Trafficking *Fraudulent* Accounts: The Role of the Underground Market in Twitter Spam and Abuse

Kurt Thomas, *University of California, Berkeley, and Twitter*
Damon McCoy, *George Mason University*
Chris Grier, *University of California, Berkeley, and International Computer Science Institute*
Alek Kolcz, *Twitter*
Vern Paxson, *University of California, Berkeley, and International Computer Science Institute*
Why create fraudulent social media accounts?

To sell stuff, marketing, spam
Visibility
Spread malware
Amplification and suppression of views
  - propaganda and political views
  - smear campaigns
  - astroturfing
Phishing
Inflating account follower statistics, ego
What are the similarities and differences between conventional e-mail spam and social media spam?

Trust
- Credibility differences dependent on platform, social media sometimes more

Degree of visibility (targeted vs open)
- Social media spam longer lasting

More evolving attack patterns in social media

Social media more about “fake engagement”

Incentive differences

Risk inhibiting free speech values to squash spam or fake accounts

Both email and social media have methods of the spammer confirming viewing

Platform type informs type of spam
So many bots

Mentioned in the paper:
3% of Twitter accounts fraudulent (2011)
1.5% of Facebook accounts fraudulent (2012)

More recently:
9-15% of active Twitter accounts are bots - Varol 2017 [2]

(Remember, a bot may not be a fraudulent account!)
Fraudulent Accounts in the News

The Follower Factory – The New York Times

How China Built a Twitter Propaganda Machine Then Let It Loose on Coronavirus – ProPublica

Are ‘bots’ manipulating the 2020 conversation? Here’s what’s changed since 2016. – The Washington Post

Bot-like Turkish accounts complement military operation in Syria – DFRLab

Suspicious third-party apps monetize fake engagement on TikTok – DRFLab

Nearly half of Twitter accounts pushing to reopen America may be bots – MIT Tech Review
Let's do an unscientific experiment. Name 3 accounts and let's see how many (potential) fraudulent accounts follow them.
Overview of the Study

Identified and purchased ~120K accounts ($5,000) from 27 different fraudulent account merchants.

Worked to understand the organization of the account marketplace and how merchants circumvented registration defenses.

Developed a classifier to fingerprint fraudulent accounts and train it starting with the 120K purchased accounts.

Deployed classifier and detected fraudulent accounts over the 10-month period.

Analyzed results of disrupting the account marketplace.

Detection of auto-generated accounts using classified improved on Twitter’s existing detection methods.
Initial Findings

Median cost of a fraudulent account: $0.04

90% of orders fulfilled within 3 days with a median delivery time of 1 day

Fraud: 13% of accounts purchased were subsequently resold

PVA (Google, FB) cost significantly more than unverified accounts (Hotmail, Yahoo, Twitter) while e-mail verified accounts measurably more expensive – but less than PVA.

Prices consistent overtime with little-to-no variation (Possible reasons why?)

- Uninterrupted supply, market competition, PPI factors less of an issue, defenses remained unchanged
Findings (cont.)

79% of accounts purchased came from unique IP addresses (evenly distributed over /16 and /24 subnet and spanned 164 countries).

Merchants use 50% of IP addresses under their control to register fewer than 10 accounts.

All but 5 merchants (77% of accounts) verified through email confirmation.

Merchants solved CAPCHAs on 35% of accounts with CAPTCHA verification slightly increasing an account’s respective price.

Only 8% of dormant accounts purchased later detected and suspected indicating limited usefulness of at-abuse time metrics against automated registration.

Median age of stockpiled accounts 31 days old

“We find that sellers like accs.biz and victorieservices have tens of thousands of IPs at their disposal on any given day, while even the smallest web storefront merchants have thousands of IPs on hand to avoid network-based blacklisting and throttling.”

<table>
<thead>
<tr>
<th>Email Provider</th>
<th>Accounts</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotmail.com</td>
<td>64,050</td>
<td>68.32%</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>12,339</td>
<td>13.16%</td>
</tr>
<tr>
<td>mail.ru</td>
<td>12,189</td>
<td>13.00%</td>
</tr>
<tr>
<td>gmail.com</td>
<td>2,013</td>
<td>2.15%</td>
</tr>
<tr>
<td>nokia.mail.com</td>
<td>996</td>
<td>1.06%</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>2,157</td>
<td>2.30%</td>
</tr>
</tbody>
</table>

Table 5: Top 5 email providers used to confirm fraudulent Twitter accounts.
Figure 3: Availability of unique IPs over time for the six merchants we track over the longest period. All but one seller we repeatedly purchase from are able to acquire new IP addresses to register accounts from over time.

