We Didn’t Start the Tweets

An Analysis of Disinformation Networks on Twitter

R Thompson

Network Science Presentation

Context

Data

Analysis

Next Steps

2016 saw a realization of the darker side of social media

- Previously, social media was seen as a great equalizer to potentially overturn authoritarian regimes
  - Arab Spring
  - Green Revolution in Iran

- That same power was turned towards destabilizing democracies

Battlefields of tomorrow will be on computer networks

- Cyber operations will be center to any great power competition

- Information operations have been injected with steroids
  - New field emerging of Computational Propoganda
  - Potential to take down governments without firing a shot or even hacking a network

Twitter removed state-backed accounts sowing disinfo

- Russia was the biggest source of coverage
  - Internet Research Agency
- Other countries that have also received attention include
  - Iran
  - Saudi Arabia
  - Venezuela
  - Qatar
- Interested in looking at Iran in addition to Russia as there have been media reports that Russian hackers have posed as Iranian hackers previously
- IMPORTANT TO NOTE: removals are focused on accounts misrepresenting who they are, spreading disinformation AND associated with state-backed entities

Context

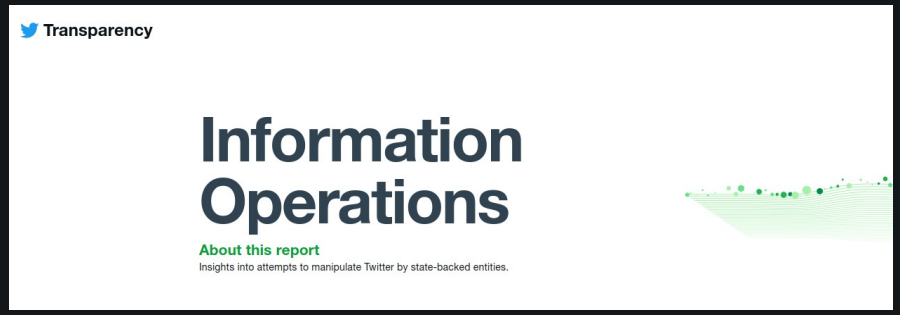
Data

Analysis

Next Steps

“In line with our principles of transparency and to improve public understanding of inauthentic influence campaigns, Twitter is making publicly available archives of Tweets and media that we believe resulted from state-backed information operations on our service.”

Twitter has released extensive amounts of data



- All taken from Twitter's Information Operations
- Data sets were collected by Twitter between October 2018 and October 2020
- Each month accounts identified are segmented by country into different files
  - Users (removed accounts)
  - Tweets (what they said)
  - Media (images they shared)
- UserIDs are hashed in many cases to protect identities
  - Tweet information tell us who interacted with these accounts' tweets

Analysis focused mainly on the combined network

	Russia		Iran		Total	
	Core	Total	Core	Total	Core	Total
Nodes	5,185	1,158,351	6,788	414,056	11,973	1,572,407
Edges	201,576	2,571,286	34,416	1,417,174	235,992	3,988,460

Definitions	
Core	Accounts that have been flagged in the users file
Total	All accounts associated with all interactions (accounts may still be active)
Edges	An edge exists if there is an interaction between accounts, retweet, like, or reply. Weights are determined by the total interactions. Direction from account posting to account interacting

