WIKIMEDIA
FOUNDATION

BIDEN CAMPAIGN STAFF

IMPERSONATION RETROSPECTIVE

★ ★ ★ ★ Table of Contents ★ ★ ★ ★

BIDEN CAMPAIGN STAFF IMPERSONATION RETROSPECTIVE

Reviewing an impersonation attempt targeting the Biden campaign on 8/9/20 on English Wikipedia, this page records the work done by WMF to improve disinformation capabilities for the Foundation and volunteer functionary teams affected by the 2020 US presidential election.

Disinformation attacks targeting Wikipedia content tied to ongoing political campaigns, more than regular vandalism of such content, are a significant threat to the well-being of the communities and the credibility of the content they create and curate to freely share, reducing readers' trust in Wikipedia. US elections are not the first encounter Wikimedia has had with politically motivated disinformation, but these elections traditionally shape the form of disinformation threats for an extended period of time across many languages as bad actors learn from what worked in the US context. Thus, the Foundation sees the 2020 presidential election as an important shared learning opportunity for itself, English Wikipedia functionaries, the stewards, and other functionaries across the movement to better protect our communities and the platform.

CASE PROFILE

In August 2020, an account impersonating the distinct real name of a leading member of Joe Biden's US presidential campaign registered and made changes to the personal article of US Senator Duckworth that could be read as preparations for a forthcoming Vice President nomination. Shortly thereafter, a new Twitter account (@VPSearchUpdate; seems set up specifically for this) started tweeting about this, making the connection explicit and trying to generate traction on Twitter.

Overall, the incident appears to have been thoughtfully designed with a fair amount of knowledge of the English Wikipedia's defense mechanisms. The attacker took care to:

1. Use a Wikimedia account name that only insiders of the US presidential election process would instantly identify as concerning by itself but prominent enough that everyone could connect it to the Biden campaign if they cared to search

2. Use that new Wikimedia account in line with the local community's editorial guidelines in contributing to BLP articles, so flying under the radar of both:

   a) the recent change patrol volunteers monitoring through Huggle and other tools, or
   b) content editors who have the Duckworth entry on their watch lists.

3. Produce an artifact that would reasonably isolate their aims from countermeasures if spotted - a screenshot of the edit history displaying the all-important piece of evidence in form of the user account name that could be easily reproduced off-platform, on Twitter.

4. Deploy the screenshot very quickly on Twitter, putting forward an explicit, misleading claim based on the inference that a Biden campaign staffer would only edit the article if Duckworth were of extreme (VP) significance to the campaign. (No claims related to Duckwork's potential role in the campaign were made on Wikimedia's platform itself.)

In terms of **impact**, Wikimedia got lucky that the Twitter account was not instantly magnified by a pre-prepared botnet on Twitter to spread the word. Not following through with the bot stage suggests that this attack might have been a test. The approach deployed did validate their method as it enabled them to combine the credibility of Wikipedia that the communities have worked hard to create for two decades with explicit disinformation claims tied to the Wikimedia user account name on Twitter. Sceptical Twitter users, including journalists and other societal influencers, could check the Duckworth article's history on Wikipedia to see that the screenshot wasn't manipulated; adding an additional layer of credibility. While Wikimedia's community and staff systems didn't identify the issue early, the Biden campaign was quick to spot the problem. They reached out to OTRS and the Foundation's Security team, which passed the issue on for timely resolution to T&S while John Bennett had a call with a security official of the campaign.

STRENGTHS

Two separate processes worked fairly well:

1. OTRS initially reacted timely to address the Biden campaign's concerns in trying to address the impersonation by blocking the account.

2. Security and T&S, once it hit their radars respectively, moved quickly in trying to identify and coordinate resolutions of an uncommon challenge.

CHALLENGES

The attack and the efforts to repel it surfaced a range of challenges:

1.

There is no shared understanding of disinformation as a type of attack the projects face and how to deal with it; the stewards, Foundation staff, English Wikipedia teams from RC to oversight, and the stewards have no shared understanding of the threat, how to identify it, and how best to protect the communities and the platform from damage it can do. Examples of these gaps surfaced in this instance include:

a) OTRS dealt with the initial user name issue like a common identity block focused on preventing future contributions, not by addressing the actual disinformation potential inherent in the user name.
b) The remedy of putting an info box onto the blocked user account was ineffective for addressing disinformation concerns. Only the most engaged observers from twitter might go several steps further than the article history; clicking their way all the way through to the user account itself, find a box the community did place to flag concerns about the account, and factor it into their evaluation.
c) T&S needed more than 15 minutes to organize itself and marshall an effective allocation of staff with shared direction to engage the stewards in an effective redress.
d) T&S needed more than an hour to first find an active steward to handle the renaming of the concerning account.
e) Once it became apparent that the database would not deliver the aimed-for indirect fix of the article history log that was at the center of the Twitter part of the attack, T&S reached out to English Wikipedia's oversight team, asking for the logs to be redacted. This request was initially rejected and ultimately never enacted because the database did catch up and displayed the renamed user name before the WMF-OS conversation could be resolved.
f) The following day, OTRS unblocked the renamed disinformation user account.

