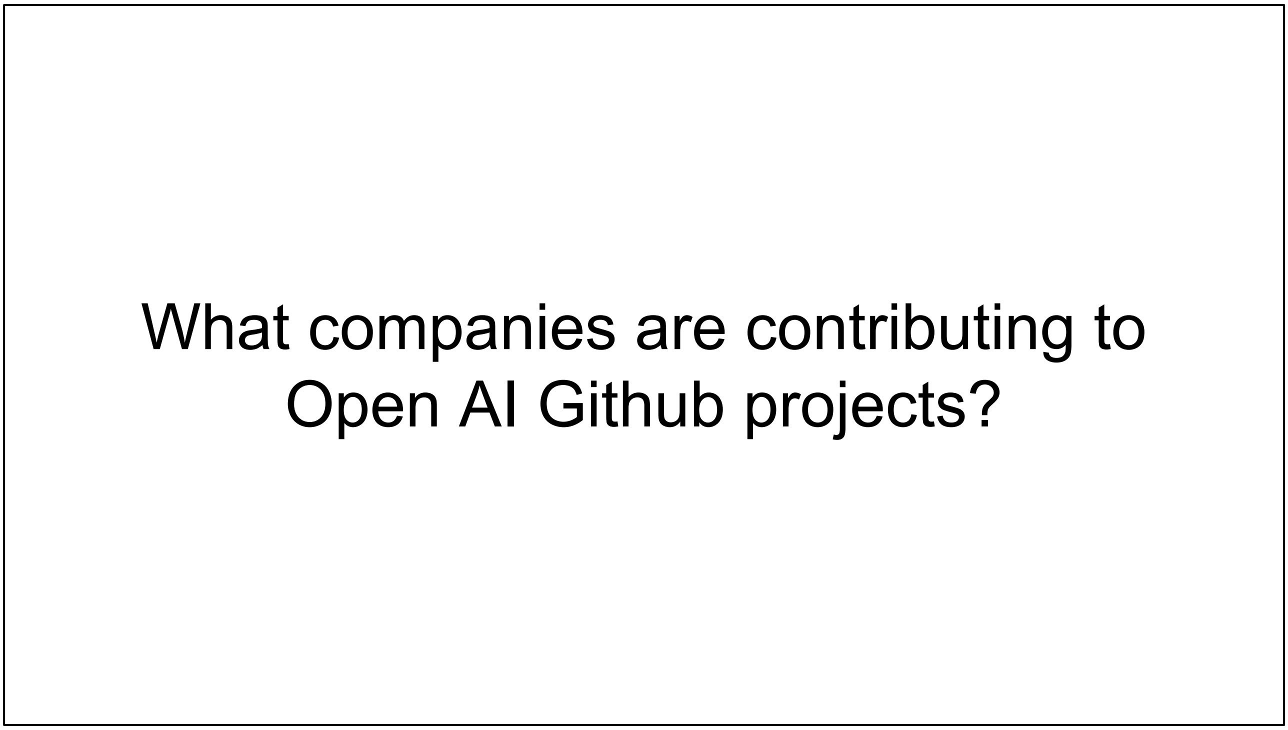
Github Contributors

Extracting Emails From Commits



What companies are contributing to Open AI Github projects?

Problem

- Not all Github Users specify a Company
- Only the current Company is available
- Github API does not provide email addresses

