Understanding Promotion-as-a-Service on GitHub

Kun Du, **Hao Yang,** Yubao Zhang, Haixin Duan,
Haining Wang, Shuang Hao, Zhou Li, Min Yang

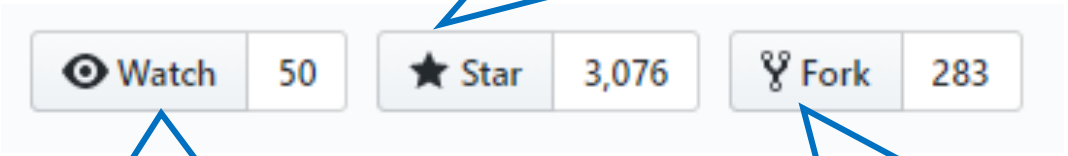
GitHub as social networks

- 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 star repositories and topics for tracking projects of interest.



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.

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

Technical director **$7,500 – 10,000 per month**

北京云族佳科技有限公司 ZP名企 北京-大兴区 | 10年以上 | 硕士 | 招1人

优先条件:(任意满足一项加分)

- 熟悉java开发技术栈
- Pythonista
- 有自己的技术博客并持续更新
- 有Github上开源项目且100star以上。

Back-end Engineer **$2,900 – 6,000 per month**

北京易联纵横科技有限公司 ZP名企 北京 | 经验不限 | 学历不限

Require more than 100 stars

加分项:

1.除了Golang 或者Elixir 之外，对其他语言也有所涉猎，比如Ruby，Rust，Scala，Clojure 等;

2.对美（产品、程序、设计）有追求;

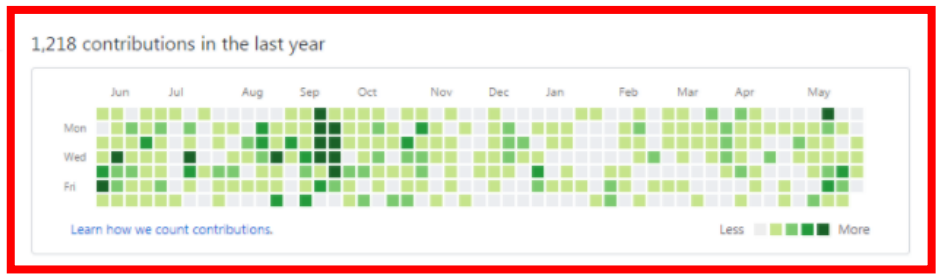
3.Github有后端项目超过100 star，或给大型开源项目贡献过代码;

4.具备TDD/BDD 实战经验;

5.具备设计公开API 接口的实战经验;

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).



Front-end Engineer **$1,500 – 2,300 per month**

多加网络科技(北京)有限公司 ZP名企 北京 | 1-3年 | 大专 | 招1人

Require even more than 500 stars

熟悉或了解 React Native ;

GitHub 有500 以上 star 的前端项目;

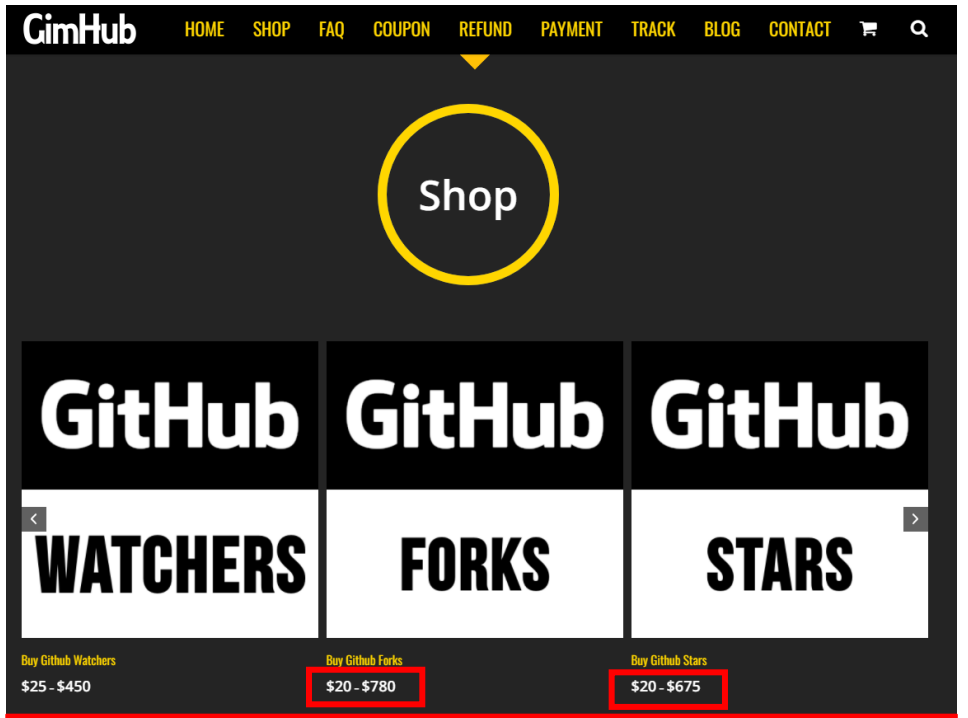
https://www.zhaopin.com/

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

GitHub Promotion-as-a-Service

- Some developers attempt to manipulate the social statistics of their own accounts by purchasing stars and forks

GitHub fake forking and staring group in IM



GimHub, an online shop selling GitHub Watchers, forks and stars

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

Changing star on GitHub with free drinks, you change it?

News 2019-08-02 12:36:09 views: null

According to Theregister reports, an open source project on Github is problematic, it was achieved through free drinks in exchange for increasing the number of star. The incident on a website, Korea's largest wireless communications provider SK Telecom operator exposure, SK Telecom is the project sponsor, this promotion has been discontinued.

The project is Metatron Discovery, which is an application of real-time data analysis, based on a customized version of Apache Druid. Achieved some results under the effect of this activity, the project has nearly 2,500 star.

