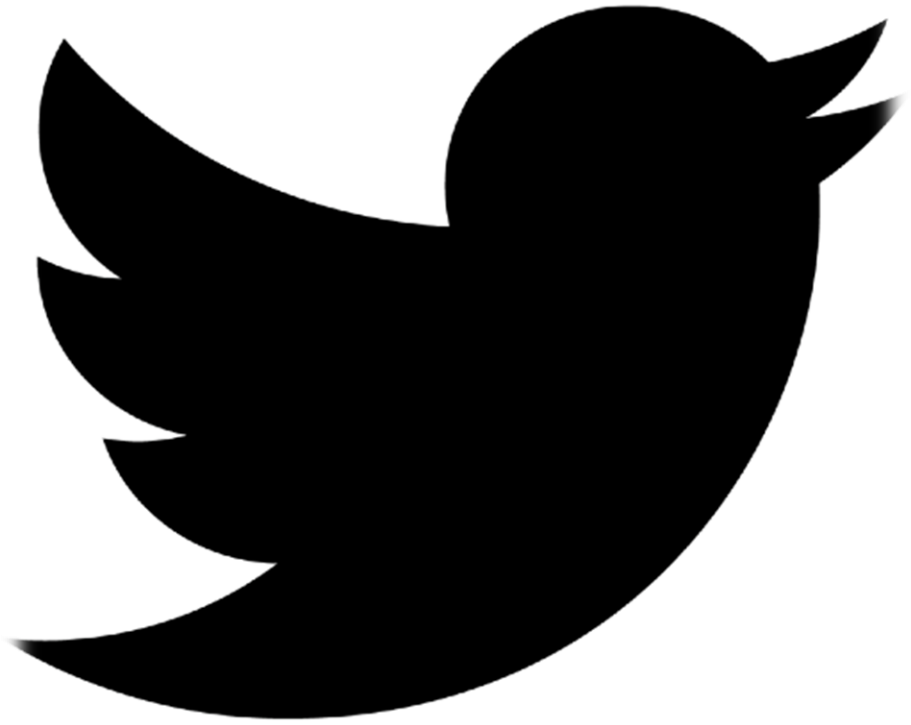
Using Data Science To dismantle ISIS's Twitter Network

17k+ tweets analyzed



Prepared by:
Ed G. Haddad

Preface:

The goal of this analysis is to explore common features exhibited by top twitter users that partially act as ISIS's propaganda network on the platform. Furthermore, we are going to investigate the qualities that distinctively mark their tweets.

Data set

The 17,410 by 8 data set at hand contains the following main variables:

- *Name*
- *Username*
- *Description*
- *Location*
- *Number of followers at the time the tweet was downloaded*
- *Number of statuses by the user when the tweet was downloaded*
- *Date and timestamp of the tweet*
- *The tweet itself*

Exploratory Data Analysis

We are going to perform exploratory data analysis on some of these variables to decide on the best approach to conduct the analysis.

One of the statistical impediments presented by this data set is the fact that observations (rows) are tweets that are associated with a group of other variables such as name, username…etc., which means a single account is going to appear multiple times in the data set.

Username	**Description**	**……**	**Followers…..**
GunsandCoffee70	ENGLISH TRANSLATIONS: http://t.co/QLdJ0ftews		640….
GunsandCoffee70	ENGLISH TRANSLATIONS: http://t.co/QLdJ0ftews		640….
YazeedDhardaa25	Observing a JIHAD NEWS mainly about Islamic State.		904….
YazeedDhardaa25	Observing a JIHAD NEWS mainly about Islamic State. . .		904….

Excerpt from the data set

As we can see in the excerpt above the account "GunsandCoffee70" would count twice in summary statistics unless we were able to find a way by which each account only count once. In order to do that we devised a way to draw only one occurrence of an account associated with multiple tweets.

After running the script on the original 17,410 by 8 data set it collapsed to a 112 by 8 data set, which matches the "100+ ISIS fanboys" specified by the original problem specification. For example the excerpt above would collapse to the one shown below.

Username	**Description**	**……**	**Followers…**
GunsandCoffee70	ENGLISH TRANSLATIONS: http://t.co/QLdJ0ftews		640….
YazeedDhardaa25	Observing a JIHAD NEWS mainly about Islamic State.		904….

