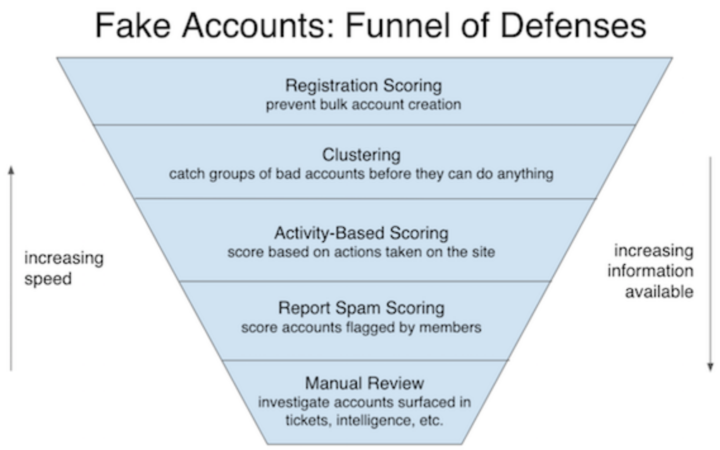
Automated Fake Account Detection at LinkedIn

Jenelle Bray September 12, 2018 Share Tweet Like 25

To maintain a safe and trusted professional community on LinkedIn, we require that every LinkedIn profile must uniquely represent a real person. One of the ways we ensure that accounts are real is by building automated detection systems at scale for detecting and taking action against fake accounts. The Anti-Abuse team at LinkedIn creates the systems that allow us to protect our members from activity by bad actors.

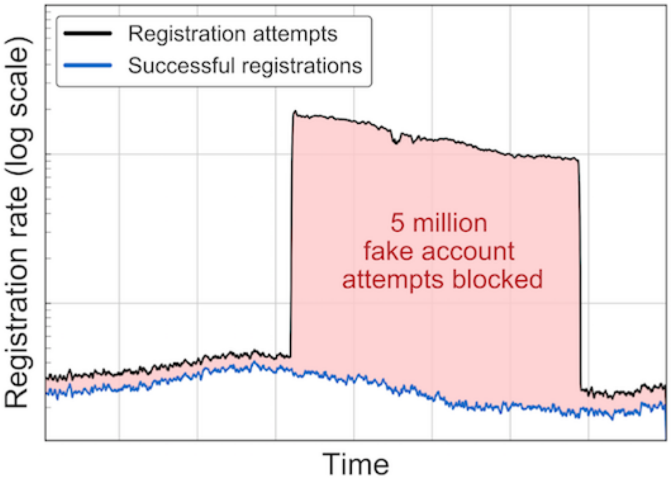
Unfortunately, LinkedIn is the target of bad actors who constantly try to create fake accounts. There is a wide range of sophistication behind these bad actors and the intent varies. Fake profiles can be used to carry out many different types of abuse: scraping, spamming, fraud, and phishing, among others. By preventing or promptly removing fake accounts on the site, we ensure that LinkedIn members are protected.

In order to build robust countermeasures against different types of attacks on our platform, we employ a funnel of defenses to detect and take down fake accounts at multiple stages. We aim to catch the majority of fake accounts as quickly as possible to prevent harm to our members.


Fake Accounts: Funnel of Defenses
Registration Scoring
prevent bulk account creation
Clustering
catch groups of bad accounts before they can do anything
Activity-Based Scoring
score based on actions taken on the site
Report Spam Scoring
score accounts flagged by members
Manual Review
investigate accounts surfaced in tickets, intelligence, etc.
increasing speed
increasing information available

At the top of the funnel is the first line of defense: registration scoring. For many types of abuse, attackers require a large number of fake accounts for the attack to be financially feasible. Thus, in order to proactively stop fake accounts at scale, we have machine-learned models to detect groups of accounts that look or act similarly, which implies they were created or controlled by the same bad actor.

Every new user registration attempt is evaluated by a machine-learned model that gives an abuse risk score. Signup attempts with a low abuse risk score are allowed to register right away, while attempts with a high abuse risk score are prevented from creating an account. Attempts with medium risk scores are challenged by our security at preventing bulk fake account creation. The figure below shows one attack where the model blocked five million fake accounts from being created in less than a day.



Although we prevent a large majority of bulk fake accounts from being created at registration, we sometimes don't have enough information at that point to determine if accounts are fake. For this reason, we have other downstream models to catch smaller batches of fakes. First, we create clusters of accounts by grouping them based on shared attributes. We then find account clusters that show a statistically abnormal distribution of data, which is indicative of being created or controlled by a single bad actor. These are supervised machine learning models that use features per cluster instead of per member. The models score the clusters, then propagate the cluster label to individual accounts. This cluster approach allows us to catch more fake accounts quickly, faster than we could if we wait for them to start taking abusive actions on the site.

Fake accounts that are not created in bulk by a single bad actor are often detected by our activity-based models. Our models require more information on these accounts' behavior to decide whether they are indeed fake. We have many models that either look for specific types of bad behavior typical to abusive accounts or behavior that is anomalous. Additionally, our systems have redundancy, which ensures that fake accounts not caught by the early stages of our defenses (top of the funnel) are eventually caught by the latter ones (bottom of the funnel).

A human element will always be necessary to catch fake accounts that have evaded our models. While we strive to take down fake accounts before they interact with our members, we also get signals from our members who report suspicious activity on the site. Members give us valuable information by reporting accounts that have not been caught by our models so that they can go through additional scoring and review. Finally, we have a team of investigators that look for accounts that have evaded all levels of automated defenses. These investigations also yield valuable signals that can subsequently be incorporated into our models.

We are constantly improving our fake account models given the adversarial nature of abuse. The quicker we catch fake accounts, the more we prevent damage to our members, which helps keep LinkedIn a safe, trusted and professional community.

machine learning, Security

Related story
Serving Top Comments in Professional Social Networks

Related story
The Statistical Modeling System Powering LinkedIn Salary

Blog Data Open Source Jobs Women in Tech

LinkedIn Corporation © 2019 About Cookie Policy Privacy Policy User Agreement

LinkedIn.com