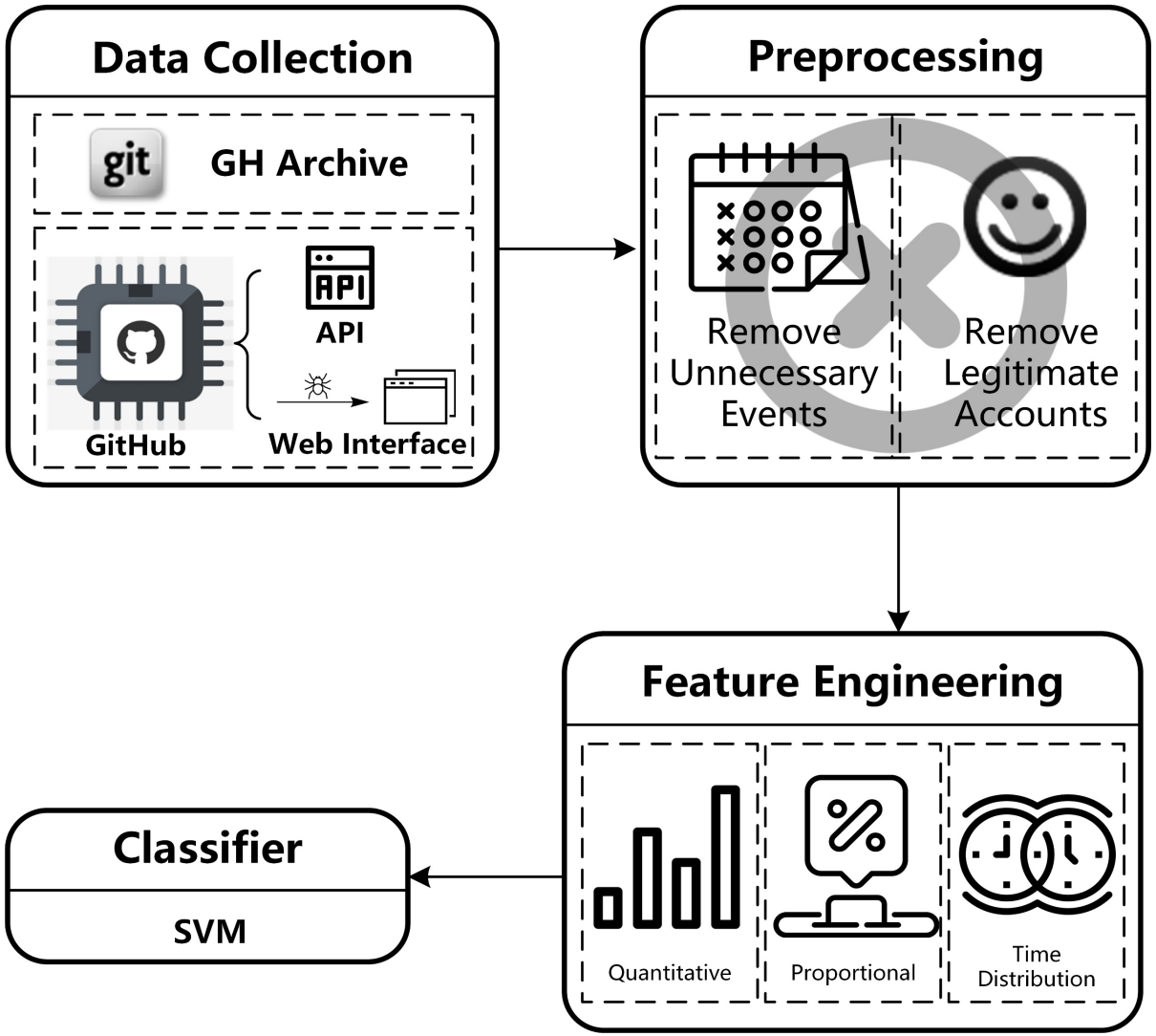
Our Contributions

- 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.

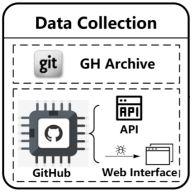
- We shed new light on how this promotion service is operated. We disclosed **a hidden functionality** of GitHub that allows a user to pretend to be a skillful developer.

System Architecture



There are 4 main components in our system:

1. Data collection.
2. Preprocessing.
3. Feature Engineering.
4. SVM-based Classifier.

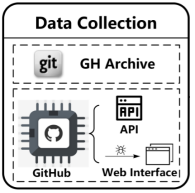


Data Collection

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


```
{
    "id":"7044401123",
    "type":"WatchEvent",
    "actor":{
        "id":1710912,
        "login":"yangwenmai",
        "display_login":"yangwenmai",
        "gravatar_id":"",
        "url":"https://api.github.com/users/yangwenmai",
        "avatar_url":"https://avatars.githubusercontent.com/u/1710912?"
    },
    "repo":{
        "id":75951828,
        "name":"wainshine/Chinese-Names-Corpus",
        "url":"https://api.github.com/repos/wainshine/Chinese-Names-Corpus"
    },
    "payload":{
        "action":"started"
    },
    "public":true,
    "created_at":"2018-01-01T00:00:02Z"
}
```

Data Collection

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

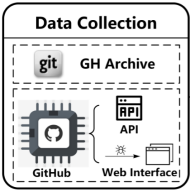
GH Archive records public GitHub event and archives it in JSON format. We collected all archive files from 2015 to 2019

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.


```
{
    "id":"7044401123",
    "type":"WatchEvent",
    "actor":{
        "id":1710912,
        "login":"yangwenmai",
        "display_login":"yangwenmai",
        "gravatar_id":"",
        "url":"https://api.github.com/users/yangwenmai",
        "avatar_url":"https://avatars.githubusercontent.com/u/1710912?"
    },
    "repo":{
        "id":75951828,
        "name":"wainshine/Chinese-Names-Corpus",
        "url":"https://api.github.com/repos/wainshine/Chinese-Names-Corpus"
    },
    "payload":{
        "action":"started"
    },
    "public":true,
    "created_at":"2018-01-01T00:00:02Z"
}
```

```
"events_url":"https://api.github.com/events",
"feeds_url":"https://api.github.com/feeds",
"followers_url":"https://api.github.com/user/followers",
"following_url":"https://api.github.com/user/following{/target}",
"gists_url":"https://api.github.com/gists{/gist_id}",
```

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.

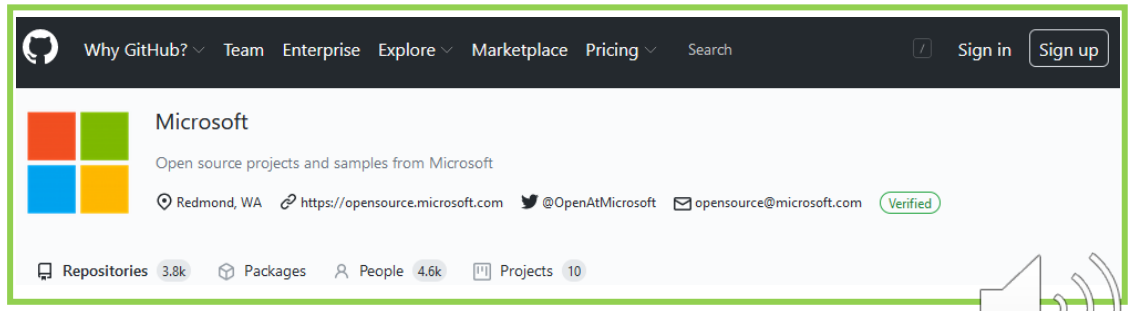
3. GitHub Web Interface (https://www.github.com/)

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.


```
{
    "id":"7044401123",
    "type":"WatchEvent",
    "actor":{
        "id":1710912,
        "login":"yangwenmai",
        "display_login":"yangwenmai",
        "gravatar_id":"",
        "url":"https://api.github.com/users/yangwenmai",
        "avatar_url":"https://avatars.githubusercontent.com/u/1710912?"
    },
    "repo":{
        "id":75951828,
        "name":"wainshine/Chinese-Names-Corpus",
        "url":"https://api.github.com/repos/wainshine/Chinese-Names-Corpus"
    },
    "payload":{
        "action":"started"
    },
    "public":true,
    "created_at":"2018-01-01T00:00:02Z"
}
```



```
"events_url":"https://api.github.com/events",
"feeds_url":"https://api.github.com/feeds",
"followers_url":"https://api.github.com/user/followers",
"following_url":"https://api.github.com/user/following{/target}",
"gists_url":"https://api.github.com/gists{/gist_id}",
```

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

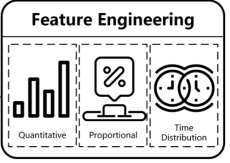
Normal Accounts

1,550 users who made major active contributions in well-known repositories.

a. the contributors of popular GitHub repositories

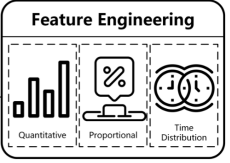
b. those who have proposed valuable issues on popular repositories.