2.

As the list of issues under the prior point indicates, the hybrid system also has no shared, effective, communication channels capable of addressing fast-moving disinformation attacks that aim to do damage off-platform and thereby eliminating the traditional coordinating factor of damage transparently done onwiki. Examples include:

a) The Foundation was not initially aware of the OTRS ticket's content.
b) There was a considerable time gap, several hours, between Security identifying the email outreach it received from the Biden campaign that it passed on to T&S and T&S picking it up in its inbox.
c) T&S needed some time to find an active steward on IRC.

d) English Wikipedia's Oversight team replied to T&S's email within the hour of its outreach but that fact didn't filter through to the Foundation because the oversight team's reply did get stuck in spam filters.

e) OTRS had no insights into any steps beyond its own initial scope, likely leading to the unblock of the disinformation account the following day.

3. A magnifying factor enabling disinformation attacks has been that there are no anti-disinformation tools supporting volunteer functionaries or staff in identifying and effectively combating attacks. This gap is a major weakness beyond communications channels the Foundation needs to address through resource allocation, including:

a) Streamlined functionary and moderation tools that make recent change patrolling, admin, steward, and office actions easier to implement and automatically communicate to other key stakeholders.

b) Rebuilding the MediaWiki log system into a live threat intelligence resource available to key volunteer functionaries and staff. Most vital information concerning activities happening on the platform gets logged in public user account creation, user block, article deletion or protection logs. In general, local admins and recent changes volunteers are doing a good job identifying and initially dealing with such issues. The structural challenge is that these logs are distributed and not easily machine-readable. Traditionally, the stewards work with bots that report live logs on IRC. Those are monitored by the Small wiki monitoring team that includes Stewards, global sysop and others. Disinformation is a challenge across all our relevant logs, on all wikis, with a tight time window for cleanup to prevent damage on- and off-wiki, and it comes with a much broader range of potential flags than a list of concerning terms.

c) Machine learning (ML) models and ML based products to automatically search for and flag potential disinformation attacks on the platform, notably including those targeting Wikidata with its weaker community defenses, for human attention.

d) Threat intelligence resources that help community and staff to quickly identify and understand off-platform components of an ongoing attack in which Wikipedia - like in the case at hand - is misused to lend credibility to a false claim on another platform.

e) A shared information vault where staff and functionaries of different groups can safely share attack patterns and long-term abuser profiles.

f) A shared online learning and training infrastructure where volunteers, starting with recent change patrollers and adminship-aspirants, can learn about topics relevant to their roles, including but not limited to disinformation.

IMPROVEMENT OPTIONS

It took the system more than 24 hours to resolve what was in effect a lightly resourced but effective disinformation attack on its most well-defended wiki. If the same hostile method would be deployed during the peak of the campaign on a high profile issue, considerable damage could have been done to the project's public reputation the Foundation's ability to continuously defend limited community self-governance as a viable T&S model for a major platform in the face of increasing regulatory pressures, and the integrity of the democratic process in the US.

Given that the Foundation, with the help of the community, is trying to address disinformation challenges on Wikimedia platforms, the following initial steps are recommended to mitigate risks:

1. Share this retrospective with key community stakeholder groups who can receive it under NDA: the stewards; the English Wikipedia Arbitration Committee, Oversight, and Checkuser teams; and the OTRS administration. Also, it should be made available to the community at large once the broader work on disinformation begins.

2. Build a common understanding of disinformation attacks through a series of shared conversations about this case and the subject more broadly with interested members of the key community stakeholders groups identified.

3. Provide an effective, shared coordination space for affected functionary groups and staff that is dedicated to identifying, triaging, and coordinating the countering of disinformation attacks tied to the US presidential election, likely a Slack channel as it is compatible with NDA requirements and already integrated with emergency@ services.

4. Include disinformation challenges within the scope of the UCoC enforcement outline conversation mandated by the Board after its ratification of a UCoC text. Transparent and clear process outlines are filat to enable effective, localized, and ideally decentrally organized redress.

Looking back at hybrid disinformation attacks that took place earlier in 2020 involving English and Indonesian language Wikipedia volunteers whose physical security was endangered in the course of their voluntary article contributions, disinformation resilience is a long-term challenge to the movement that is encompassing both onwiki and offwiki components. This combination of asymmetric threats requires not just hand-in-glove collaboration between the Foundation and the communities it supports but ongoing Foundation investment.

www.wikimediafoundation.org