Understanding Promotion-as-a-Service on GitHub

Kun Du, Hao Yang, Yubao Zhang, Haixin Duan, Haining Wang, Shuang Hao, Zhou Li, Min Yang
GitHub as social networks

• GitHub
  ➢ The most important code management and sharing website

  ➢ More than 40M developers, more than 44M repositories created, and 2.9M organizations by 2019
GitHub as social networks

- GitHub
  - The most important code management and sharing website
  - More than 40M developers, more than 44M repositories created, and 2.9M organizations by 2019

One can watch a repository to receive notifications for new pull requests and issues that are created.

A fork is a copy of repository. Forking a repository allows one to freely change without affecting the original project.

One can star repositories and topics for tracking projects of interest.
Job Recruiting

**Forking:** copying projects from other users with the aim of using the code yourself. Forks are personal copies of another user’s repository which live in your account. It’s possible to make changes to forks without affecting the original project. **Attention:** Stars and forks are a sign of good, usable code.

- **Stars** (quick reminder: good code is forked and starred a lot, so pay attention to these elements).
- Contribution calendar (you might be tempted to think it’s useless for you, but read on to find out why you should like it A LOT).

https://devskiller.com/source-developers-from-github/
Job Recruiting

• Impact on Job Recruiting

Forking: copying projects from other users with the aim of using the code yourself. Forks are personal copies of another user's repository which live in your account. It's possible to make changes to forks without affecting the original project. **Attention:** Stars and forks are a sign of good, usable code.

- Stars (quick reminder: good code is forked and starred a lot, so pay attention to these elements).
- Contribution calendar (you might be tempted to think it's useless for you, but read on to find out why you should like it A LOT).

https://devskiller.com/source-developers-from-github/

https://www.zhaopin.com/
GitHub Promotion-as-a-Service

• Some developers attempt to manipulate the social statistics of their own accounts by purchasing stars and forks
GitHub Promotion-as-a-Service

- In 2019, SK Telecom, the biggest mobile service provider in South Korea, was reported abusing GitHub stars by giving free drinks to accounts for starring a specified repository.
Our Contributions

• We performed the first comprehensive study on GitHub promotion and uncovered the strategies used by suspicious promoters.

• We conducted a large-scale measurement on more than 40 million GitHub accounts, and identified more than 63K suspected promotion accounts.

• We shed new light on how this promotion service is operated. We disclosed a hidden functionality of GitHub that allows a user to pretend to be a skillful developer.
There are 4 main components in our system:
1. Data collection.
2. Preprocessing.
3. Feature Engineering.
4. SVM-based Classifier.
Data Collection

1. GH Archive (https://www.gharchive.org/)

GH Archive records public GitHub event and archives it in JSON format. We collected all archive files from 2015 to 2019.
Data Collection

1. GH Archive (https://www.gharchive.org/)
GH Archive records public GitHub event and archives it in JSON format. We collected all archive files from 2015 to 2019.

2. GitHub API. (https://api.github.com/)
GH Archive only archives the history since 2015 and also misses the registration and profile information. So, GitHub API is utilized to collect these missing data, such as the profile of a specific GitHub account.
Data Collection

1. GH Archive (https://www.gharchive.org/)
GH Archive records a public GitHub event and archives it in JSON format. We collected all archived files from 2015 to 2019.

2. GitHub API. (https://api.github.com/)
GH Archive only archives the history since 2015 and also misses the registration and profile information. So, GitHub API is utilized to collect these missing data, such as the profile of a specific GitHub account.

3. GitHub Web Interface (https://www.github.com/)
Through the web interface, we can crawl the information directly from GitHub web pages, such as a specific GitHub account’s avatar, the popular repository list, and the hot trend list.

We extracted a total of **23,375,824** accounts from 2015 to 2019.
Training Data

Promoter Accounts
1,023 bought from 10 different promoters
4 from the QQ group
3 from WeChat
3 from telegram
Training Data

Promoter Accounts
1,023 bought from 10 different promoters
4 from the QQ group
3 from WeChat
3 from telegram

Normal Accounts
1,550 users who made major active contributions in well-known repositories.
  a. the contributors of popular GitHub repositories
  b. those who have proposed valuable issues on popular repositories.

Scikit-learn: https://github.com/scikit-learn/scikit-learn
Jquery: https://github.com/jquery/jquery
Redis: https://github.com/antirez/redis
Bootstrap: https://github.com/twbs/bootstrap
Tensorflow: https://github.com/tensorflow/tensorflow
Bitcoin: https://github.com/bitcoin/bitcoin
Vscode: https://github.com/Microsoft/vscode
Cocos2d: https://github.com/cocos2d/cocos2d-objc
React: https://github.com/facebook/react
Linux: https://github.com/torvalds/linux
Qemu: https://github.com/qemu/qemu
Elasticsearch: https://github.com/elastic/elasticsearch
Homebrew: https://github.com/Homebrew/homebrew-core
Rails: https://github.com/rails/rails
Go: https://github.com/golang/go
Ant-design: https://github.com/ant-design/ant-design
Opencv: https://github.com/opencv/opencv
Swift: https://github.com/apple/swift
Elixir: https://github.com/elixir-lang/elixir
# Feature Engineering