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

No.	Type	Created_at	Repo	Star Count	Repo Type
1	WatchEvent	2018-12-23T07:37:38Z	mahmoud/awesome-python-applications	6,369	Popular Author
2	WatchEvent	2018-12-23T07:37:44Z	FavioVazquez/ds-cheatsheets	3,597	Popular Author
3	WatchEvent	2018-12-23T07:38:02Z	facebookresearch/flashlight	762	Organization
4	WatchEvent	2018-12-23T07:39:19Z	trekhleb/homemade-machine-learning	8,749	Popular Author
5	WatchEvent	2018-12-23T07:39:30Z	ghost1****/t****	716	Promotion Target
6	WatchEvent	2018-12-23T07:39:38Z	FAQGURU/FAQGURU	3,529	Popular Author
7	WatchEvent	2018-12-23T07:39:46Z	alibaba/x-deeplearning	2,145	Organization
8	WatchEvent	2018-12-23T07:40:00Z	alash3al/redix	689	Popular Author

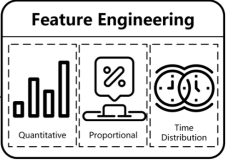


Feature Engineering

1. Action types

Promotion accounts: star and fork operations occupy most action types

Normal accounts: action types are much more diverse



Feature Engineering

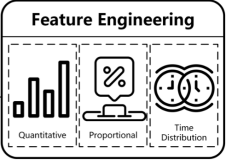
1. Action types

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

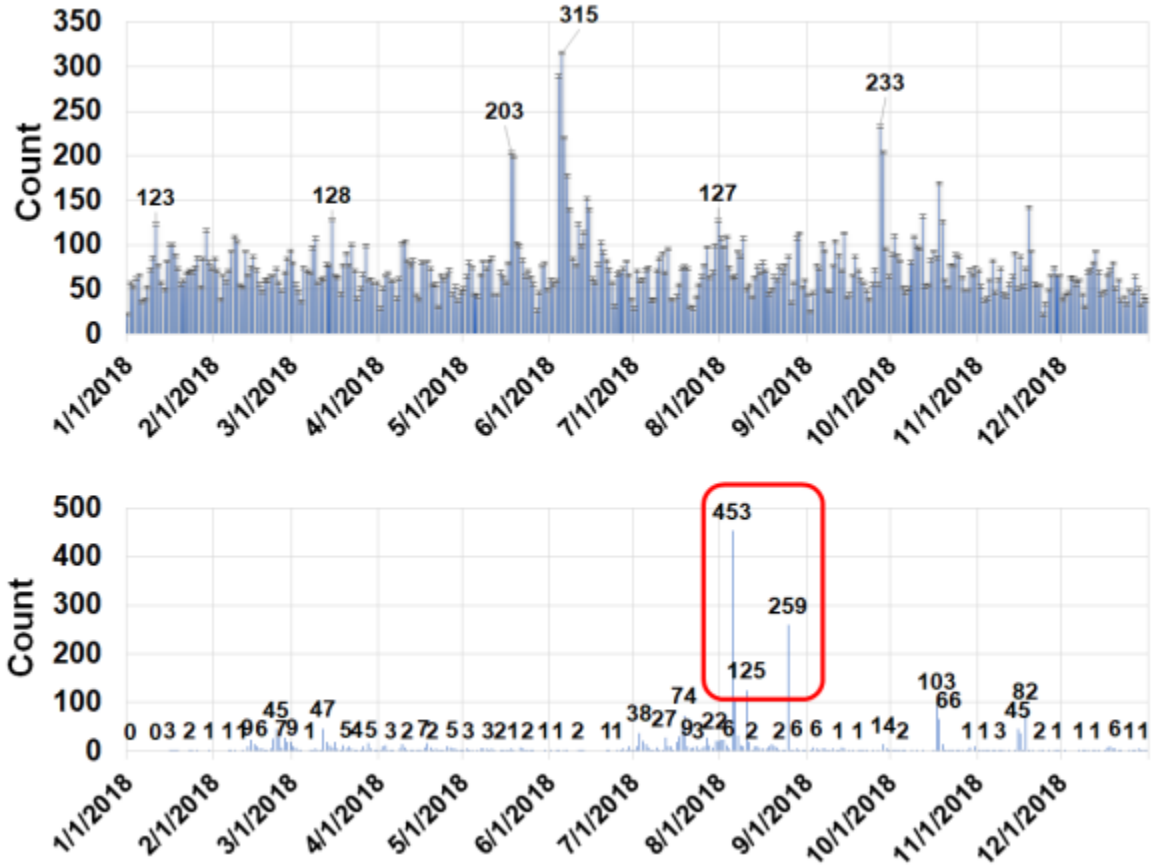
A promotion account performed star operations and the interval between two operations is less than 20 seconds.

Detection and Result Validation

- We built SVM-based classifier and detected **63,872** suspected promotion accounts