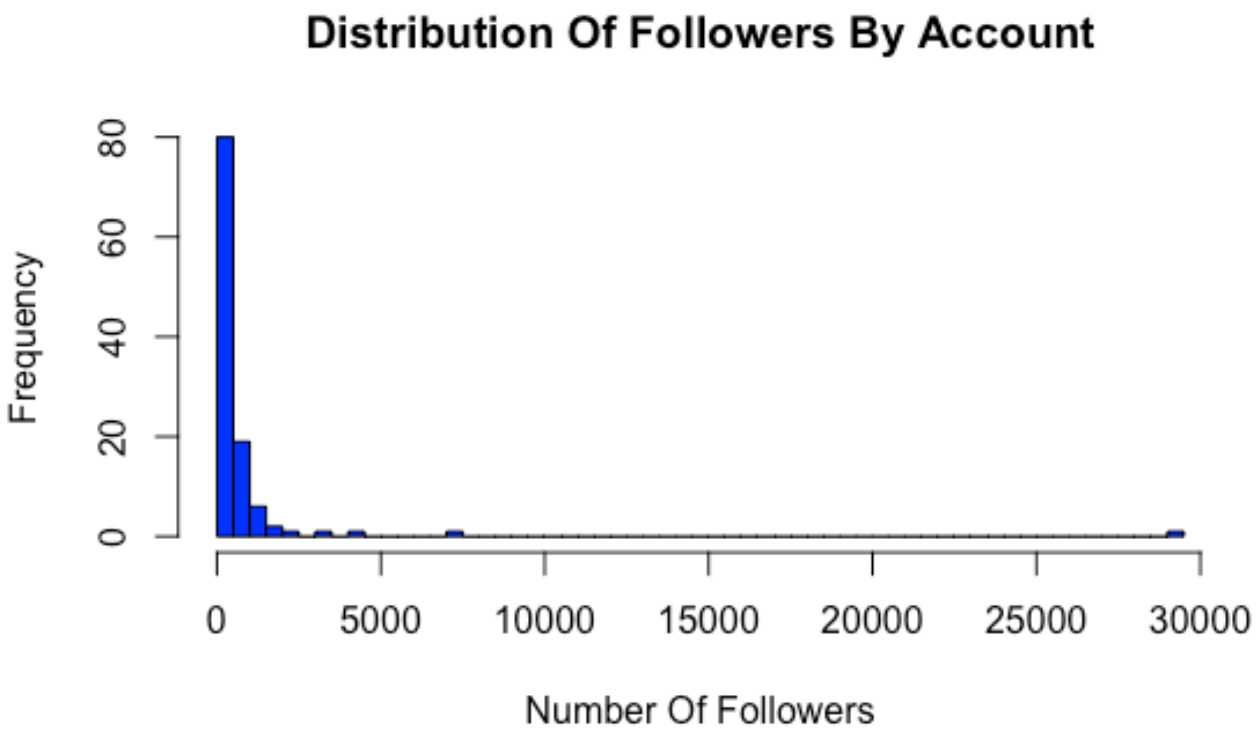
Excerpt from the collapsed data set

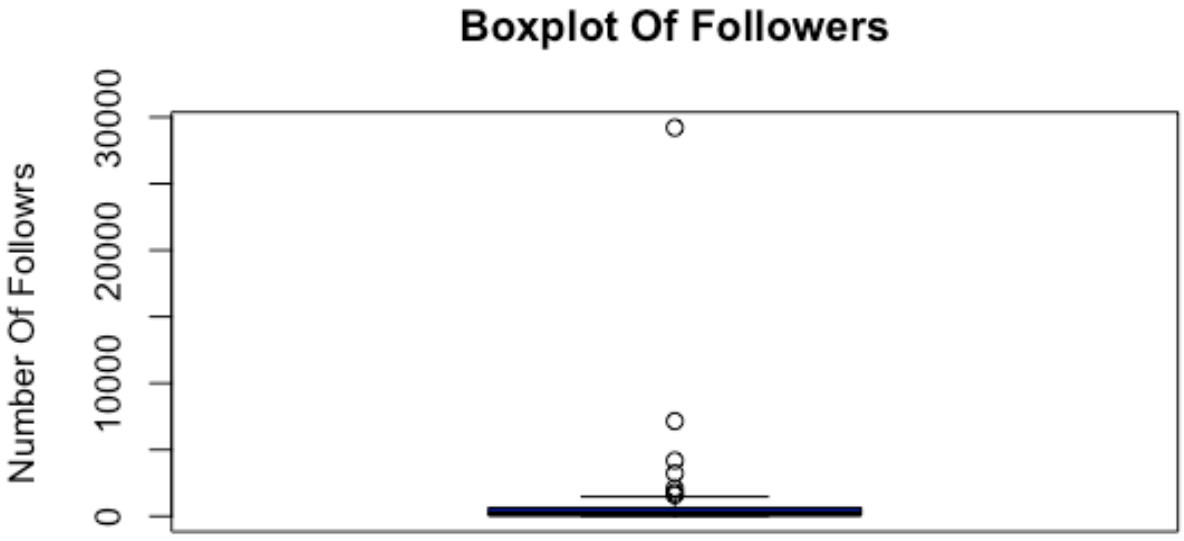
Now the data set is ready for further exploration. Running summary statistics on the data would yield the following.