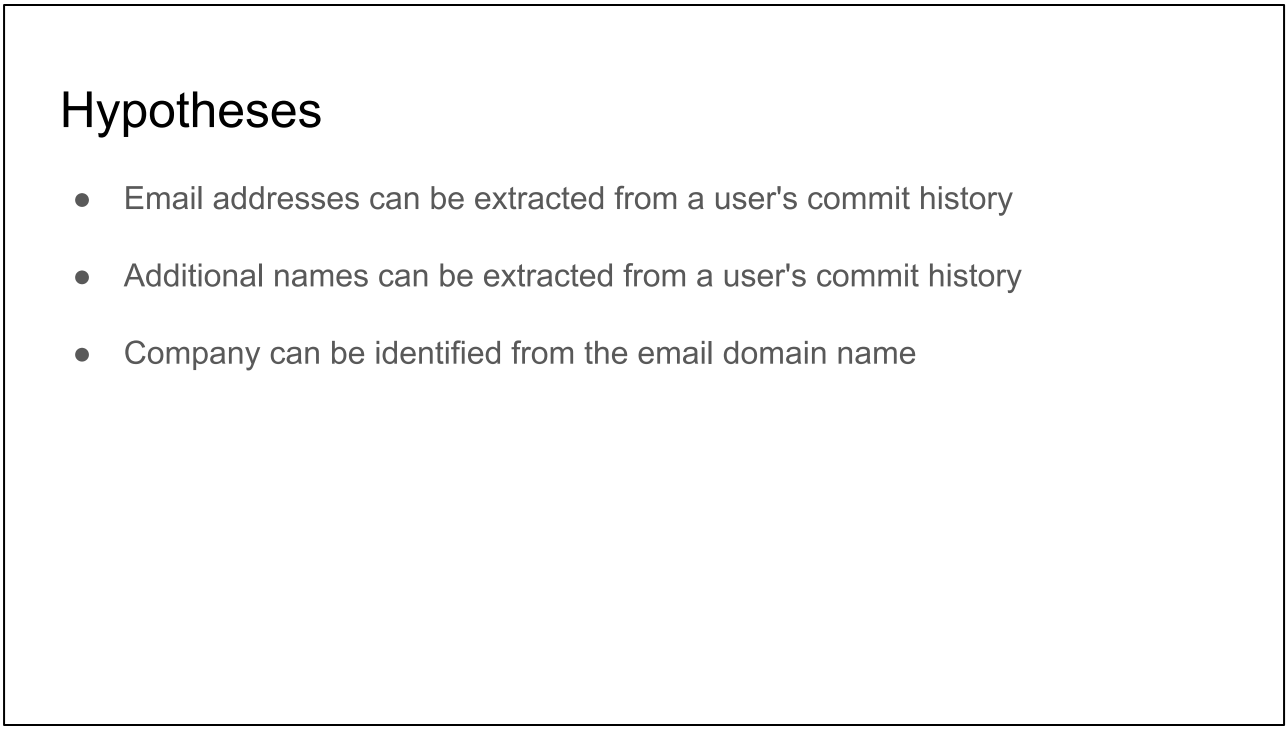
Git SHA’s require email as input

```
sha1(
    metadata
        commit message
        committer
        commit date
        author
        authoring date
    Hash-Of-Entire-Working-Directory
)
```

Solution

Get Email from Commit History!

```
$ git shortlog -s -ne
   532  Tianqi Chen <tqchen@users.noreply.github.com>
   438  Bing Xu <antinucleon@gmail.com>
   321  muli <muli@cs.cmu.edu>
   304  Chiyuan Zhang <pluskid@gmail.com>
   263  Eric Junyuan Xie <piiswrong@users.noreply.github.com>
   262  tqchen <tianqi.tchen@gmail.com>
   208  Yizhi Liu <javelinjs@gmail.com>
   206  Mu Li <muli@cs.cmu.edu>
   171  Junyuan Xie <eric.jy.xie@gmail.com>
   105  sneakerkg <xiaotj1990327@gmail.com>
    97  Bing Xu <antinucleon@users.noreply.github.com>
    91  Yutian Li <hotpxless@gmail.com>
    86  Chuntao Hong <chuntao.hong@gmail.com>
    71  terrytangyuan <terrytangyuan@gmail.com>
    66  Xingjian Shi <xshiab@ust.hk>
    60  yajiedesign <yajiedesign@gmail.com>
    54  winsty <winsty@gmail.com>
    53  Yao Wang <kevinthesunwy@gmail.com>
```



Hypotheses

- Email addresses can be extracted from a user's commit history
- Additional names can be extracted from a user's commit history
- Company can be identified from the email domain name