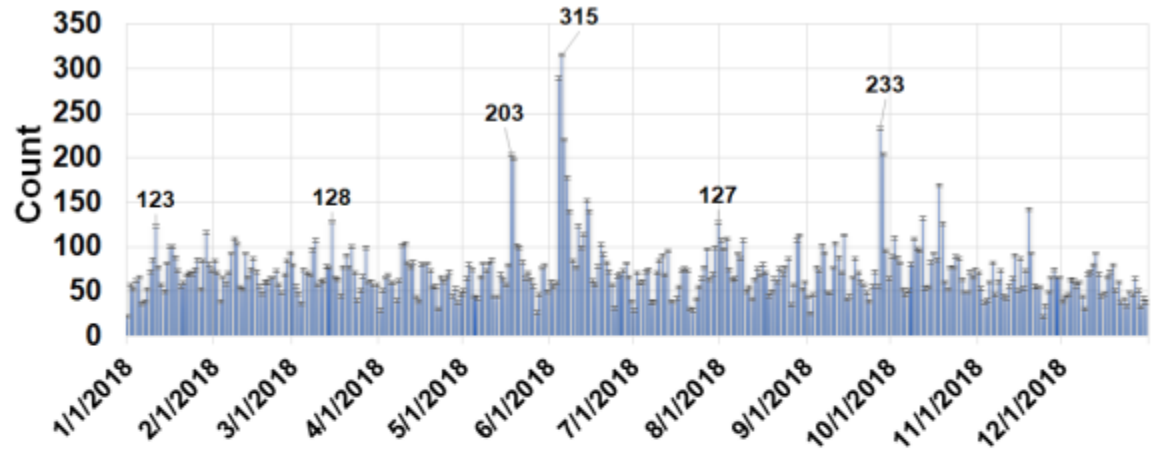
Detection and Result Validation

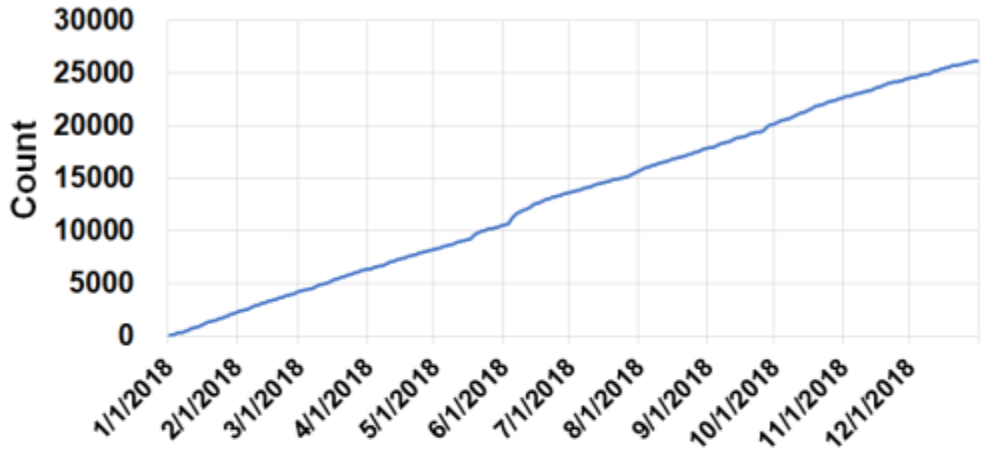
- We built SVM-based classifier and detected **63,872** suspected promotion accounts

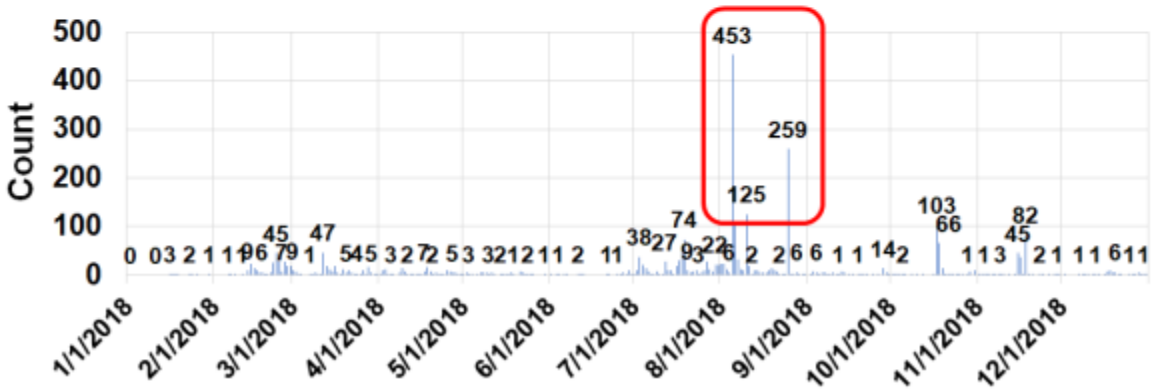


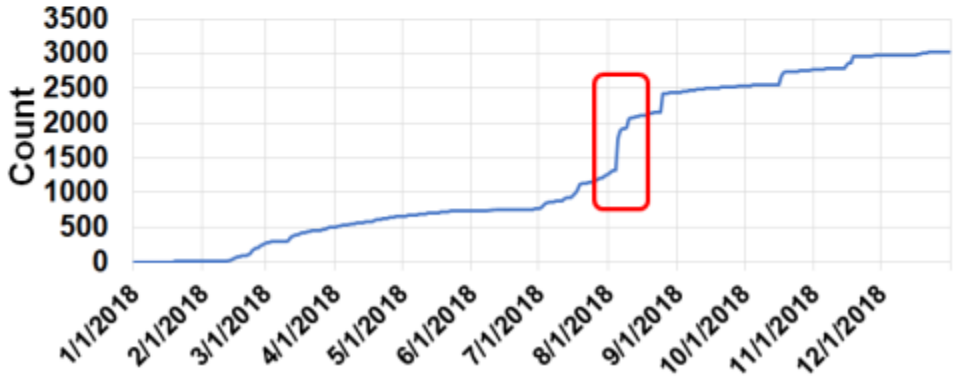
Detection and Result Validation

- We built SVM-based classifier and detected **63,872** suspected promotion accounts



