51b.m3zeluykl.

5. Awan, Imran. “Cyber-Extremism: Isis and the Power of Social Media.” SpringerLink, Springer US, 15 Mar. 2017, https://link.springer.com/article/10.1007/s12115-017-0114-0Sec7.

6. “ISIS Fast Facts.” CNN, Cable News Network, 6 Nov. 2019, https://www.cnn.com/2014/08/08/world/isis-fast-facts/index.html.

7. Cocco, Federica. “How Many Foreign Fighters Have Joined Islamic State?” The Week UK, Dennis Publishing Limited, 24 June 2015, https://www.theweek.co.uk/islamic-state/64120/how-many-foreign-fighters-have-joined-islamic-state.

8. Ozeren, Suleyman, et al. “An Analysis of ISIS Propaganda and Recruitment Activities Targeting the Turkish-Speaking Population.” Cambridge Core, Cambridge University Press, 4 Sept. 2018, https://www.cambridge.org/core/journals/international-annals-of-criminology/article/an-analysis-of-isis-propaganda-and-recruitment-activities-targeting-the-turkishspeaking-population/604BCDA57I reader.

9. Zaman, Khuram. “How ISIS Uses Twitter.” Kaggle, Kaggle, 14 Nov. 2019, https://www.kaggle.com/fifthtribe/h isis-uses-twitter.

10. “How to Use Hashtags.” Twitter, Twitter, n.d., https://help.twitter.com/en/using-twitter/how-to-use-hashtags.

11. Na. “GET Media/Upload (STATUS) - Twitter Developers.” Twitter, Twitter, n.d., https://developer.twitter.co media/api-reference/get-media-upload-status.

12. Bhasin, Jatin. “Graph Analytics-Introduction and Concepts of Centrality.” Medium, Towards Data Science, 19 Aug. 2019, https://towardsdatascience.com/graph-analytics-introduction-and-concepts-of-centrality-8f5543b55de3.

13. Wilson Center. “Timeline: the Rise, Spread, and Fall of the Islamic State.” Wilson Center, 31 Oct. 2019, www.wilsoncenter.org/article/timeline-the-rise-spread-and-fall-the-islamic-state.

14. “Anaconda Python/R Distribution - Free Download.” Anaconda, https://www.anaconda.com/distribution/.

15. Li, S. (2018, June 1). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. Retrieved from https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24.