

# Proposal: Building A System To Visualize and Analyze Misinformation Communities on Twitter with Domain Knowledge

Ammar Ahmed  
ammara2@illinois.edu  
Computer Science, MCS

Roy Chiu  
rchiu7@illinois.edu  
Computer Science, MCS

Kathleen Isenegger  
kti3@illinois.edu  
Computer Science, PhD

Devin Tark  
devinv2@illinois.edu  
Computer Science, MCS

October 2, 2020

## 1 Background

Numerous studies have attempted to analyze and understand the network structures that comprise modern forms of communication, including social media platforms like Twitter. Prior investigations on these social networks range from analyzing reciprocity, user interactions, user communities, or posts and have yielded practical methods for understanding social media networks. For example, there have been a handful of studies on information dissemination on social media between credible and non-credible sources and discovered distinguishing characteristics to classify the two sources. Yet, many of these studies focus on the broader attributes of these networks rather than focus on specific network communities. Many network analysis tools, such as Gephi, also focus on generalizing metrics across any graph or network. Building a tool that can offer similar features as these network analysis tools but geared towards these specific, untouched communities will be helpful in uncovering and further understanding these networks and the patterns behind them.

## 2 What Problem Are We Trying to Solve?

The problem we are trying to solve is the lack of interfaces for analyzing specific misinformation communities. This is an important issue because misinformation is becoming extremely common on many social media platforms. Many researchers find benefit in analyzing misinformation networks through graphical analysis due to social networks' general adherence to network characteristics. Regular social media users may want to monitor the topics and key users in these networks to be aware of the most prevalent misinformation themes and sources. The vastness of social media sources requires an adequate set of tools for analyzing any particular subspace. While there are plenty of advanced search tools available for social networks, there is no top-down, whole-network view that allows users to view a misinformation community in near real time that is also content and topic-aware. We are designing a dashboard that will be specifically built to allow users and researchers to get a big-picture view for their analysis of misinformation on social media.

### 3 Why is this Useful and Cool?

There is ambiguity around how people become dedicated supporters of misinformation. One method of recruitment known to be used by QAnon, a large conspiracy theorist group, is through the use of the hashtag “SaveTheChildren”. They covertly open up audiences to their conspiratorial messages by first establishing contact with this seemingly mainstream social issue of child trafficking. Our dashboard will analyze the trending topics within misinformation communities to offer greater transparency about the origin of certain viral topics, messages, and/or hashtags to follow the trends of the recruitment technique. This will be useful to researchers who are interested in how these communities grow. It will also be useful to casual social media users as they can be more aware about the role misinformation communities play in the content they may see and it can help them to make informed decisions about what they believe or share. Additionally, it may be useful to social media moderators to clearly see these pockets of misinformation communities when considering accounts to ban due to malicious political manipulation, a problem which is becoming more and more common. It will also allow those who study misinformation communities to get a one-stop-shop big picture view of the community at large.

### 4 Feasibility & Proposed Approach

This will be a web-based tool that utilizes a robust framework of software to allow its implementation within our limited time frame. The Twitter API will be used in conjunction with an AWS server to create an automatic process of streaming tweets into our database. Two of our team members already have applied and been accepted for developer access. In terms of tools, we plan to use a graph-based database like Neo4j to store most if not all of the data and relationships we collect, and to built-in graphical analysis algorithms. We plan to use React to help us create a dynamic and stateful front-end; Python will be used for the back-end calculations and analysis. Flask will likely provide us the API framework. To analyze the topic similarity of Q drops with network tweets, we will utilize natural language processing to uncover major entities embedded within documents and compare the similarity between those within tweets and those from the Q poster themselves.

Twitter network analysis, in general, is an immensely popular computing / social media analysis topic, due in part to its robust API and (mostly) public content. Therefore, we assess that collecting enough good quality data to perform the analysis and visualization tasks is very feasible. In short, our tool at a minimum will:

1. Visualize either the entire or a sample of the Qanon network on Twitter
2. Display Network metrics like clustering, connectedness, activity, distribution of in-degree and out-degree relationships
3. Display top users by 4 main metrics: Followers, HITS, Influence-Passivity, Topic-Based PageRank
4. Display trending topics within this community
5. Compare trending topics on Twitter to trending topics in Q messaging

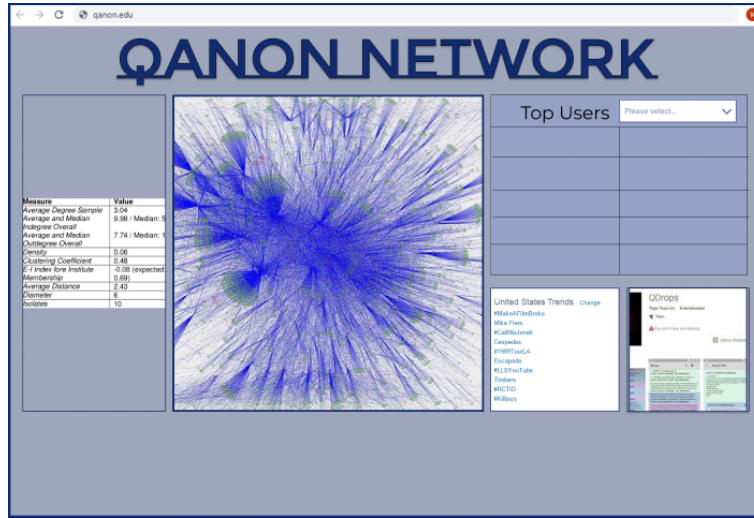


Figure 1: Major Functionality of our Tool

## 5 Literature Survey Overview

The question our literature will seek to answer is “What metrics are used to measure influence in a network and what tools exist to analyze and visualize online communities with these metrics?”