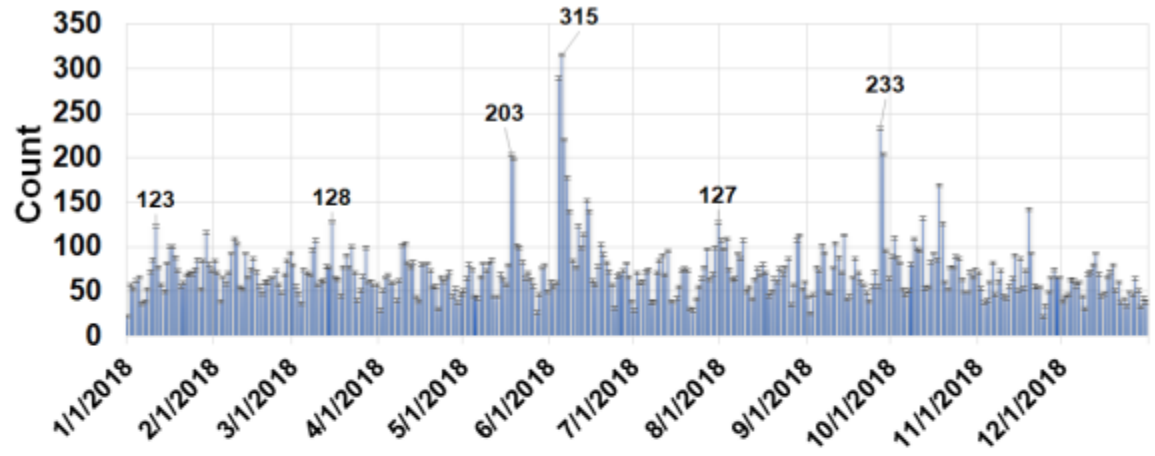


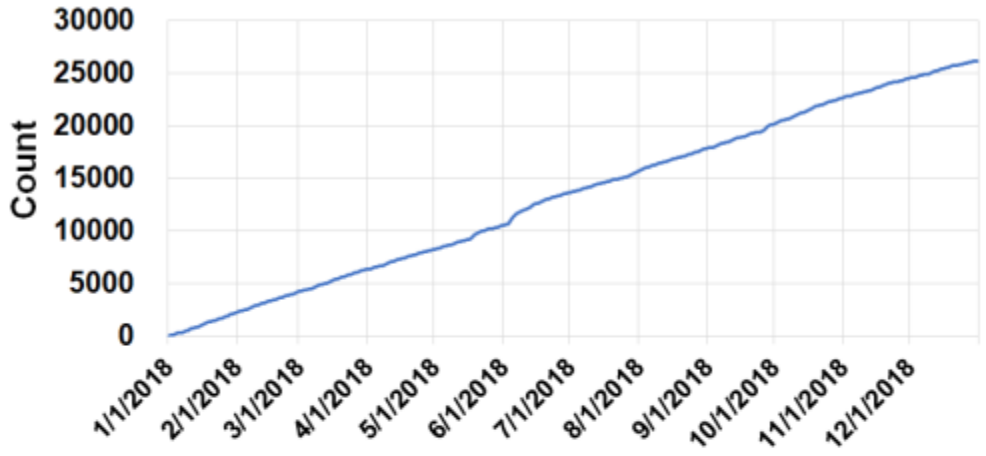


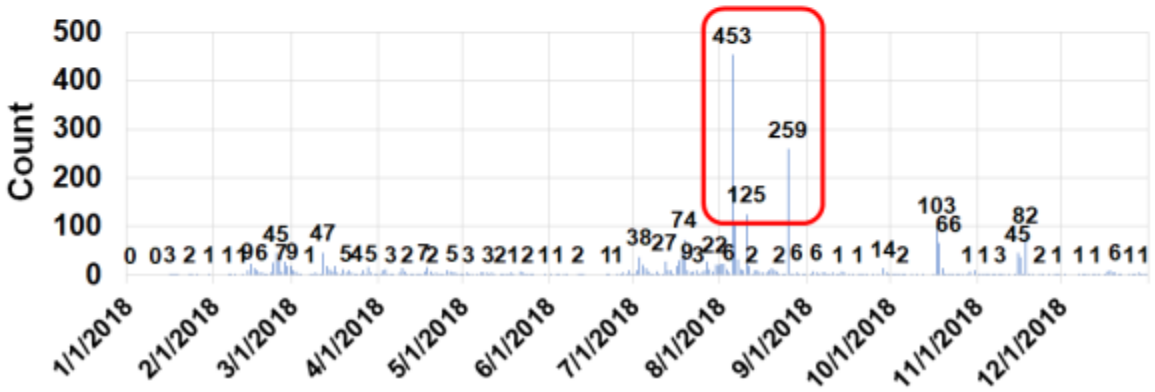


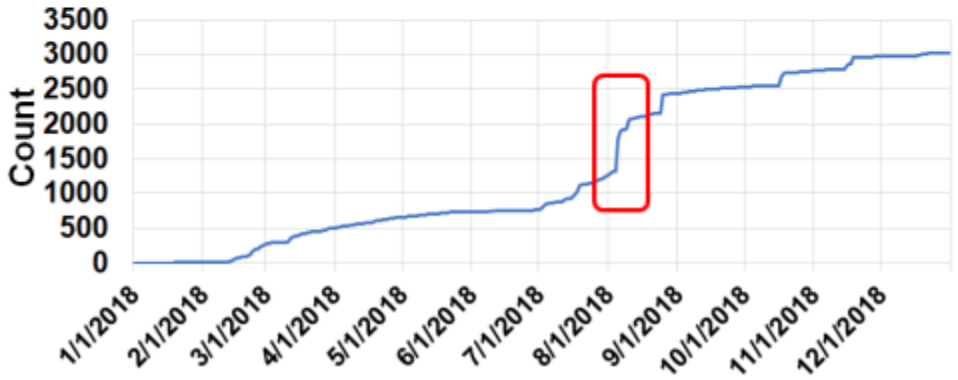
Detection and Result Validation

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

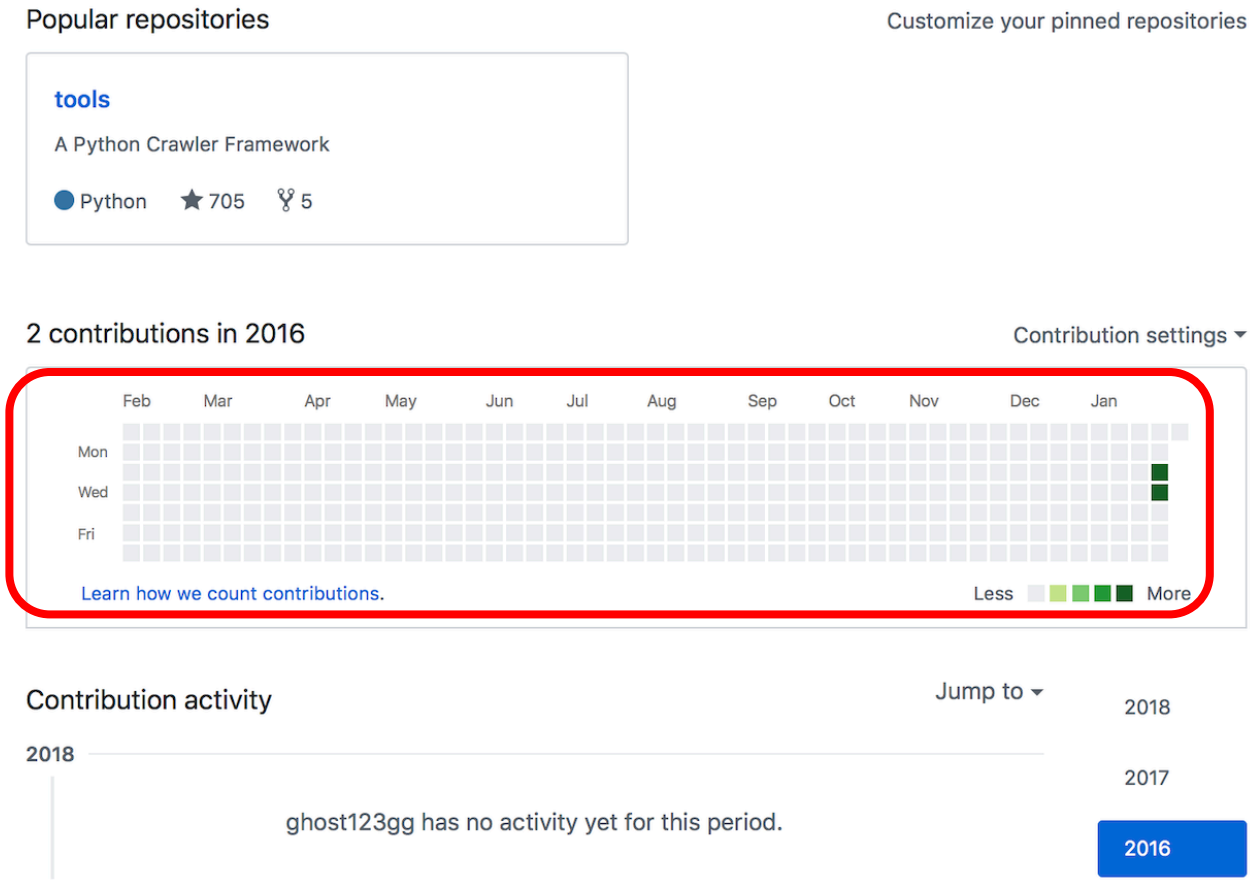




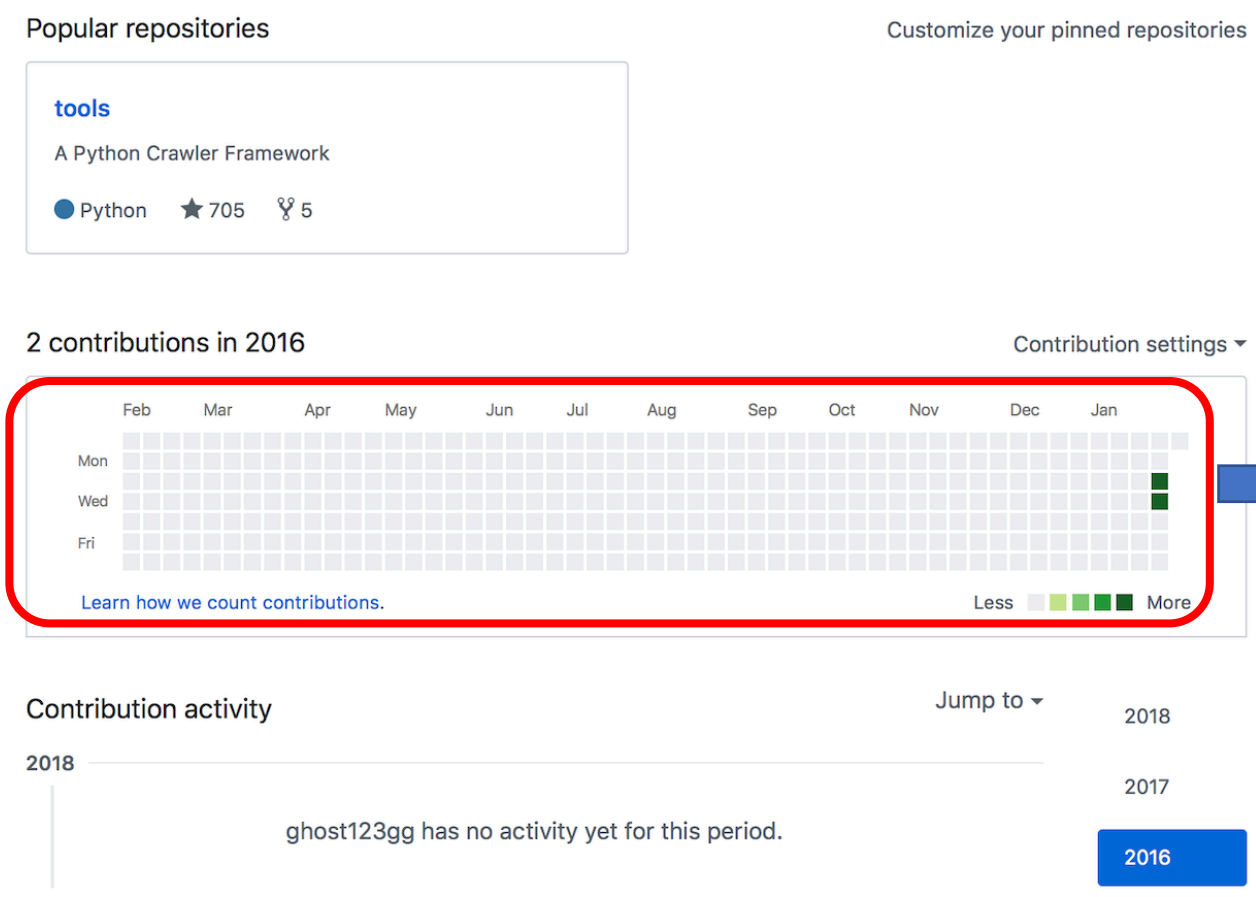


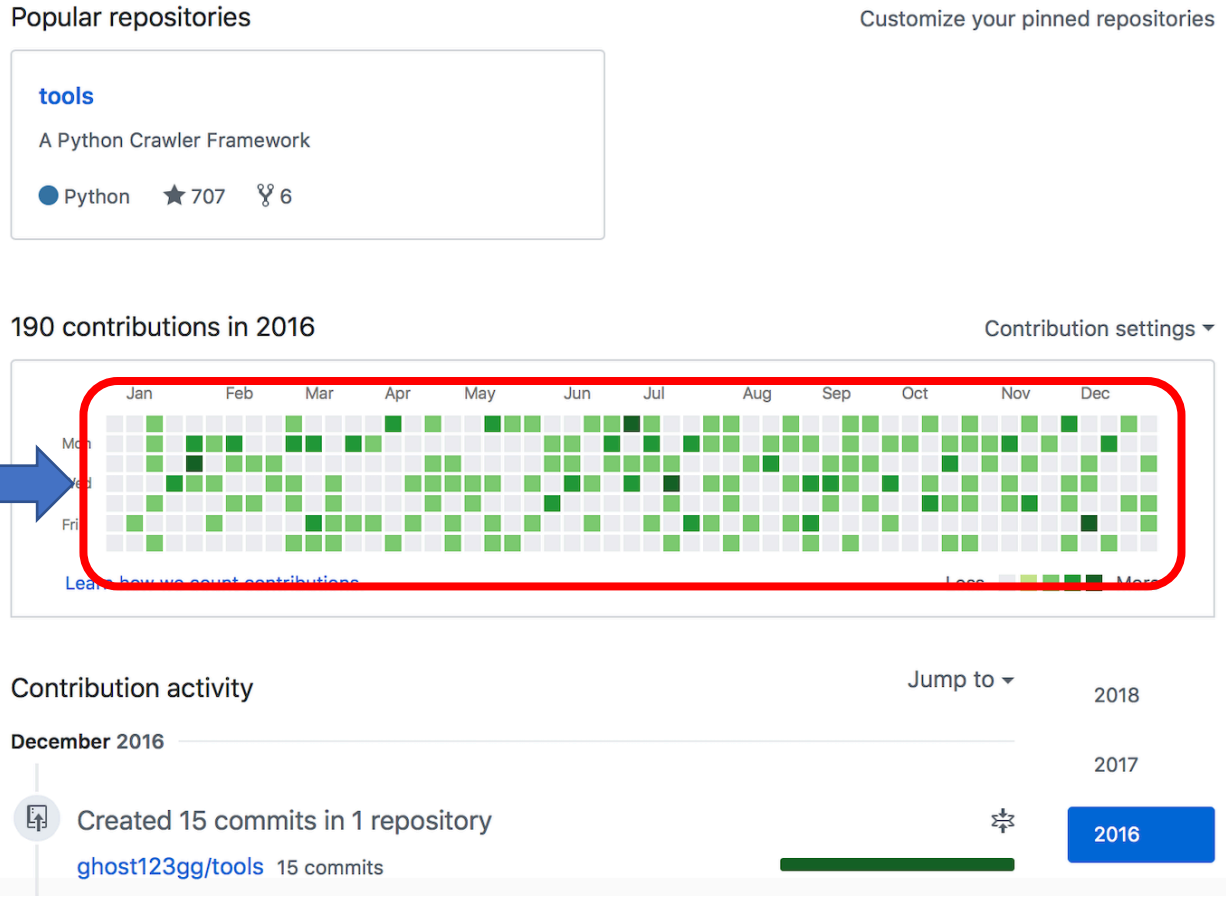


Forging Retroactive Commits

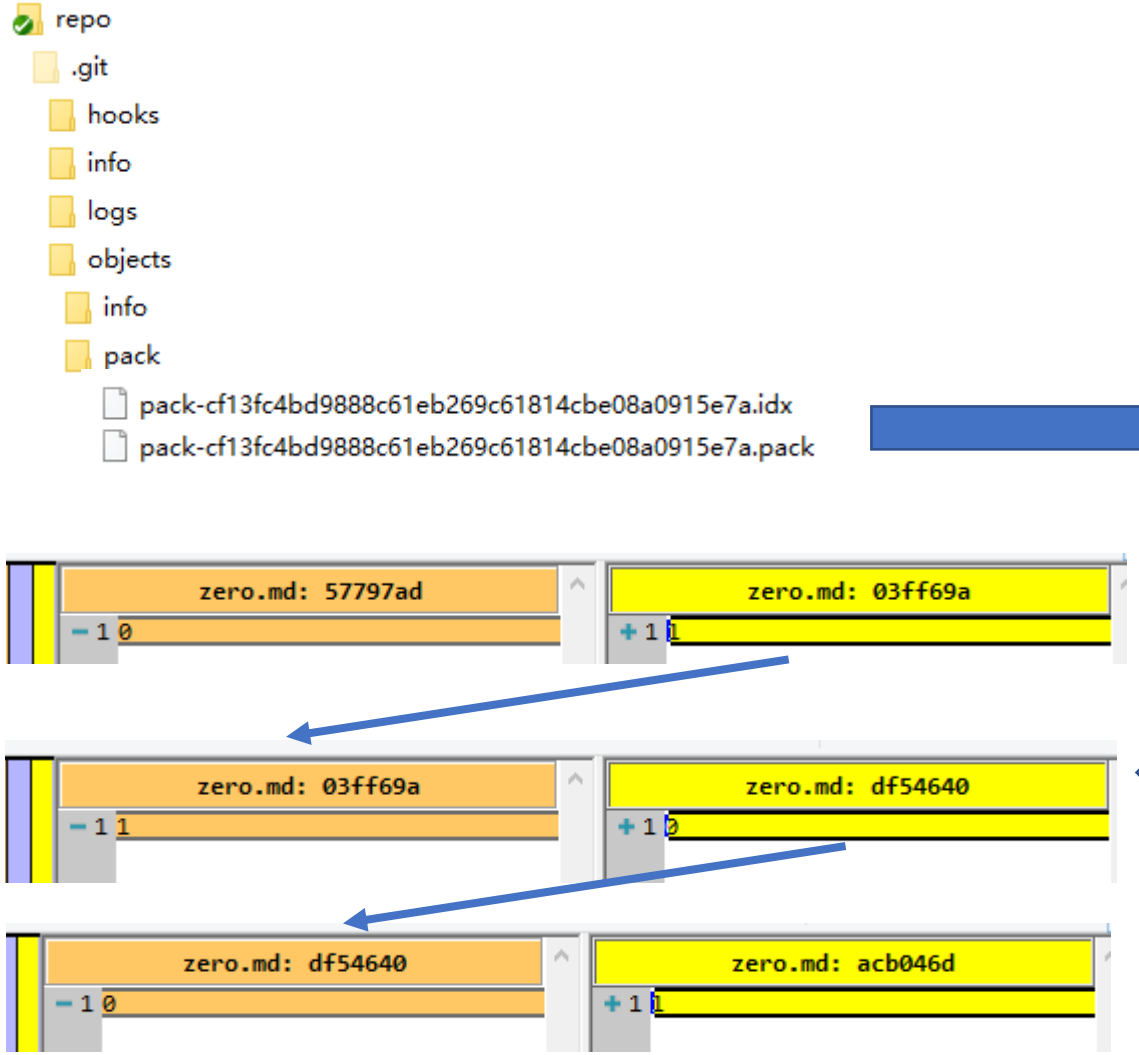


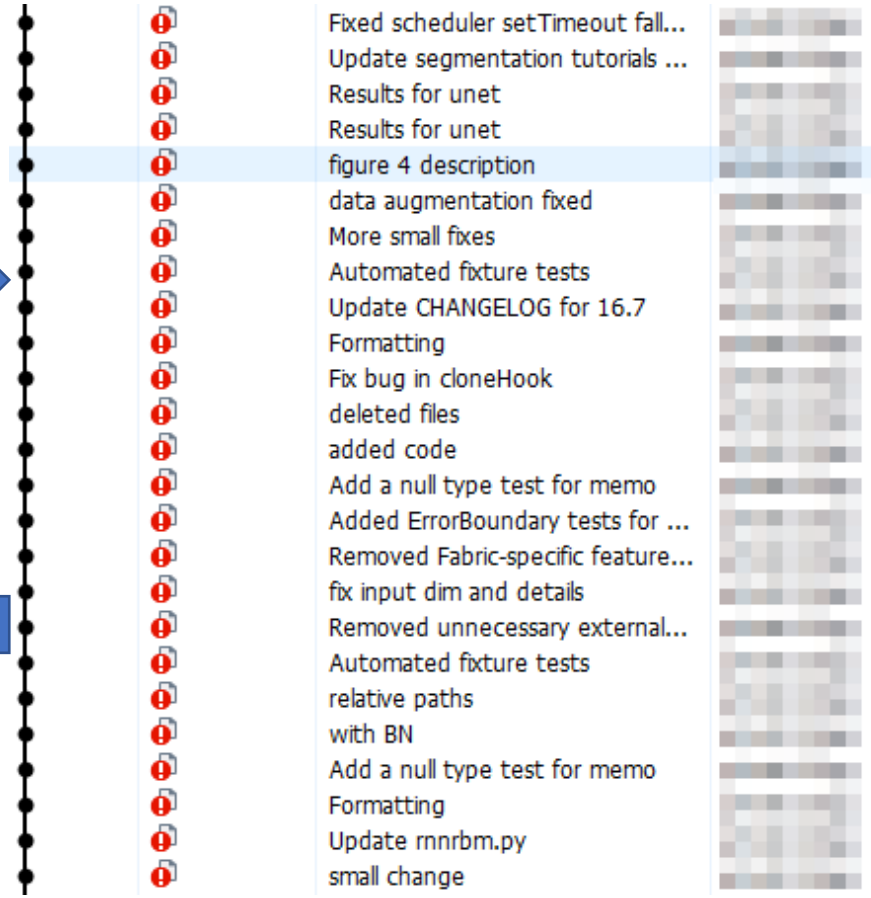
Forging Retroactive Commits



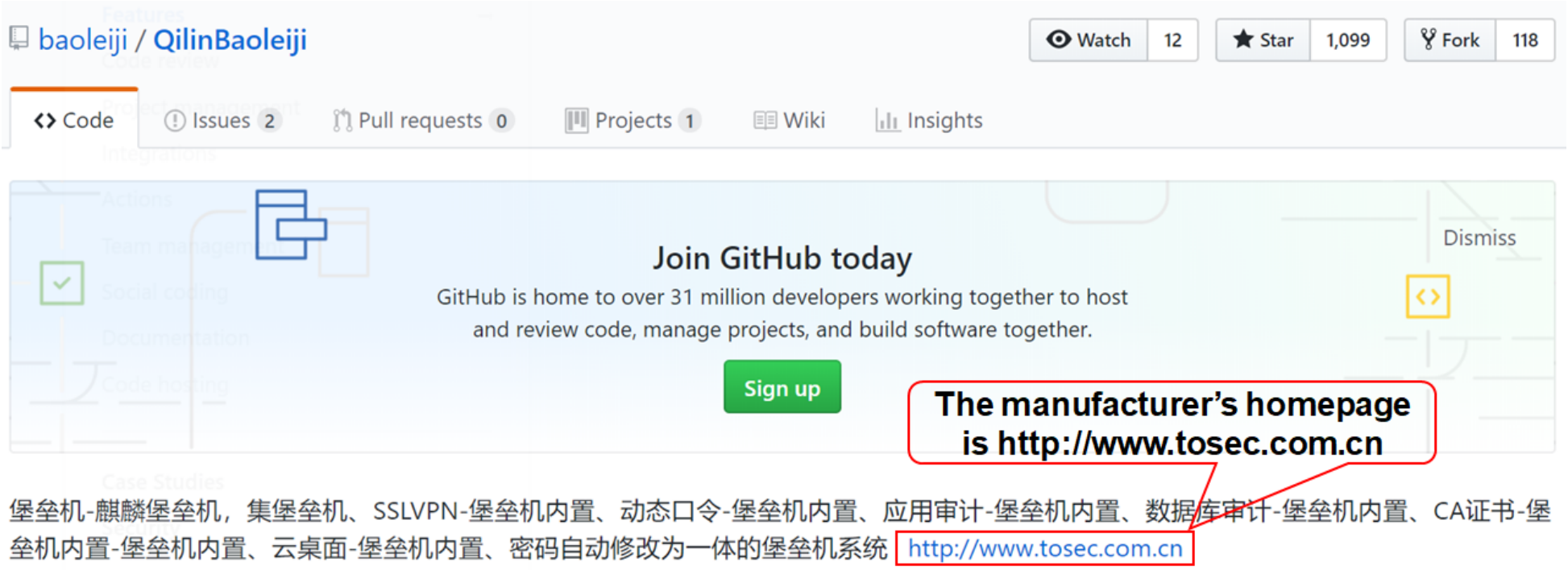


Forging Retroactive Commits





Small business promotion

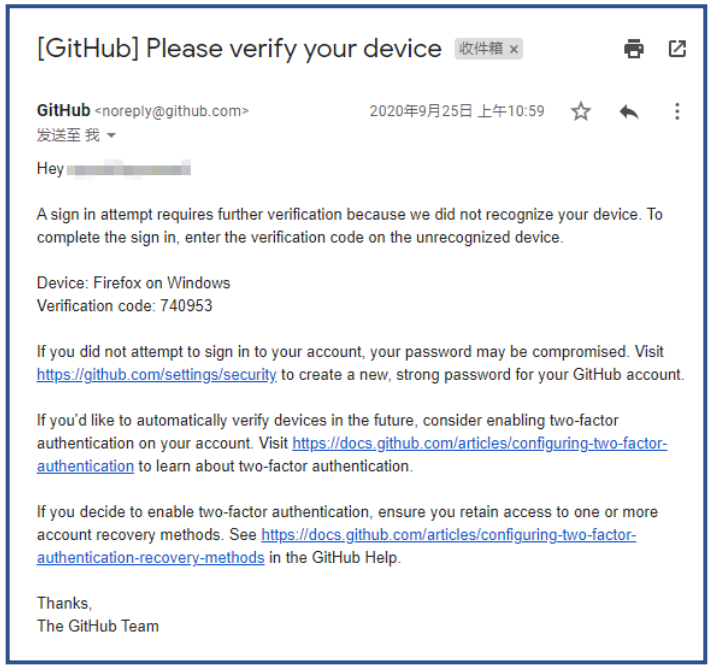


The company first publishes the repository on GitHub and places the link in their source code, homepage hyperlink or “readme.md” to their online shop’s homepage, showing a hyperlink from GitHub.