Context

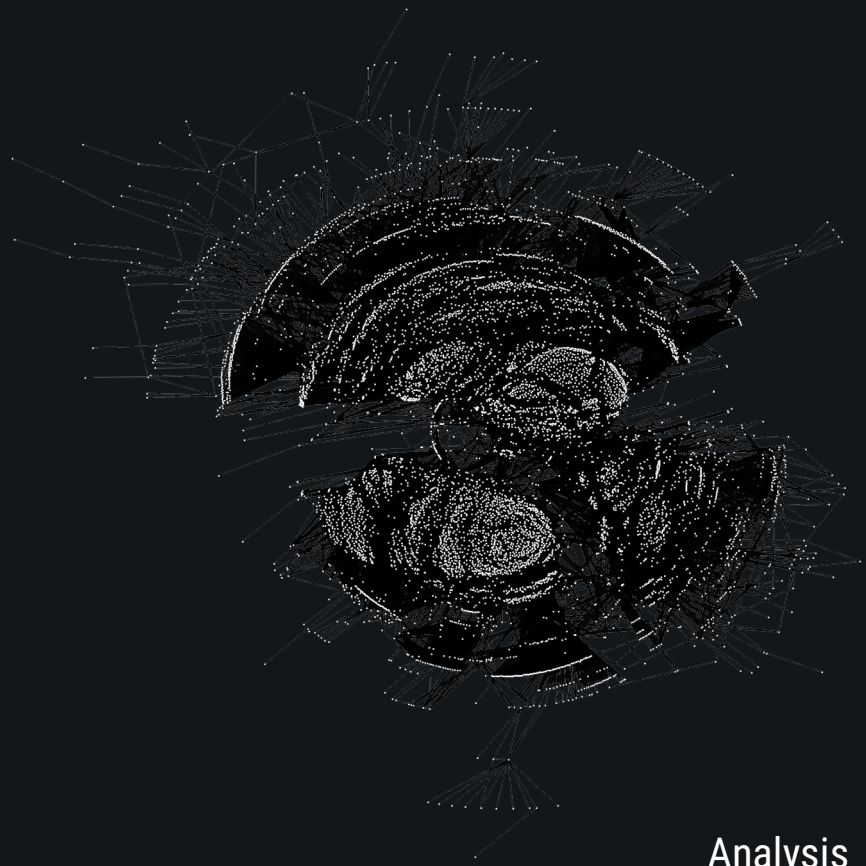
Data

Analysis

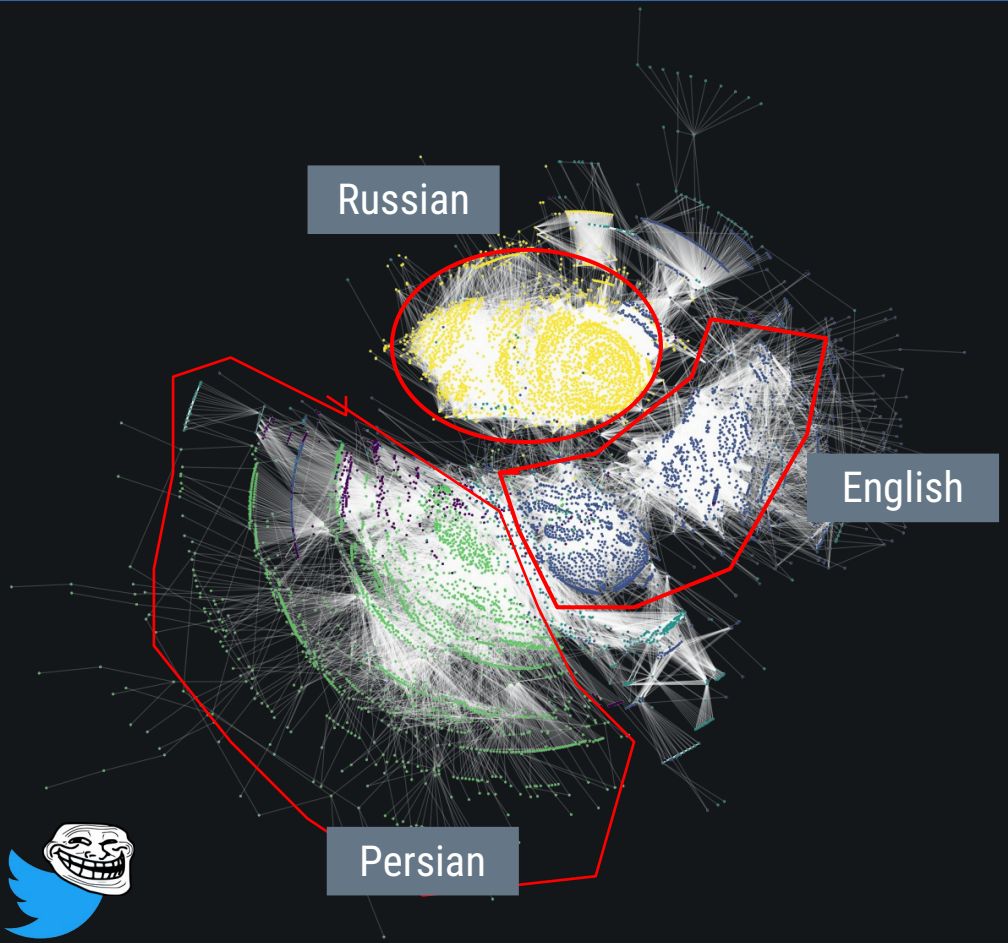
Next Steps

BLUF: Russian accounts were more effective at messaging

- While Russia had less core nodes, their reach was clearly much further when you look at the accounts that interacted with their material
- Russian network had more distinctive communities
  - Partially explained by primary language of activity
  - Potentially could be content/issue driven
- Russian activity was generally “tighter” than Iranians
  - Median clustering coefficient was nearly 17x bigger

