Case Study

NBC News

NBC News Analyzes Hundreds of Thousands of Russian Troll Tweets Using Neo4j

INDUSTRY
Media & Publishing

USE CASE
Social Media and Social Network Graphs

GOAL
- Expose troll networks behind Russian meddling in 2016 U.S. election

CHALLENGE
- Restore and analyze more than 200,000 tweets

SOLUTION
- Troll networks exposed using Neo4j graph database

RESULTS
- Surfaced Russian trolls posing as U.S. citizens, local media and local political groups
- Uncovered retweet patterns, hashtags and activity spikes during Moscow business day

During the 2016 U.S. election, Russian trolls infiltrated online conversations. NBC News sought to investigate and encountered two challenges: recovering deleted tweets and analyzing the data to detect patterns. Reporters used Neo4j to scrutinize hundreds of thousands of tweets and expose tactics of Russian troll networks.

The Company

NBC News is one of the leading journalistic outlets in the world. It broadcasts iconic shows, such as "NBC Nightly News," "Today" and "Meet the Press," and runs the 24-hour news station MSNBC plus an array of digital news platforms. Russia's attempt at meddling in U.S. politics was confirmed in 2016 and an NBC team sought to understand how Kremlin-backed trolls exploited Twitter.

The Challenge

There's no question that Russian Twitter trolls interfered in the 2016 U.S. Presidential election. But determining precisely how they did so has been difficult due to the shadowy nature of cyber warfare, the anonymity of the internet, the ease of hiding behind counterfeit identities and the vast volume of social media data.

In November 2017, the U.S. House of Representatives Permanent Select Committee on Intelligence released a list of 2,752 Twitter accounts associated with the Internet Research Agency, the Kremlin-linked "troll farm." (Twitter later expanded the list to 3,814 accounts.) Russian agents impersonated U.S. citizens, news organizations and political groups, and set up fake accounts to spread disinformation and incite division.

By the time the list was released, Twitter had suspended the accounts and deleted tweets and user profiles. NBC reporters needed to find the missing troll tweets.

How could the data be recovered and analyzed? How did the networks operate? How did trolls infiltrate the online conversations of everyday Americans and attempt to sway public opinion? The questions were of paramount public interest – and the answers elusive without tools to rescue and analyze the data.

The Strategy

NBC News turned to Neo4j, a graph database platform ideally suited to illuminating connections in large datasets. The first task was to recover as much of the missing data as possible.

Investigators recovered tweets from the Internet Archive (popularly known as the Wayback Machine) and independent groups that monitored Twitter during the election. These sources yielded a database of 202,973 tweets from 454 accounts.

Case Study

"Using a connections-first approach to analyzing these sorts of datasets, both governments and social media platforms can more proactively detect and deter this sort of meddling behavior before it has a chance to derail democracy or poison civil conversation."

– Will Lyon,
Developer Relations Engineer, Neo4j

This database represented only a fraction of overall activity, but sufficiently enabled the reporters and analysts to begin answering the big question: What were the trolls doing?

The Solution

The graph showed the relationships between entities such as tweets, users (some exposed as known trolls), hashtags, source applications and links.

Graph algorithms measured centrality of nodes based on connections with other entities. Community detection algorithms revealed networks of users who frequently interacted – and identified which trolls were influencers and which simply amplified other trolls. PageRank identified the most influential accounts within each cluster.

The reporters began to see the troll networks in action. Each community featured a small core of content generators and a larger body of retweeters. Only about 25 percent of the troll tweets were original; the rest were retweets. Trolls took advantage of common hashtags and replied to popular accounts in order to amass followers and build influence.

The trolls left lots of footprints. Legitimate Twitter users often tweet from their phones, but the investigators discovered a disproportionately high number of tweets from the Twitter web client. When plotted by time, troll tweets spiked during working hours in Russia.

The Results

The investigators found several types of Russian troll accounts. Some were concocted to resemble typical Americans such as @LeroyLovesUSA. Others mimicked news sites such as @OnlineCleveland. A third category pretended to be political organizations such as @TEN_GOP, which represented itself at the Tennessee Republican Party.

In reality, all were operated by the Internet Research Agency in Russia. Neo4j illuminated how hundreds of fake accounts coordinated in networks.

Within weeks of the release of the troll list, NBC and Neo4j generated a database of more than 200,000 tweets. NBC published an exposé based on the Neo4j analysis. The story revealed how Russian trolls impersonated Americans, interacted with legitimate users, attracted hundreds of millions of followers and injected propaganda into American politics.

Reflecting on what social media platforms and governments can do to prevent future abuse, Will Lyon, Developer Relations Engineer at Neo4j, said, "Using a connections-first approach to analyzing these sorts of datasets, both governments and social media platforms can more proactively detect and deter this sort of meddling behavior before it has a chance to derail democracy or poison civil conversation."

Ben Popken, NBC Senior Business Reporter, tweeted his appreciation to the Neo4j team, saying, "Huge props and thank you to Neo4j for helping compile and analyze the deleted Twitter data, surfacing trends and uncovering new angles."

Questions about Neo4j?

Contact us across the globe:
info@neo4j.com
neo4j.com/contact-us