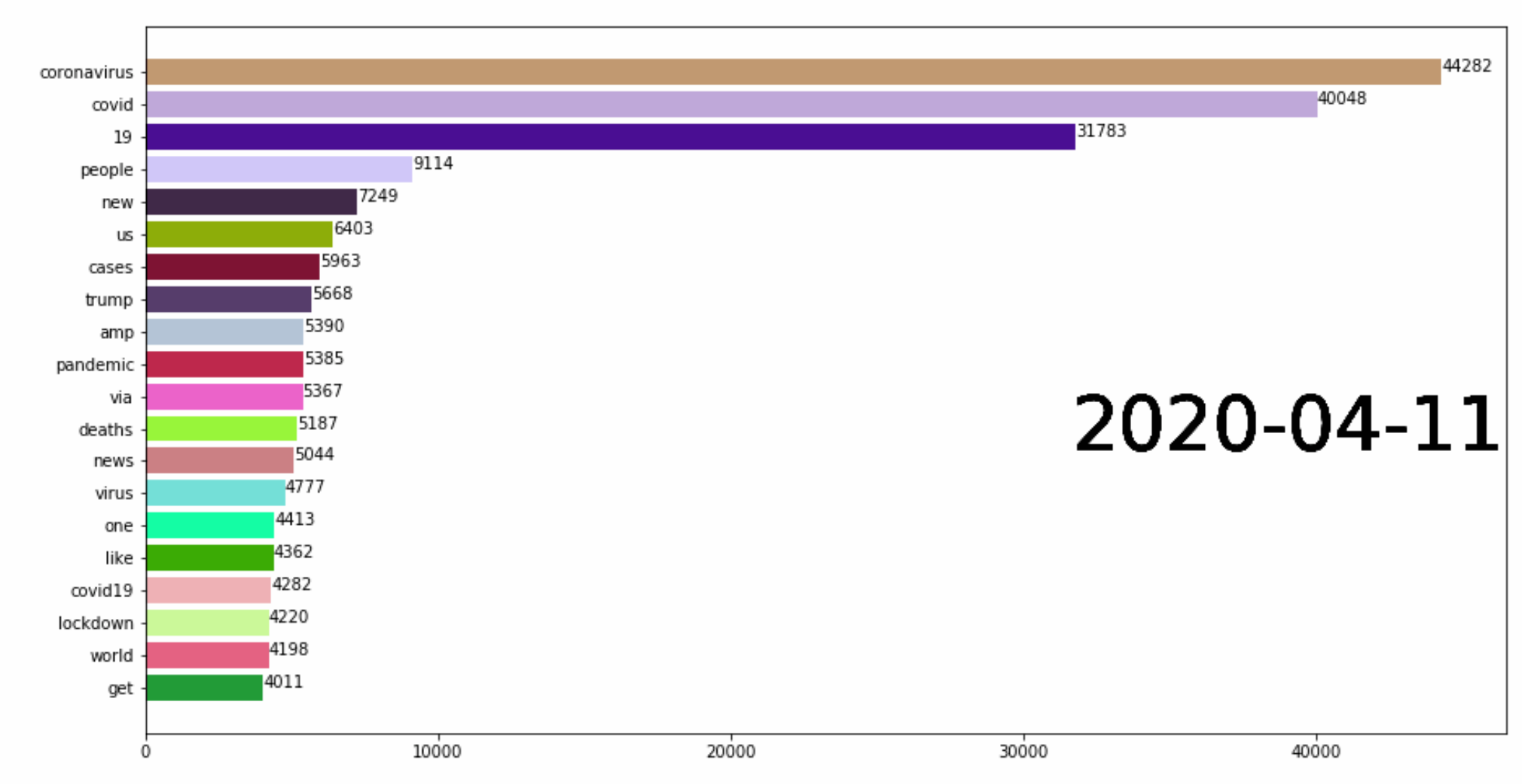
TACC COVID-19 Twitter Dataset Enables Social Science Research about Pandemic

Leading supercomputing center partners with scholars to collect, distribute, and analyze social media communications



An animation showing the top 20 daily n-grams (common words in Twitter post) changing overtime.