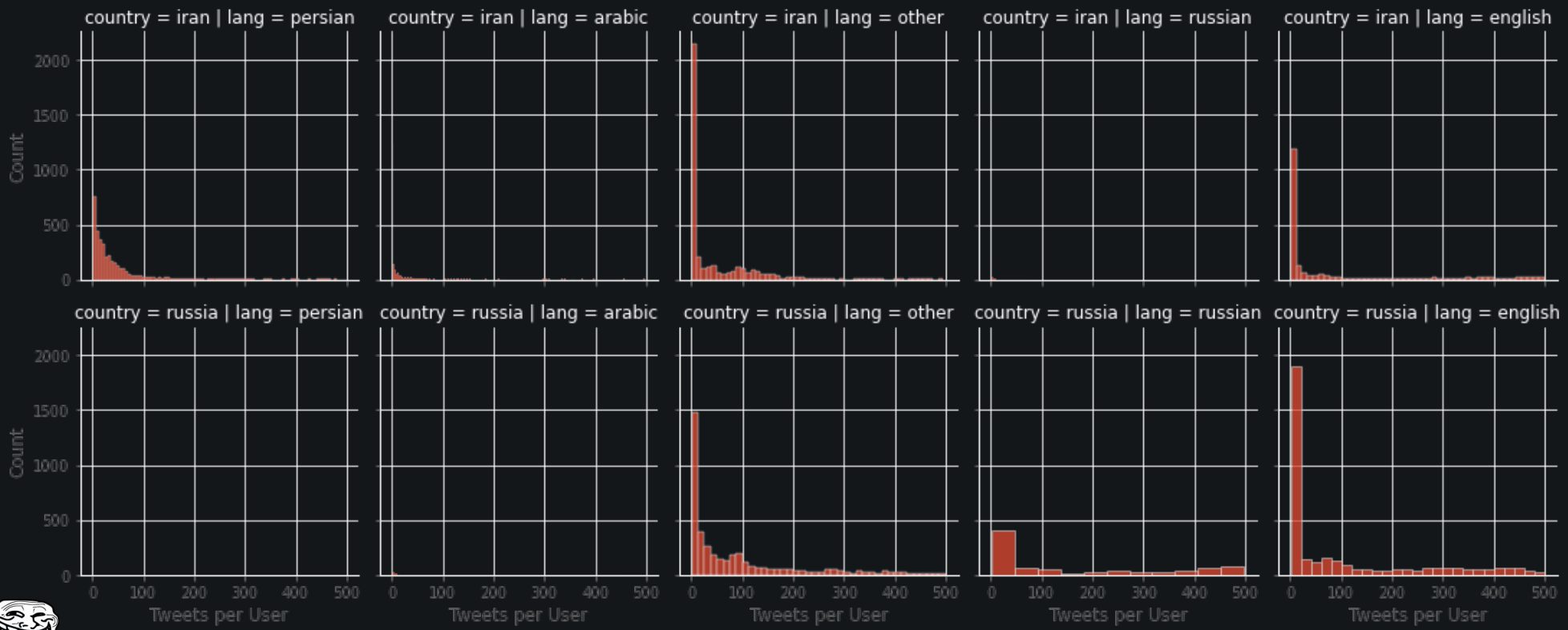


Russian backed accounts were in Russian & English



- Accounts that primarily tweeted in Russian and English can be traced back to Russian backed accounts
- English accounts more sparse
- Iranian accounts mainly tweeted about domestic issues as well as regional
  - Small number of Iranian run accounts primarily tweeted in Arabic