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
16.0	87.5	210.5	767.3	642.8	29210.0

From summary statistics above we realize that the data is heavily positively skewed i.e. 75% of the accounts has lower than 642 followers, 24 accounts are above the mean of (767) and only 2 accounts enjoy more than 5000 followers.

In order to visualize that let's look at the histogram and box plot below:

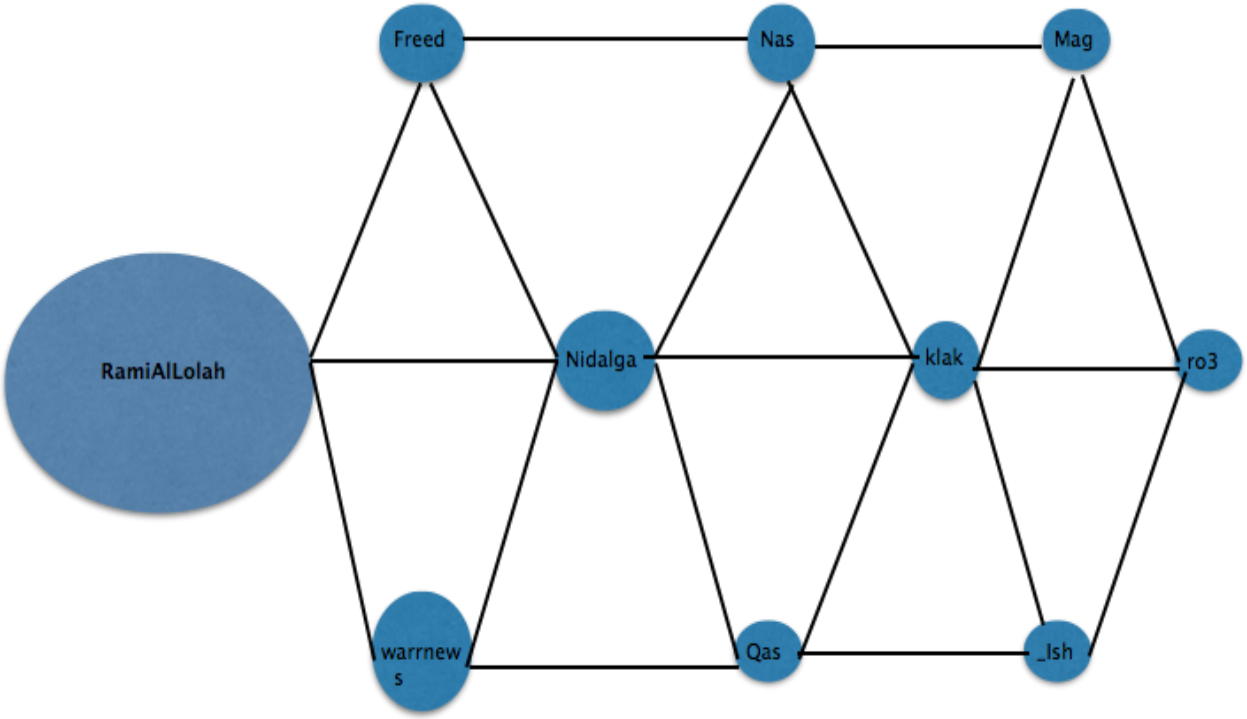

Distribution Of Followers By Account
Frequency
0
20
40
60
80
0
5000
10000
15000
20000
25000
30000
Number Of Followers


Boxplot Of Followers
Number Of Followrs
0
10000
20000
30000

Thus, few users carry disproportionate influence on ISIS's twitter network, the table and visualization below illustrate the top 10 influencers and their sizes scaled relative to their weights on ISIS's twitter network.