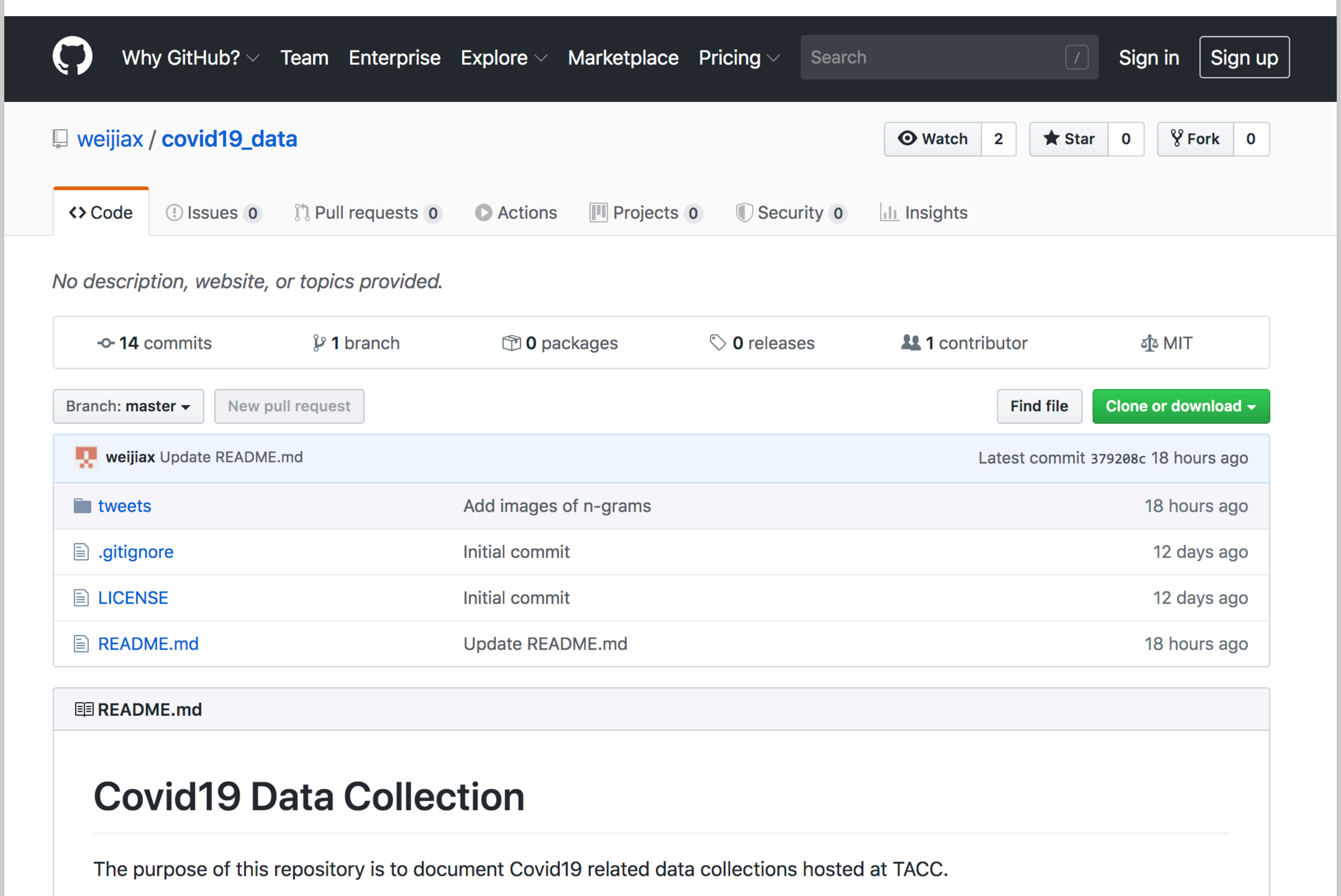
Of the myriad ways researchers are fighting the spread of the coronavirus, studying Tweets may not be the first that come to mind. But now, as in past crises, tapping into one of the world's leading real-time messaging service can help identify new pandemic hotspots, highlight new symptoms, or interpret how people and communities are responding to orders to practice social distancing.

The Texas Advanced Computing Center (TACC)'s expert data science team has facilitated social media analysis in the past, and has developed machine learning tools to better pull needles of insight out of the vast haystacks of the Twitterverse.