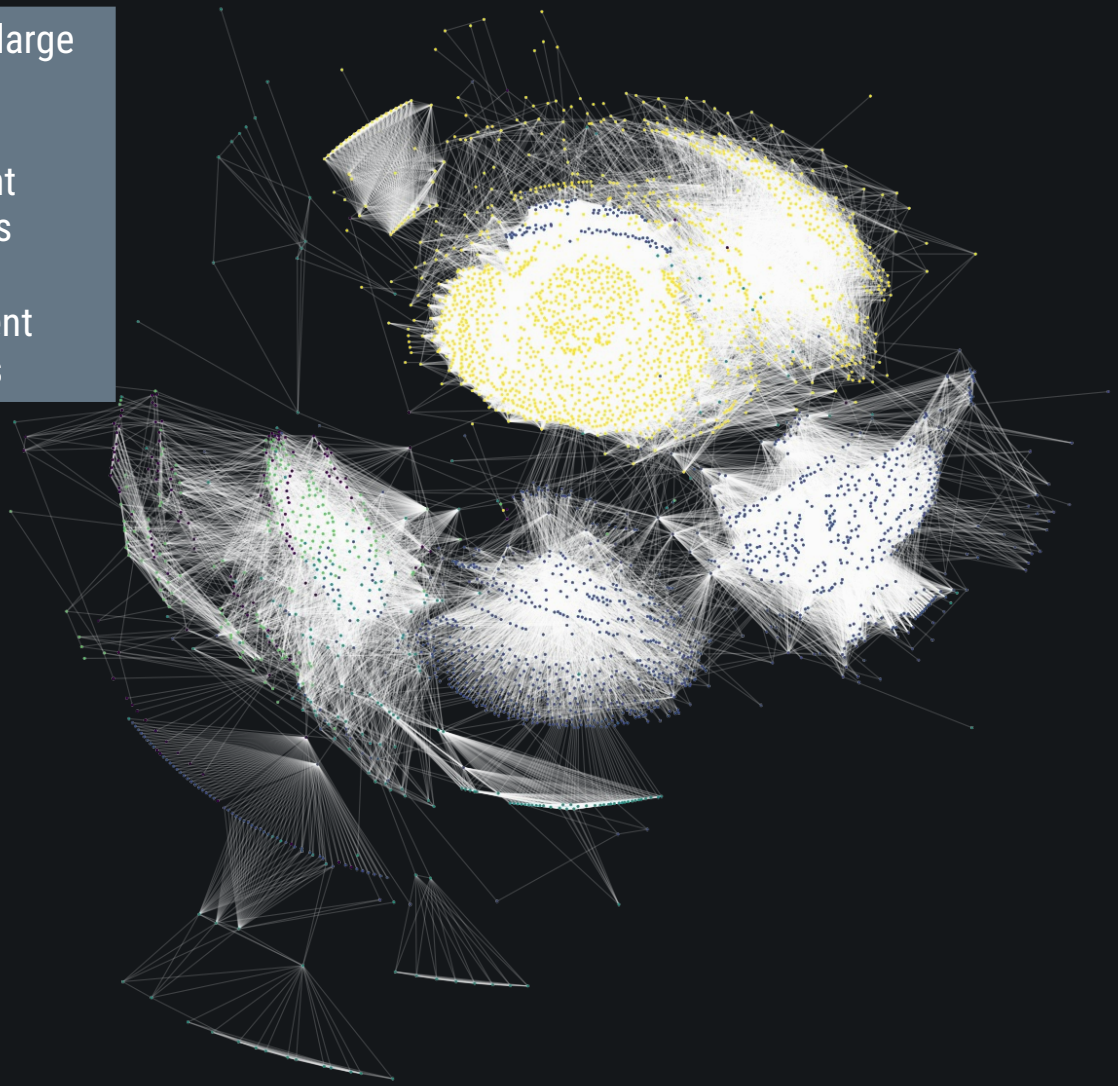
Many accounts produced only a small number of tweets



Subset of Nodes produced a large amount of content

92% of Iranian-backed content produced by 10% of accounts

95% of Russian-backed content produced by 22% of accounts



Core Communities are distinct between Russia and Iran

Top 10 Communities		
Community	Country	Nodes
1	Iran	862
2	Iran	1302
3	Iran	899
4	Iran	366
5	Russia	949
6	Iran	116
7	Russia	937
8	Russia	867
9	Russia	369
10	Russia	169