Figure 4: CAPTCHA solution rates per each IP address abused by a variety of merchants as well as the rates for all merchants combined.
What are some different strategies social media companies take to reduce abuse?

- **Blacklisting**
- **E-mail verification**
- **PVA**
- **CAPTCHAs**
- **Re-verification**
- **KYC (Know Your Customer)**

The diagram illustrates the effectiveness and user inconvenience of these strategies, with a gradient from **not effective** to **very effective** and from **soft** to **hard** user inconvenience.
Recommendations of the Authors

**E-mail confirmation**: raises cost of accounts by 56%, soft measure for users, and reselling e-mail addresses becomes harder.

**CAPTCHAs**: prevent merchants from fraudulently registering 92% of accounts and provides a blacklisting signal

**IP Blacklisting**: Useful, but public and commercial blacklists have false positive rates that are too high.

**Phone verification**: Phone verified Facebook and Gmail accounts cost 150 times more than their non-PVA counterparts.
Platform incentives?

Would
- Quality of service
- Public opinion and perception
- Legal implications, government pressures
- Computational resource savings

Wouldn’t
- Inflating statistics, publicly-traded companies’ relationship to growth metrics
- Tinder, Ashley-Madison shaping a platform’s perception or image
- Computational resource savings (machine learning $$$)
- Cost > Benefit
What were some of the clues the researchers used in their model to identify fraudulent accounts at-registration?
Impact of the Classifier

During the 10-month period, 99.99% of flagged accounts were spam (very precise!) and 95.08% of the spam accounts were recalled.

Those 27 merchants turned out to be responsible for 10-20% of all accounts flagged as spam.

Of the accounts from 27 merchants, 73% were actively tweeting while 37% remained dormant.

Blackhat forums and web storefronts more damaging and sophisticated than Fiverr and Freelancer sellers.
Total annual revenue from the sale of fraudulent accounts on Twitter estimated to be between $127,000–$459,000, apparently much less than revenue from spam.

Despite depleting total stock (well 95%) of the 27 merchant’s fake Twitter accounts, **within 2 weeks they were back** with fresh, working, fraudulent accounts.

“While Twitter’s initial intervention was a success, the market has begun to recover. Of 6,879 accounts we purchased two weeks after Twitter’s intervention, only 54% were suspended on arrival. As such, long term disruption of the account marketplace requires both increasing the cost of account registration and integrating at-signup time abuse classification into the account registration process.”
Conclusion: We need better at-registration defenses.
References


Figure 2: CDF of registrations per IP tied to purchased accounts, legitimate accounts, and suspended (spam) accounts.

Table 5: Top 5 email providers used to confirm fraudulent Twitter accounts.

<table>
<thead>
<tr>
<th>Email Provider</th>
<th>Accounts</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>hotmail.com</td>
<td>64,050</td>
<td>68.32%</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>12,339</td>
<td>13.16%</td>
</tr>
<tr>
<td>mail.ru</td>
<td>12,189</td>
<td>13.00%</td>
</tr>
<tr>
<td>gmail.com</td>
<td>2,013</td>
<td>2.15%</td>
</tr>
<tr>
<td>nokia-mail.com</td>
<td>996</td>
<td>1.06%</td>
</tr>
<tr>
<td>Other</td>
<td>2,157</td>
<td>2.30%</td>
</tr>
</tbody>
</table>

Figure 3: Availability of unique IPs over time for the six merchants we track over the longest period. All but one seller we repeatedly purchase from are able to acquire new IP addresses to register accounts from over time.

Figure 4: CAPTCHA solution rates per each IP address abused by a variety of merchants as well as the rates for all merchants combined.

Figure 5: Recall of generated merchant patterns for all purchased accounts as a function of training the classifier on data only prior to time t.

Figure 6: Fraction of all suspended accounts over time that originate from the underground market.