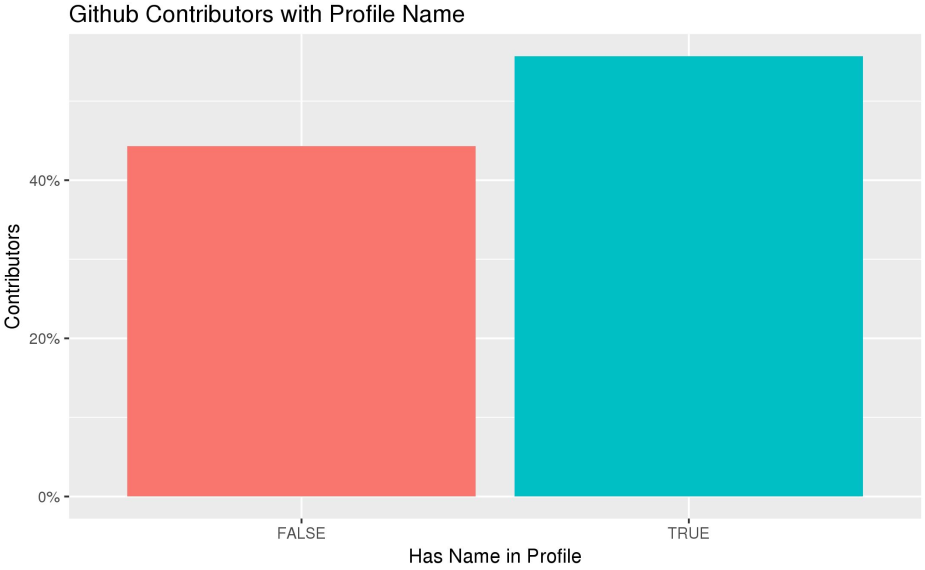
Null Hypotheses

- Users don't have sufficient commit history to yield email addresses or names
- The majority of email addresses are the default obfuscated Github ones
- Users don't use their company email, therefore company cannot be derived from the email domain.

Methodology

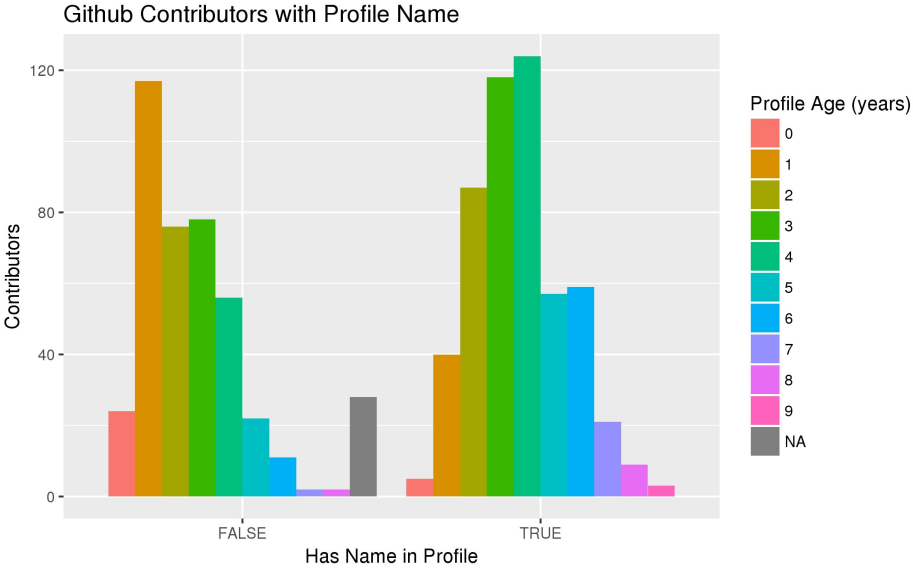
- Top Contributors to MXNet Project (~1000)
- Frequency and variety of Github Events, January 2015 - April 2017
- Github API used to collect profile data