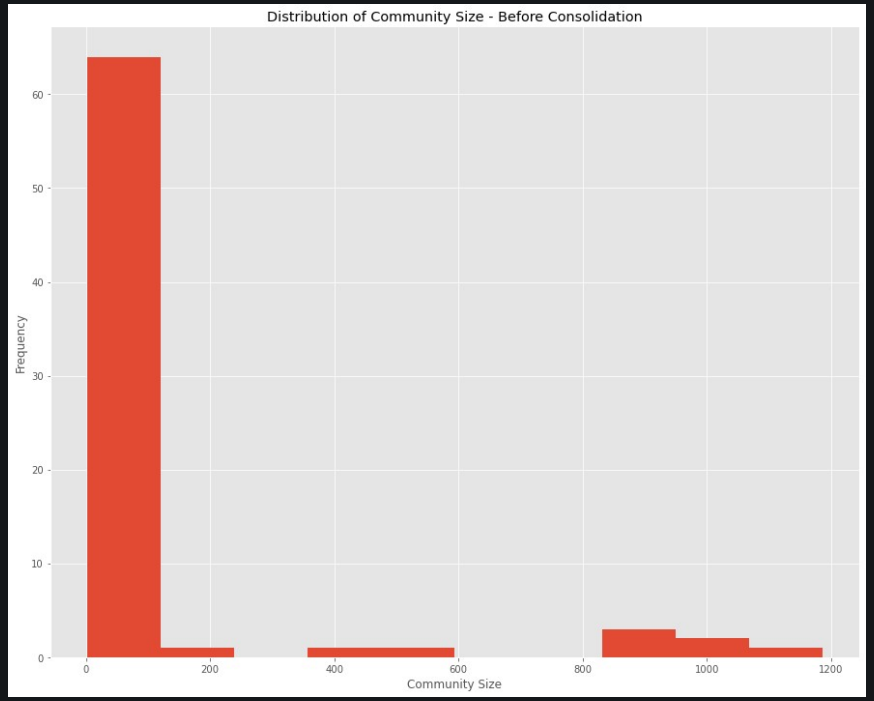
- Using Louvain Community Detection

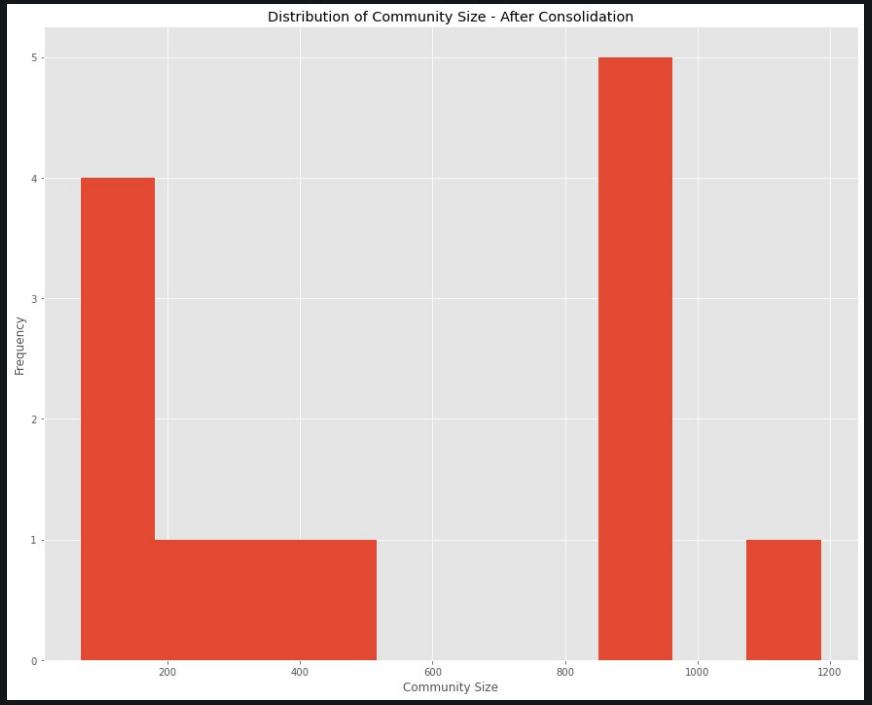
$$Q=\frac{1}{2m}\sum_{ij}\left[A_{ij}-\frac{k_i k_j}{2m}\right]\delta(c_i,c_j)$$