<table>
<thead>
<tr>
<th>No.</th>
<th>Type</th>
<th>Created_at</th>
<th>Repo</th>
<th>Star Count</th>
<th>Repo Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WatchEvent</td>
<td>2018-12-23T07:37:38Z</td>
<td>mahmoudawesome-python-applications</td>
<td>6,369</td>
<td>Popular Author</td>
</tr>
<tr>
<td>2</td>
<td>WatchEvent</td>
<td>2018-12-23T07:37:44Z</td>
<td>FavioVazquez/ds-cheatsheets</td>
<td>3,597</td>
<td>Popular Author</td>
</tr>
<tr>
<td>3</td>
<td>WatchEvent</td>
<td>2018-12-23T07:38:02Z</td>
<td>facebookresearch/flashlight</td>
<td>762</td>
<td>Organization</td>
</tr>
<tr>
<td>4</td>
<td>WatchEvent</td>
<td>2018-12-23T07:39:19Z</td>
<td>trekhlebhomemade-machine-learning</td>
<td>8,749</td>
<td>Popular Author</td>
</tr>
<tr>
<td>5</td>
<td>WatchEvent</td>
<td>2018-12-23T07:39:30Z</td>
<td>ghost1****/t****</td>
<td>716</td>
<td>Promotion Target</td>
</tr>
<tr>
<td>6</td>
<td>WatchEvent</td>
<td>2018-12-23T07:39:38Z</td>
<td>FAQGURU/FAQGURU</td>
<td>3,529</td>
<td>Popular Author</td>
</tr>
<tr>
<td>7</td>
<td>WatchEvent</td>
<td>2018-12-23T07:39:46Z</td>
<td>alibaba/x-deeplearning</td>
<td>2,145</td>
<td>Organization</td>
</tr>
<tr>
<td>8</td>
<td>WatchEvent</td>
<td>2018-12-23T07:40:00Z</td>
<td>alash3a/redis</td>
<td>689</td>
<td>Popular Author</td>
</tr>
</tbody>
</table>
Feature Engineering

1. Action types
   Promotion accounts: star and fork operations occupy most action types
   Normal accounts: action types are much more diverse
Feature Engineering

1. Action types
   Promotion accounts: star and fork operations occupy most action types
   Normal accounts: action types are much more diverse

2. Action boosts
   Promotion accounts are usually associated with a large number of star and fork operations in a short time period.
Feature Engineering

1. Action types
   Promotion accounts: star and fork operations occupy most action types
   Normal accounts: action types are much more diverse

2. Action boosts
   Promotion accounts are usually associated with a large number of star and fork operations in a short time period.

3. Action intervals
   A promotion account performed star operations and the interval between two operations is less than 20 seconds.
Detection and Result Validation

• We built SVM-based classifier and detected 63,872 suspected promotion accounts
Detection and Result Validation

• We built SVM-based classifier and detected **63,872** suspected promotion accounts
Detection and Result Validation

• We built SVM-based classifier and detected 63,872 suspected promotion accounts
Detection and Result Validation

- Randomly sampled 1,000 from 63,872 suspected promotion accounts, and found the detection accuracy is 97.3%
Forging Retroactive Commits

### Popular repositories

<table>
<thead>
<tr>
<th>tools</th>
<th>A Python Crawler Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td>705</td>
</tr>
<tr>
<td>GitHub stars</td>
<td>5</td>
</tr>
</tbody>
</table>

### 2 contributions in 2016

<table>
<thead>
<tr>
<th>Month</th>
<th>Less</th>
<th>More</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Learn how we count contributions.

### Contribution activity

- ghost123gg has no activity yet for this period.

Forging Retroactive Commits
Small business promotion

The company first publishes the repository on GitHub and places the link in their source code, homepage hyperlink or “readme.md” to their online shop’s homepage, showing a hyperlink from GitHub.
Recommendations

We suggest that GitHub should take action to regulate the existing promotion.

• One possible action is to **send emails with confirmed information** to suspected accounts and ask them for an active response. Most of these suspected promotion accounts are acquired through selling and buying. Thus, they are likely to ignore emails.
Recommendations

We suggest that GitHub should take action to regulate the existing promotion.

• One possible action is to **send emails with confirmed information** to suspected accounts and ask them for an active response. Most of these suspected promotion accounts are acquired through selling and buying. Thus, they are likely to ignore emails.

• Another action could be requiring suspected promotion accounts to complete a **more complicated captcha** during registration. This can prevent automatic login and automatic actions such as forking and starring.
Summary

• We have conducted the *first* comprehensive investigation on a new promotion service on GitHub called “Promotion-as-a-Service,” so as to improve social status and earn advantages in career development.

• We have developed a behavior pattern model by purchasing services from actual GitHub promotion service providers and have detected 63K suspected promotion accounts from 2015 to 2019.

• We believe that our findings will help the security community to pay more attention to all kinds of fraudulent promotion methods. Moreover, our work will help to retain a fair and objective recruitment in the IT industry.
Thank You!

Q & A