Starting in March, TACC began ingesting large amounts of tweets daily — roughly 40 million messages, of which one million are unique. Combining their collection with similar efforts from groups at UT Austin, the University of Southern California, and Georgia State University, they have extended their collection of COVID-19 related tweets back to January. (Last week, Twitter announced that it would be releasing new API endpoints to its own COVID-19 related tweets collection for approved developers and researchers.)

"There's a large amount of interest in these types of collections. It's very useful in data science," said Weijia Xu, who manages the Scalable Computational Intelligence group at TACC.

Today, TACC announced a new GitHub repository where interested researchers can access both pointers to raw Twitter data related to COVID-19 and large-scale analyses facilitated by TACC's supercomputers.



No description, website, or topics provided.

14 commits · 1 branch · 0 packages · 0 releases · 1 contributor · MIT

weijiax Update README.md — Latest commit 379208c 18 hours ago

tweets	Add images of n-grams	18 hours ago
.gitignore	Initial commit	12 days ago
LICENSE	Initial commit	12 days ago
README.md	Update README.md	18 hours ago

README.md

Covid19 Data Collection

The purpose of this repository is to document Covid19 related data collections hosted at TACC.

The "Covid19 Data Collection" Github page allows researchers to access pointers to a large dataset of COVID-19-related tweets and analyses of the collection.

The first of the analyses available to researchers is a set of n-grams: contiguous sequences of words from a given sample of tweets. The top 1,000 one-, two-, and three-word sequences have been assembled for each day of the pandemic. Assembling even a single 1-gram from several million tweets could take up to an hour on a laptop due to the amount of data processing involved, but can be done in minutes on TACC's supercomputers.

The TACC research team, led by Xu, has also been working on topic modeling analyses, identifying terms that frequently appear in connection with each other, though not necessarily in order. These will be added to the GitHub repository in the coming weeks.

Both methods of clustering can be helpful in identifying trends in how the pandemic, and people's response to it, are evolving.

Future projects using the data include a searchable public database; entity analysis — inspecting tweets for known entities such as public figures or organizations and returning information about those entities; and event detection — automatically detecting the occurrence of events and categorizing them.

These efforts will be facilitated by tools developed at TACC, like the Domain Information & Vocabulary Extraction project, a National Science Foundation-funded effort to extract biological entities from publication and other text documents using machine learning, which has been adapted for other types of extraction.