- Shown to be effective with Twitter from Pujol et al *Divide and Conquer*

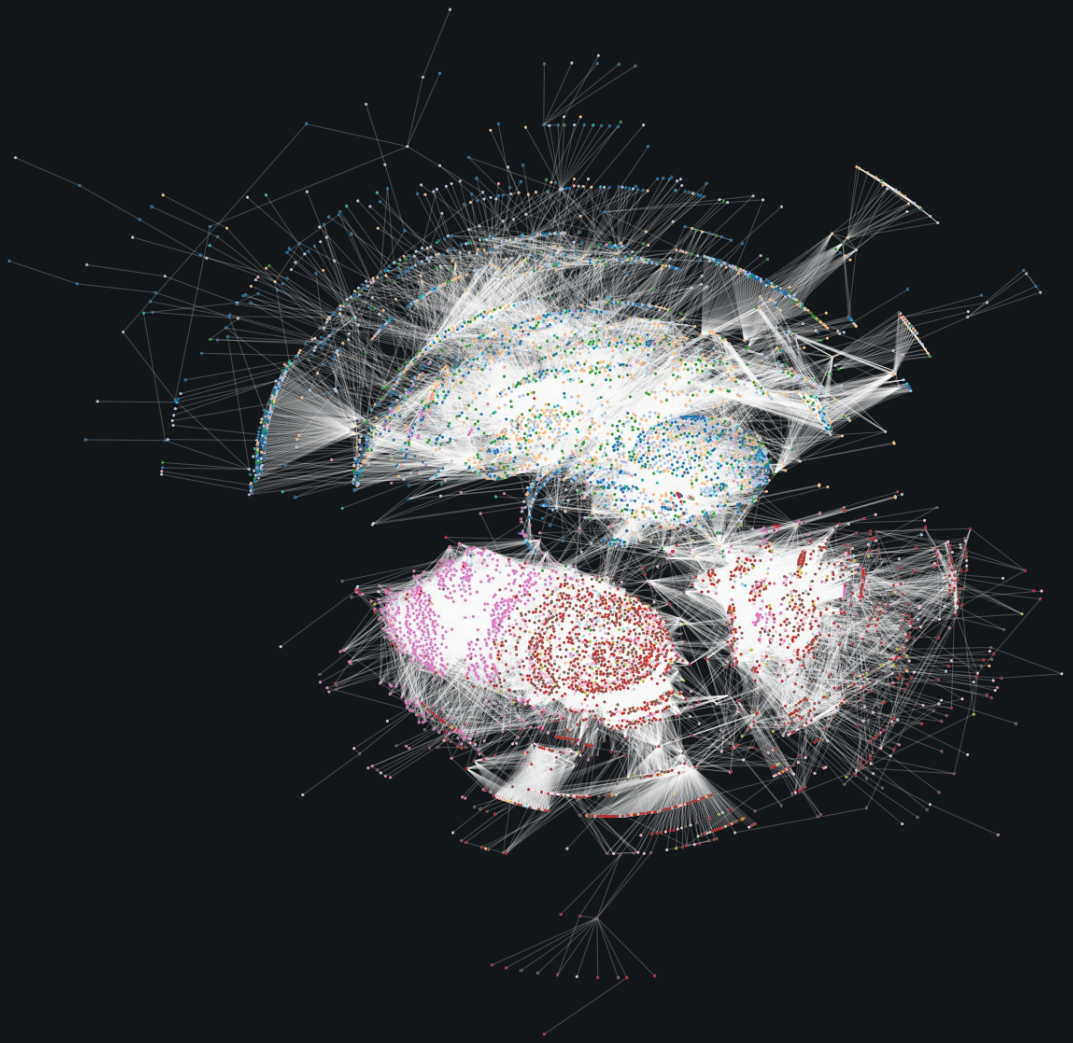
- Run on it's own over 170 "communities", but many are under 5 nodes
  - Consolidated groups that had under 50 nodes by country of origin to create additional group
  - 14 consolidated communities
  - Only two communities had any crossover (less than 10)