This is interesting to us as a group because it is at the core of what problem we are trying to solve. By studying what metrics already exist, and the shortfalls of currently existing tools, we can build a useful tool that builds on existing domain knowledge.

In general, this is useful and cool for the same reason our system is. Online communities, and the misinformation that passes through them has become remarkably influential in the past few years. Traditional forms of media restricted influence to only those very few that were on TV, radio, or could make it into newspapers. There were gatekeepers to who could reach the masses, how and when. Now, social media allows regular users to reach more people than ever before regardless of previous forms of celebrity or fame, and we therefore need new tools and metrics to track this from a big picture perspective.

### 5.1 Initial Literature List

*See References*

## 6 Plan

### 6.1 Primary Team Responsibilities

Devin Tark	-	Captain + Data Processing
Ammar Ahmed	-	Backend + Data Processing
Roy Chiu	-	Frontend + API
Kathleen Isenegger	-	Frontend + Visualizations

## 6.2 Timeline of Deliverables

Date	Item
October 9, 2020	Wireframe Diagram
October 16, 2020	Data Collection / Infrastructure
November 13, 2020	Initial working Dashboard Prototype(frontend)
December 9, 2020	Final Working Dashboard System
December 9, 2020	Peer Reviews
December 9, 2020	Final Project Report

## 6.3 Grading Rubric

1. Wireframe Diagram (5%)
  - The wireframe is clear and organized
  - The wireframe clearly and simply identifies all planned functionality of the tool
2. Data Collection / Data Infrastructure (10%)
  - There exists a store of data collected that includes at least 1000 tweets and at least 200 users
  - We have a complete pipeline process for automatically collecting future / more tweets
3. Initial Working Dashboard Prototype (10%)
  - The group can demonstrate, locally, a working front end interface that displays the core functionality of the wireframe/proposal, either with or without data.
  - The front end prototype clearly conveys what information *will be* available in the final product
4. Final Tool Demonstration (15%)
  - The team can demonstrate, locally, a working front end that displays the core functionality of the proposal
  - The tool provides useful and relevant analysis of data in accordance with the proposal
  - The tool provides functionality close to or similar as defined in Section 4 of this proposal
5. Project Report (5%)
  - The team provides a final report, 2-3 pages in length describing the final product and the process in building it
  - The report covers successes, failures, challenges, architecture, and future work
6. Peer Review (10%)
  - Each team member provides a rating out of 10 for each other team member based on their **effort and contribution** to the project and team

## References

- [1] M. Smith, B. Schneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave. Analyzing (Social media) Networks with NodeXL. In *CT '09: Proceedings of the fourth international conference on Communities and technologies*. 255-264. (2009).
- [2] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. Lopez Segui. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *Journal of Medical Internet Research*. 22(5). 1-9, (2020)
- [3] C. Shao, P.M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, G.L. Ciampaglia. Anatomy of an online misinformation network. *PLOS ONE* 13(4). (2020)
- [4] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, S. Havlin. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*. (2020)
- [5] L. Wu, F. Morstatter, K. Carley, and H. Liu. Misinformation in Social Media: Definition, Manipulation, and Detection. *ACM SIGKDD Explorations Newsletter*. 80-90. (2019)
- [6] F. Chowdhury, L. Allen, M. Yousuf, and A. Mueen. On Twitter Purge: A Retrospective Analysis of Suspended Users. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion)*. 371-378. (2020)
- [7] Vosoughi, S., Roy, D., Aral, S. The spread of true and false news online. *Science* 359(6380): 1146-1151 (2018)
- [8] A. Pitts. Polinode: A web application for the collection and analysis of network data. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 1422-1425. (2016)
- [9] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli. Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *PloS one*. (2015)
- [10] T. Schatto-Eckrodt. Use and assessment of sources in conspiracy theorists' communities. *Multidisciplinary International Symposium on Disinformation in Open Online Media*. 25-32. (2019)
- [11] E. Adar. GUESS: a language and interface for graph exploration. *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 791-800. (2006)
- [12] C. Shao. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16 Companion*. 745-750. (2016)
- [13] M. Bastian and S. Heymann. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third International ICWSM Conference*. 361-362. (2009)
- [14] A. Bruns. HOW LONG IS A TWEET? MAPPING DYNAMIC CONVERSATION NETWORKS ON TWITTER USING GAWK AND GEPHI. *Information, Communication Society*. 1323-1351. (2012)
- [15] J. Tang, M. Musolesi, C. Mascolo, and V. Latora. Temporal distance metrics for social network analysis. *Proceedings of the 2nd ACM workshop on Online social networks*. 31-36. (2009)