Top 10 influencers

Username	Weight
RamiAlLolah	0.33987666
warrnews	0.08322085
Nidalgazaui	0.04872004
Freedom_speech2	0.03761927
NaseemAhmed50	0.02468001
klakishinki	0.02084012
QassamiMarwan	0.01853619
MaghrebiHD	0.01708168
IshfaqAhmad	0.01631371
ro34th	0.01588317



Node size is based on relative weight

Analysis

Given the disproportionate influence of users in the top quartile, we are going to investigate common qualities about their tweets, specifically we are going to analyze tweets from accounts that scored above the average number of followers of 767, as this group represents 21% of the total number of accounts on ISIS's network, yet it enjoys approximately 80% of the total number of followers.

Before analyzing the tweets a number of preprocessing steps have to be taken:

1. Remove all punctuation and special characters from the tweets.
2. Remove numbers.
3. Turn all characters to lower case.
4. Remove "stopwords"; "stopwords" are common words (e.g., a, and, also, the, etc.) that are frequent and carry little analytic value.
5. Remove white space that was left by the removed words.
6. Stem the tweets. Stemming refers to removing common word endings (e.g., "ing", "es", "s") so that R recognizes the same word regardless of different forms of endings.

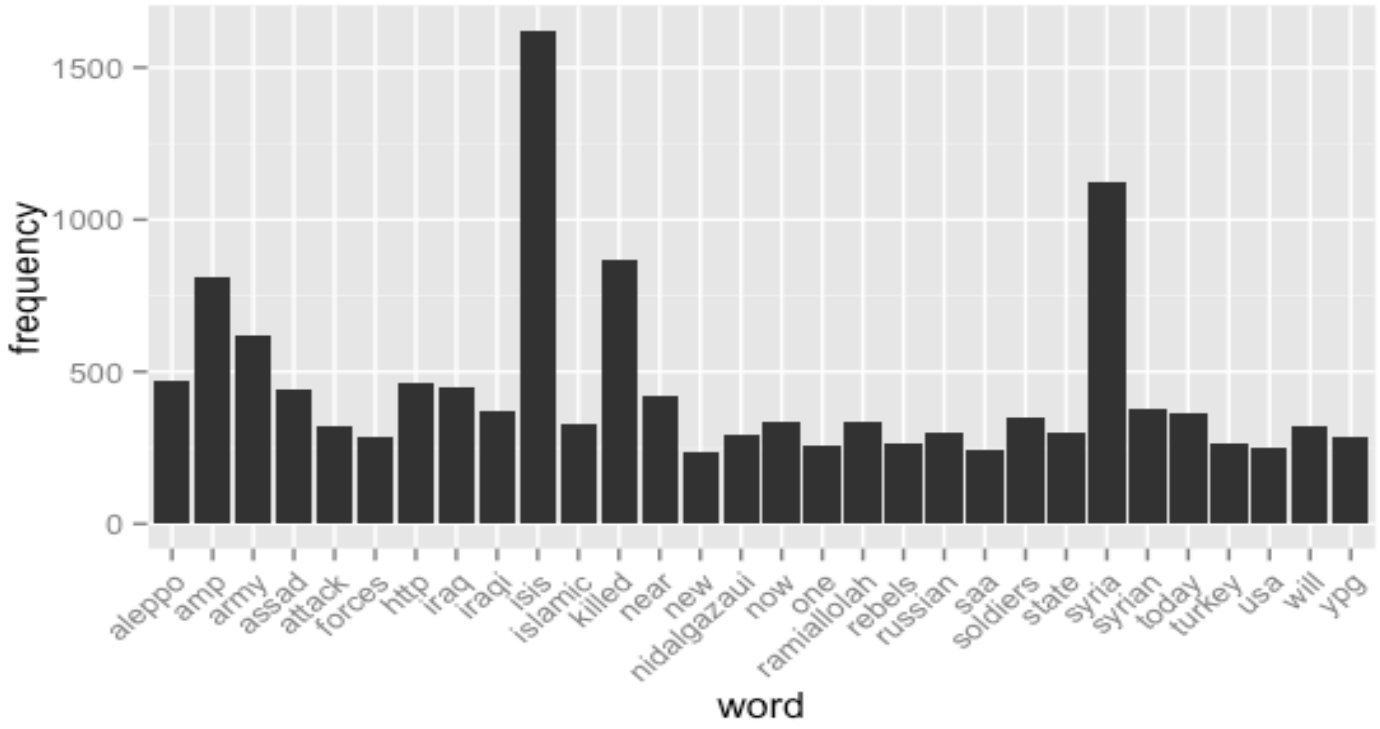
After taking these steps tweets are ready for processing, we are going to process 10,000+ stemmed tweets that were made over a period of approximately a year and a half in order to calculate, summarize and visualize the most frequent words used by the top influencers.