Consolidation of smaller groups improved distribution





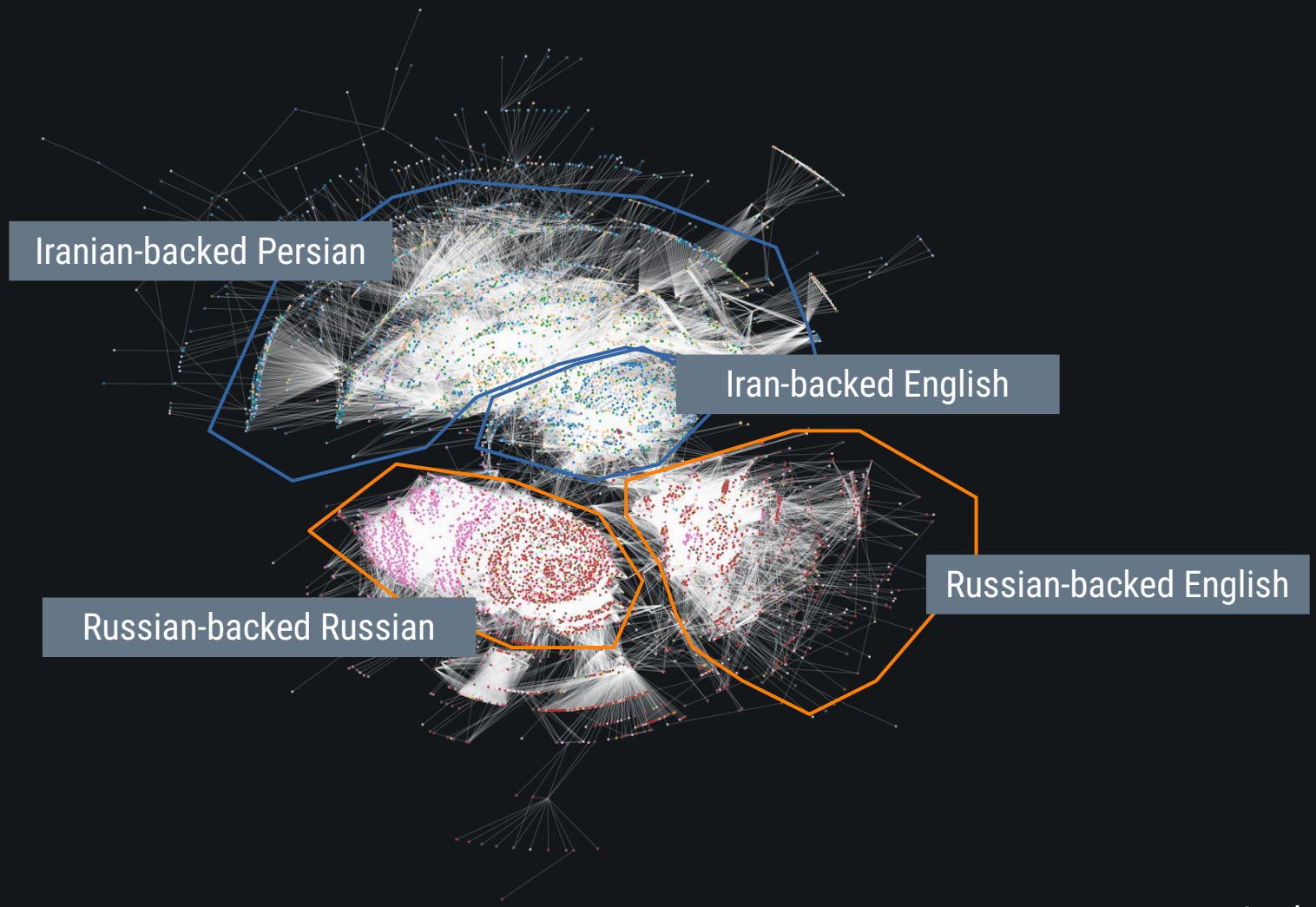
Russian & Iranian Suspended Twitter Accounts Grouped by Community



*I removed the self-connected only nodes from the visualization

Analysis

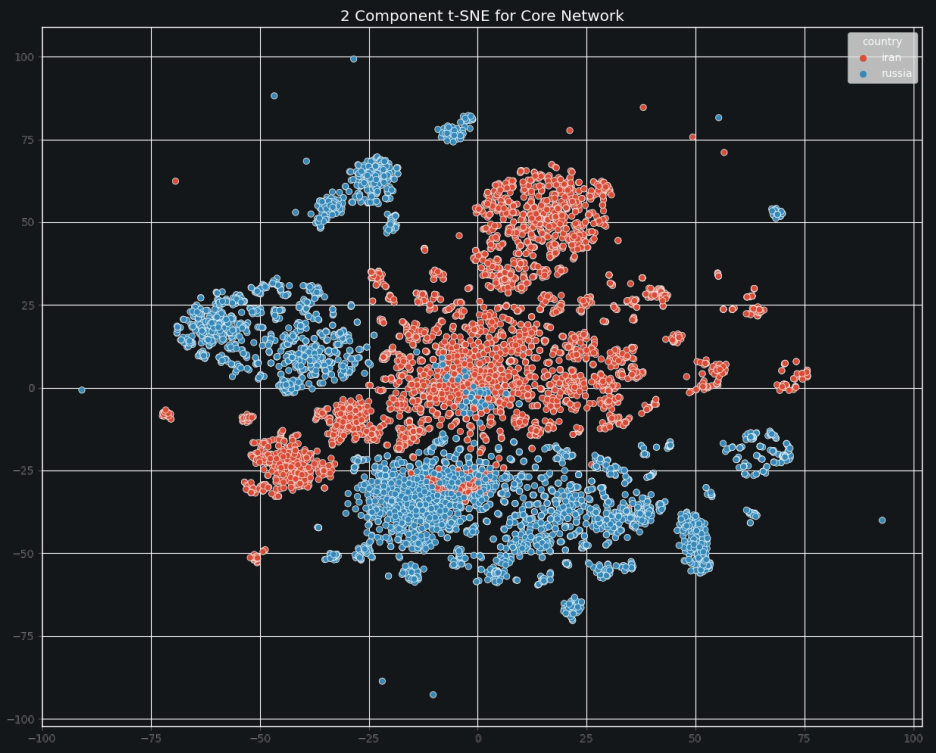
Russian & Iranian Suspended Twitter Accounts Grouped by Community