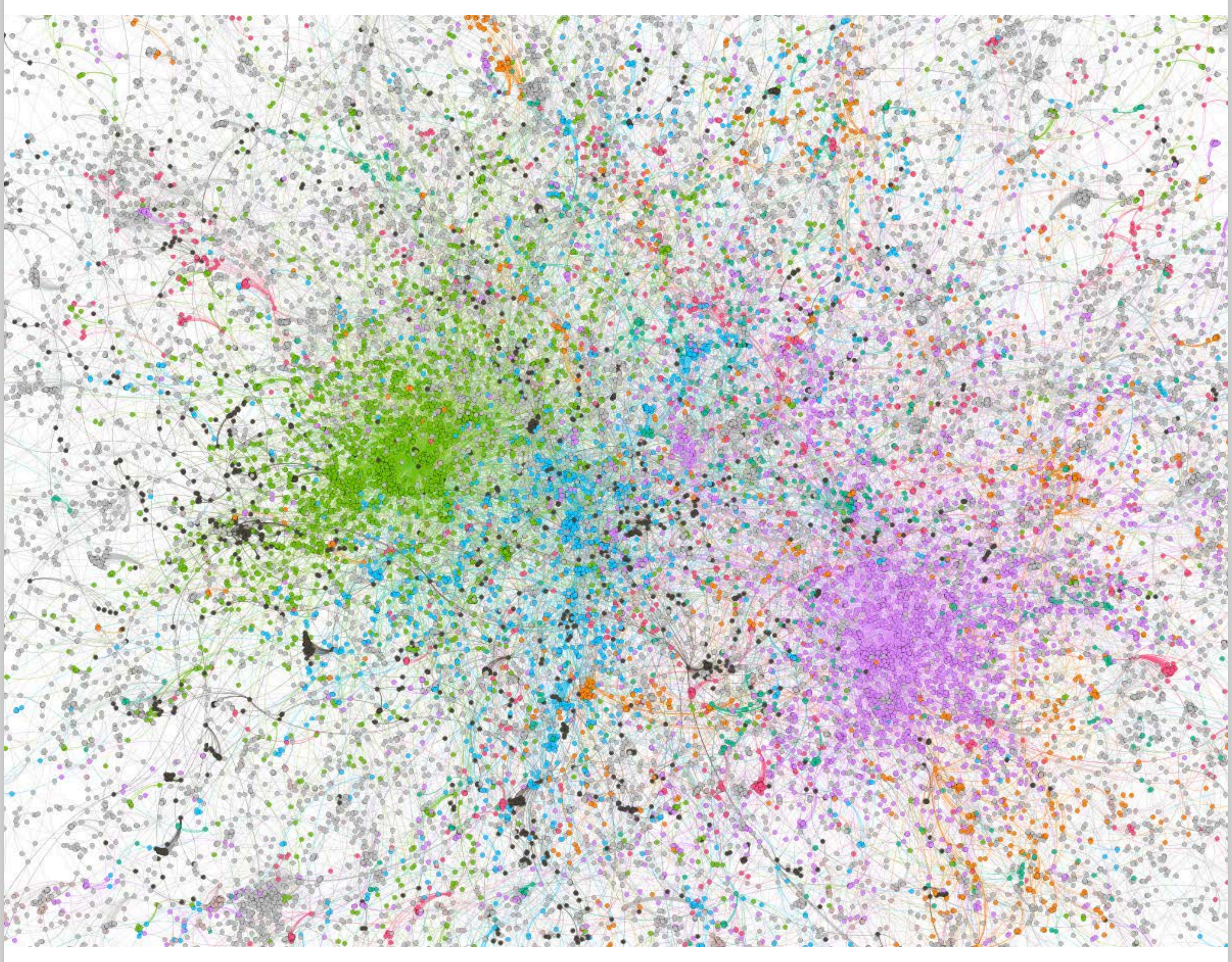
TACC's main goal — here, as in most things — is to facilitate the research of others and power discoveries. "We're mostly interested in letting people access curated datasets and helping them do research," Xu said. "We're collecting, cleaning up, and processing data so it's ready for others to use."

Researchers from The University of Texas at Austin (UT Austin) are among the first to express interest in using the TACC COVID-19 Twitter datasets for targeted research.

"The TACC COVID-19 Twitter collection will be invaluable in enabling us to model communication patterns and topics that emerge across stages of the disease," said Sharon Stover, a professor in the Moody College of Communication. "We may be able to compare the timeline to similar data from other countries such as China that experienced the epidemic earlier. This may lead us toward understanding when typical responses occur and help us to characterize how populations make sense of health pandemics at certain stages in an epidemic's process."

Strover is particularly interested in learning how one might segment tweets by certain population features to learn more about sub-networks that pass along certain information — or ignore it.

Dhiraj Murthy, an associate professor of Journalism and Sociology at UT Austin and author of the first scholarly book about Twitter, plans to use the dataset for his academic work.



Network-analysis figure derived from a sample of 100,000 tweets with 'covid' in the tweet; nodes colored in green are alt-right/strongly conservative Twitter users/organizations. [Credit: Dhiraj Murthy, UT Austin]

"My lab is in the very initial stages of using these data to study two research questions: To what extent is fake news, misinformation, and disinformation regarding COVID-19 present on social media platforms? And: Are social media platforms being used as venues for racist messaging against people of Chinese/Asian origin within COVID-19-related posts?"

Matt Lease, from the UT School of Information, has been using the database to research misinformation in collaboration with Murthy, and also to identify incidents of racist messaging. "The large dataset TACC is collecting, along with its computing and storage services, plus excellent researchers and staff, makes it a fantastic resource for researchers interested in studying and combatting the spread of racist messaging on Twitter."

Both in the moment, and for retrospective analyses, Twitter data can be an incredible resource.

Said TACC research associate Ruizhu Huang: "The large volume of tweets collected at TACC provides a valuable date source to explore various perspectives on COVID-19. And the storage and supercomputing power at TACC will tremendously speed up the data analysis process."

TACC welcomes interested scholars to reach out to collaborate or investigate the datasets. Please contact Weijia Xu [xwj@tacc.utexas.edu] for more information.