Results: 24,513 unique terms along with their frequencies were generated. Ranking those unique terms based on how frequently they appear in the tweets yielded the following list and visualization, which will show only a small subset of top words

Top 36 unique words:

Rank	Word	Frequency
1	isis	1623
2	syria	1122
3	killed	865
4	amp	811
5	army	617
6	aleppo	472
7	http	466
8	iraq	447
9	assad	439
10	near	419
11	syrian	376
12	iraqi	372
13	today	367
14	soldiers	347
15	now	332
16	ramiallolah	332
17	islamic	331
18	attack	323
19	will	318
20	state	301
21	russian	298
22	nidalgazaui	294
23	ypg	283
24	forces	282
25	turkey	264
26	rebels	262
27	one	257
28	usa	248
29	saa	244
30	new	239
31	palmyra	237
32	abu	232
33	people	228
34	north	227
35	breaking	226
36	video	219

The 30 most frequent words visualized

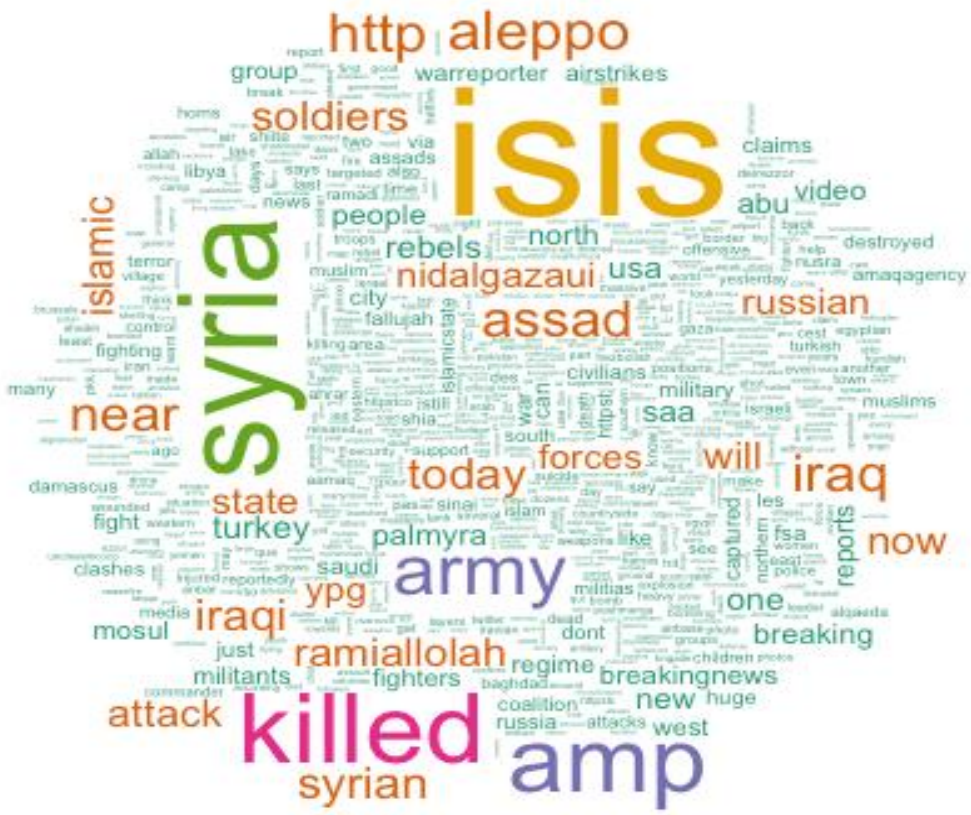


Top influencers exhibit the following patterns in their tweets:

1. They report heavily and promptly on the news, words like “today” “breaking” and “breaking news” are ranked 13th 35th and 38th respectively among a total of 24,513 unique words that were mined from the tweets.
2. Top influencers were tweeted at the most for example top influencers “ramiallolah” and “nidalgazaui” were ranked 16th and 22nd on the list of 24,513 words. This can be a reporting mechanism by which marginal players report to top influencers who in turn redistribute the content tweeted at them.
3. They distribute media in multiple forms; unique words such as “http” and “video” were ranked 7th and 36th.

What are ISIS's top influencers saying on twitter?