Names in Profile

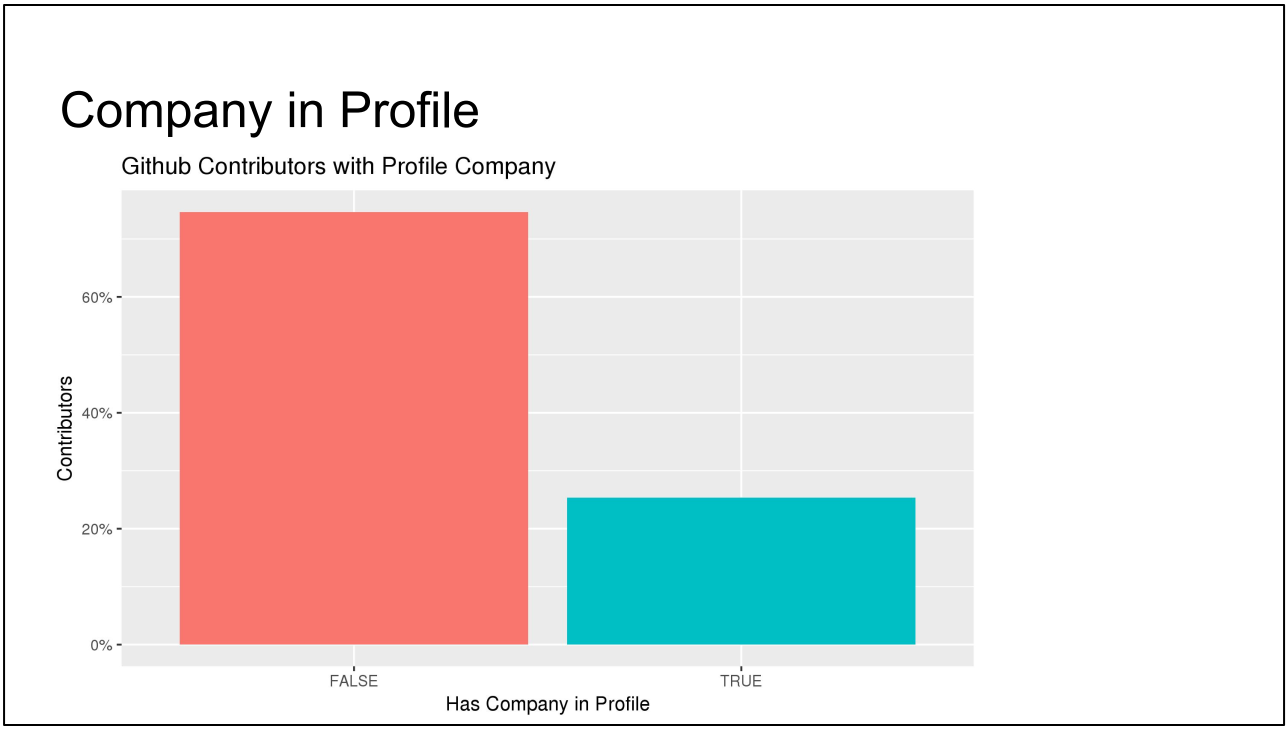


Just under 60% of the contributors had provided a name in their Github profile.

Age vs Name



Profiles with names skew towards being older with the majority being about 2-4 years old.