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.



[GitHub] Please verify your device 收件箱 ×

GitHub <noreply@github.com> 2020年9月25日 上午10:59

发送至 我

Hey [redacted]

A sign in attempt requires further verification because we did not recognize your device. To complete the sign in, enter the verification code on the unrecognized device.

Device: Firefox on Windows
Verification code: 740953

If you did not attempt to sign in to your account, your password may be compromised. Visit https://github.com/settings/security to create a new, strong password for your GitHub account.

If you'd like to automatically verify devices in the future, consider enabling two-factor authentication on your account. Visit https://docs.github.com/articles/configuring-two-factor-authentication to learn about two-factor authentication.

If you decide to enable two-factor authentication, ensure you retain access to one or more account recovery methods. See https://docs.github.com/articles/configuring-two-factor-authentication-recovery-methods in the GitHub Help.

Thanks,
The GitHub Team

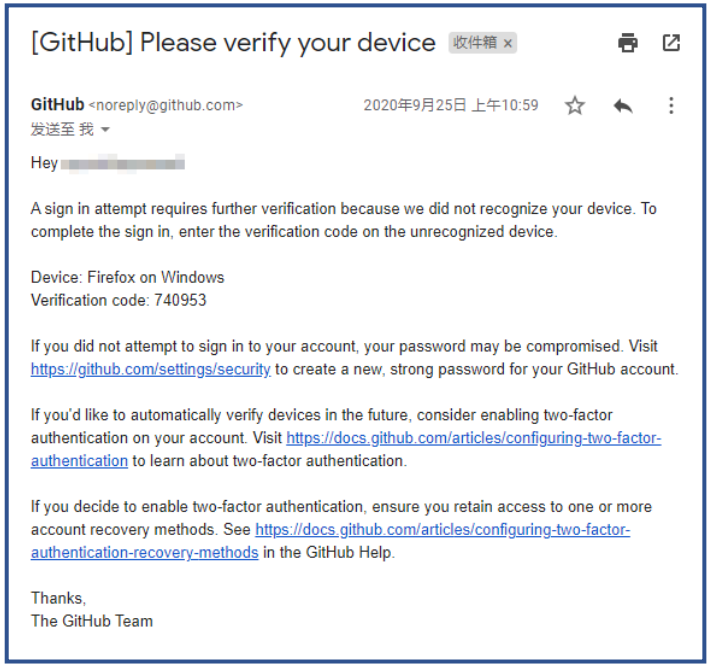
Confirm email we received

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.


[GitHub] Please verify your device 收件箱 ×

GitHub <noreply@github.com> 2020年9月25日 上午10:59
发送至 我

Hey [redacted]

A sign in attempt requires further verification because we did not recognize your device. To complete the sign in, enter the verification code on the unrecognized device.

Device: Firefox on Windows
Verification code: 740953

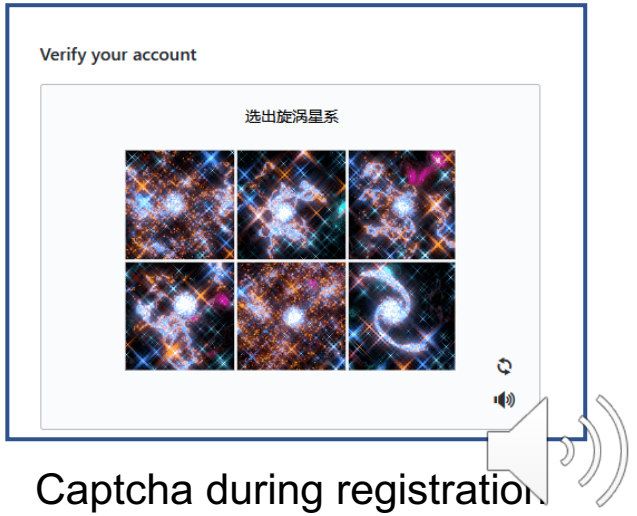
If you did not attempt to sign in to your account, your password may be compromised. Visit https://github.com/settings/security to create a new, strong password for your GitHub account.

If you'd like to automatically verify devices in the future, consider enabling two-factor authentication on your account. Visit https://docs.github.com/articles/configuring-two-factor-authentication to learn about two-factor authentication.

If you decide to enable two-factor authentication, ensure you retain access to one or more account recovery methods. See https://docs.github.com/articles/configuring-two-factor-authentication-recovery-methods in the GitHub Help.

Thanks,
The GitHub Team

Confirm email we received



Captcha during registration

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