Most frequent words visualized

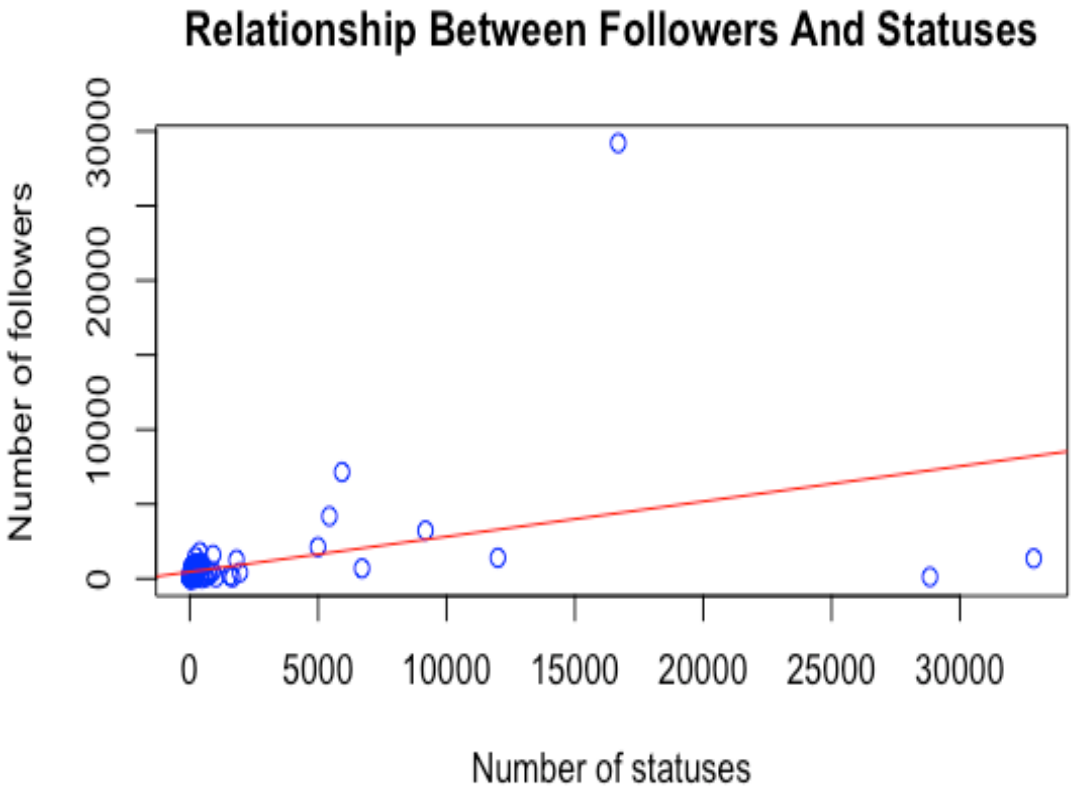


Unique words scaled according to their frequency

4. The message delivered by top influencers and conveyed in the tweets is inflaming–the word "killed" is ranked 3rd on the list of most frequent words.
5. The message is divisive as it focuses on ISIS's sectarian, religious and geopolitical enemies for example ("assad", "ypg" ,"russian","shia","usa" and "shiite") are highly ranked unique terms.

Model

One plausible hypothesis is that high level of activity on twitter would lead to high following. Let's take "*Number of statuses by the user when the tweet was downloaded"* as a shorthand for the level of activity on the website and build a linear model that regresses the "*Number of followers" on "Number of statuses".*



Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	459.1352	259.75773	1.768	0.0799
Num. OF Statuses	0.23620	0.05419	4.358	2.96e-05

Although it only explains 14% of the variation in number of followers, a statistically significant relationship exists between "Number of statuses" and "Number of followers". In other words, the biggest influencers on ISIS's twitter network are highly active tweeters relative to marginal influencers and that can partially explain their high number of followers.

Conclusion:

High following on ISIS's twitter network seems to be content as well as activity play; top influencers are avid and frequent tweeters, they break the news really fast, they employ a very targeted message that galvanizes ISIS's fan base and they seem to have a reporting mechanism that help keep them abreast on the latest development.

Knocking down top influencers such as RamiAlLolah and Nidalgazaui would represent a major hit to the network and its ability to redistribute news and propaganda. Also, understanding the messaging that this network employs and countering it promptly will reduce its recruitment effectiveness.

One important dimension that is lacking from the data set is the length of time a certain user has been on the platform. High following could be partially explained by being early on the platform establishing reputation and following among ISIS's "fanboys".

More sophisticated but time consuming techniques such as Clustering, SVD, MLR and PCA could be employed to further understand the dynamics of this network.