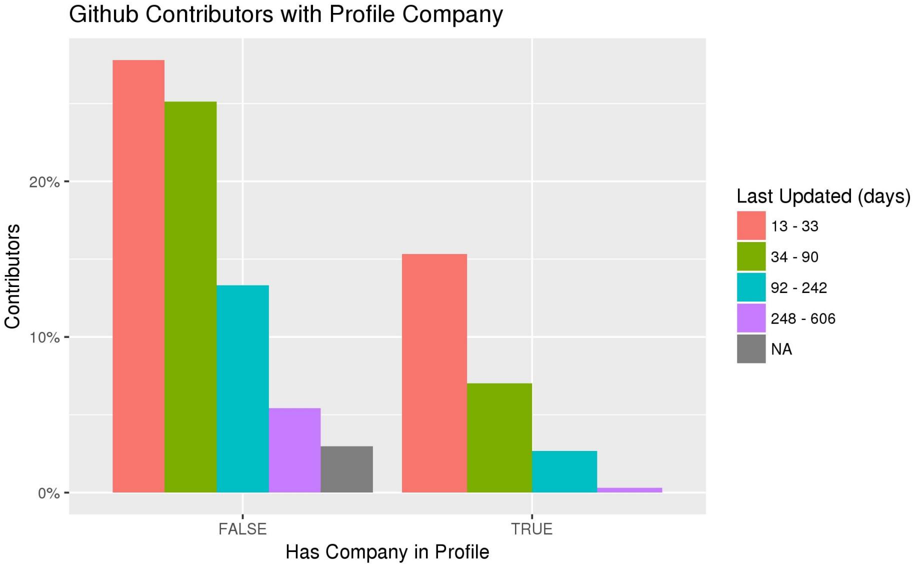


Company in Profile

Github Contributors with Profile Company

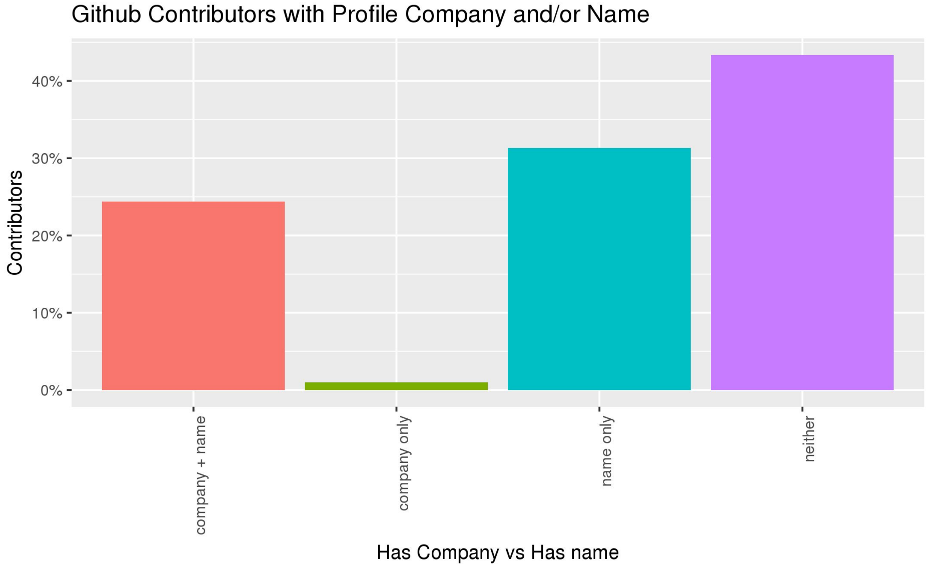
Most of the contributors don’t specify a company name in their profile.

Is Company Up to Date?



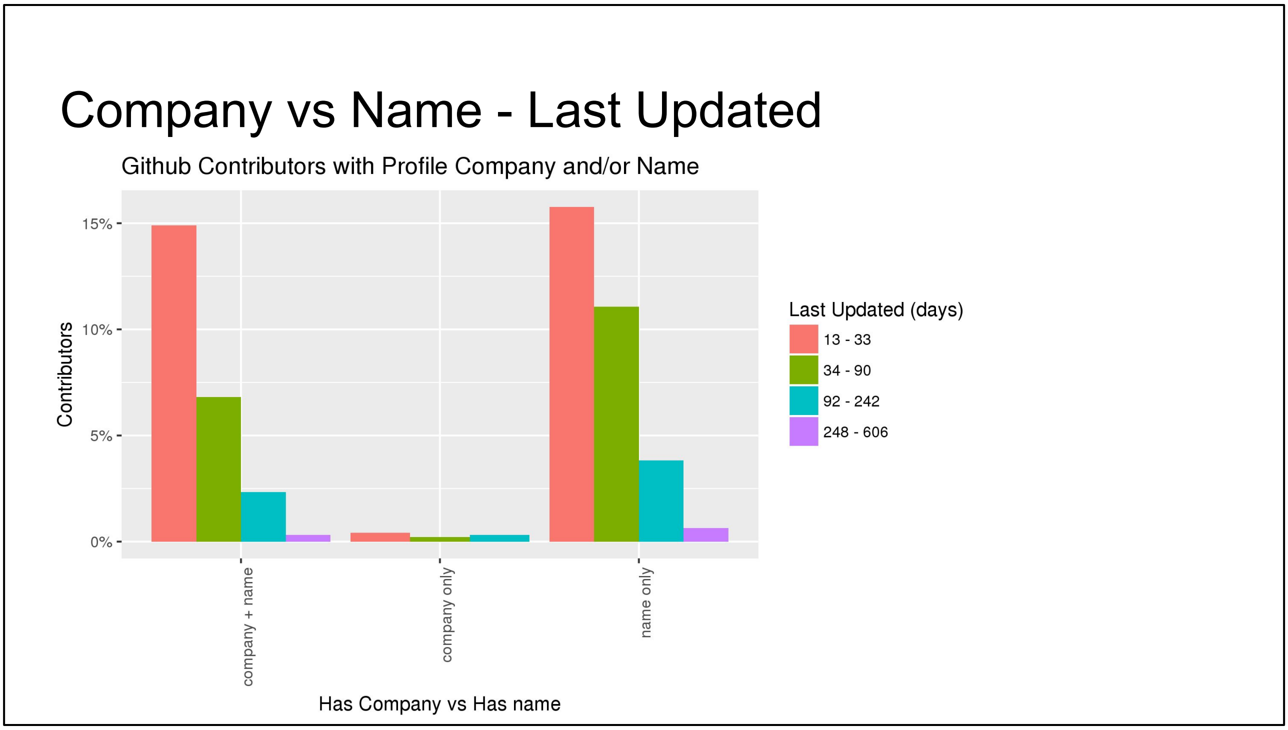
The majority of profiles, regardless of whether they had a company or not, were updated within 3 months. Profiles that had not been updated for the longest periods did not tend to have company information. This suggests the company information should be up to date.