*I removed the self-connected only nodes from the visualization

Other methods for clustering revealed similar results

- Experimented with similarity measures such as:
  - SimRank
  - Personalized PageRank
  - DSD
- Results then reduced to 2 and 3 components with Principal Component Analysis and t-distributed Stochastic Neighbor Embedding
- Russian accounts formed more distinctive groupings compared to Iranians

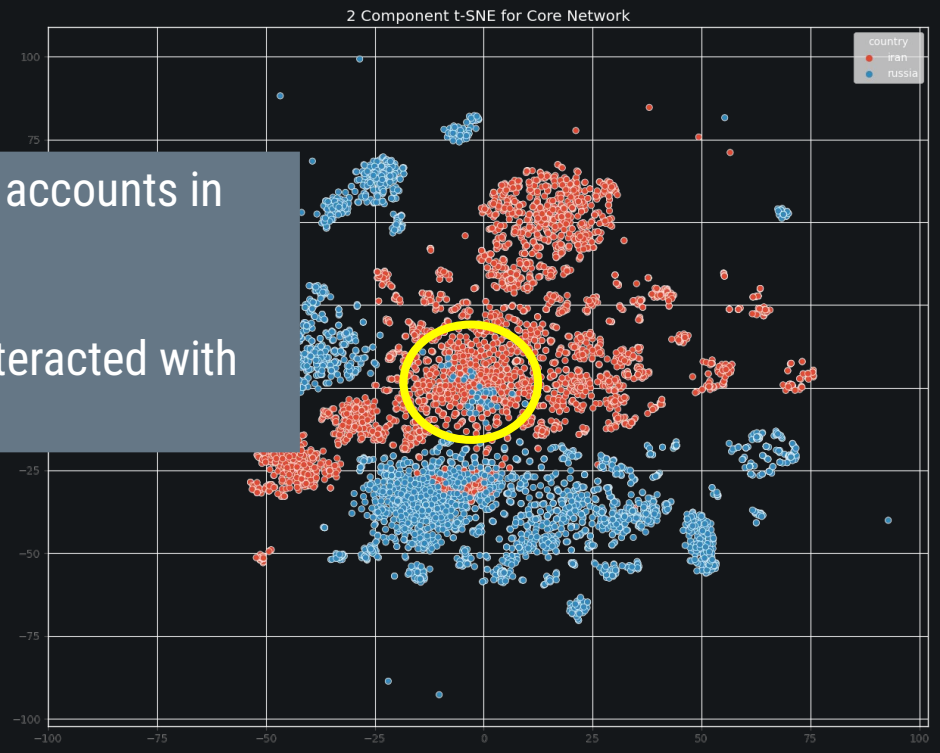


Other methods for clustering revealed similar results

- Experimented with similarity measures such as:
  - SimRank
  - Personalized PageRank
  - DSD
- Results then reduced to Principal Component Analysis and t-distributed Stochastic Neighbor Embedding
- Russian accounts formed more distinctive groupings compared to Iranians

Curious about these Russian accounts in the Iranian network

Initial analysis shows they interacted with the Iranian accounts



Context

Data

Analysis

Next Steps

More to be done, but better analytical methods needed

- Running analysis on the whole graph proved...difficult
  - Memory errors abound!
  - Best was t-SNE: Unable to allocate 17.1 TiB
- Experimented with Google Cloud servers, but some metrics took days on end to run
  - Personalized PageRank using core network to model potential troll accounts took 3 days to get half way through...then it failed
- Tried Spark/Scala on local machine before attempting cluster approach
  - GraphFrames has languished
  - GraphX has limited algorithms implemented
  - Custom implementation was trickier than time allowed

Focusing on core network behavior

- Come up with some potential explanations for the clustering that has been found
  - Language of tweets
  - Hashtags as proxy for issues
  - Tweet content – for future studies
- Are the clusters consistent with each other?
  - Curious to see what nodes continue to stick together
  - Additional dimensionality reduction work
- Explore centrality of some nodes, are there clusters that are more central?
- Also using alternative method to model potential troll accounts with random subgraphs
  - Train logit model on sample of troll and unknown accounts

Next Steps

Other ideas for what’s next?