Company vs Name



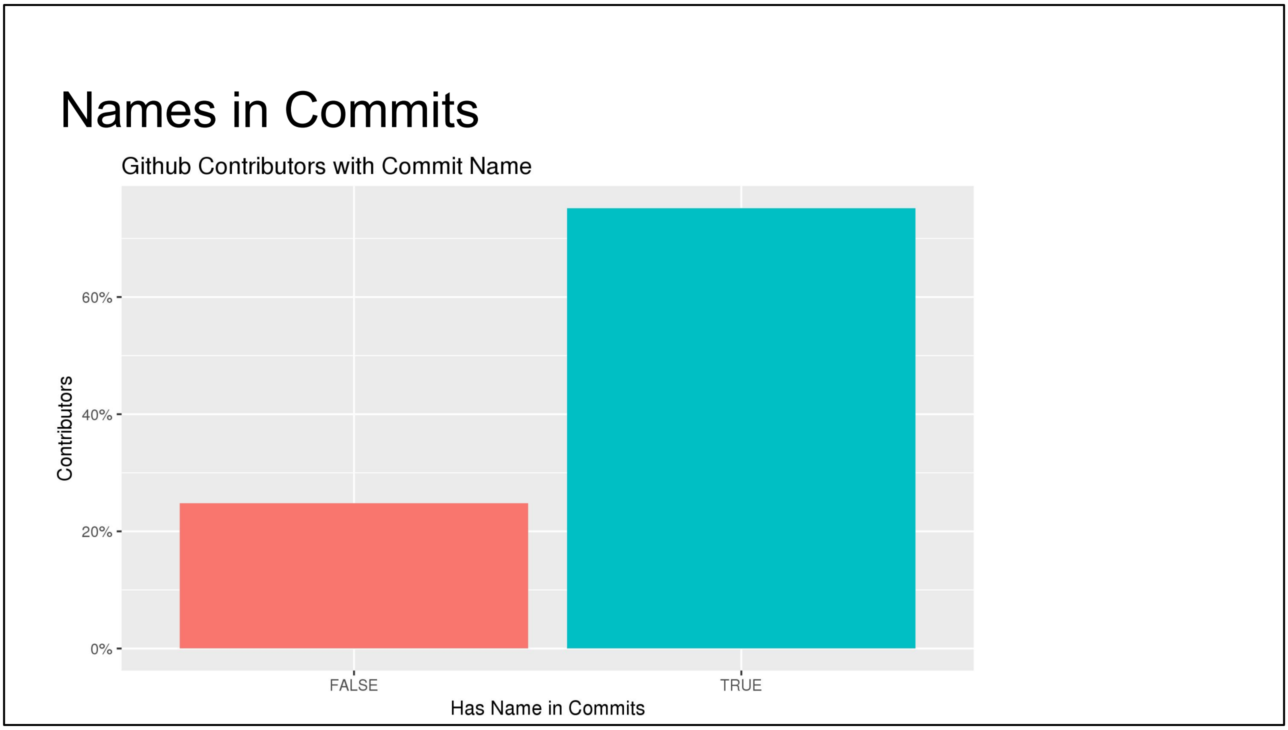
Another interesting question is whether profiles with a company name also have a name. For ones without a name, is the company information potentially less accurate?

Just over 55% of the profiles had either a company or a name. 30% had a name only and 25% had both a company and a name. A very small percent had company only. Because these appear in such a small proportion, we should look at the value of the field for those.



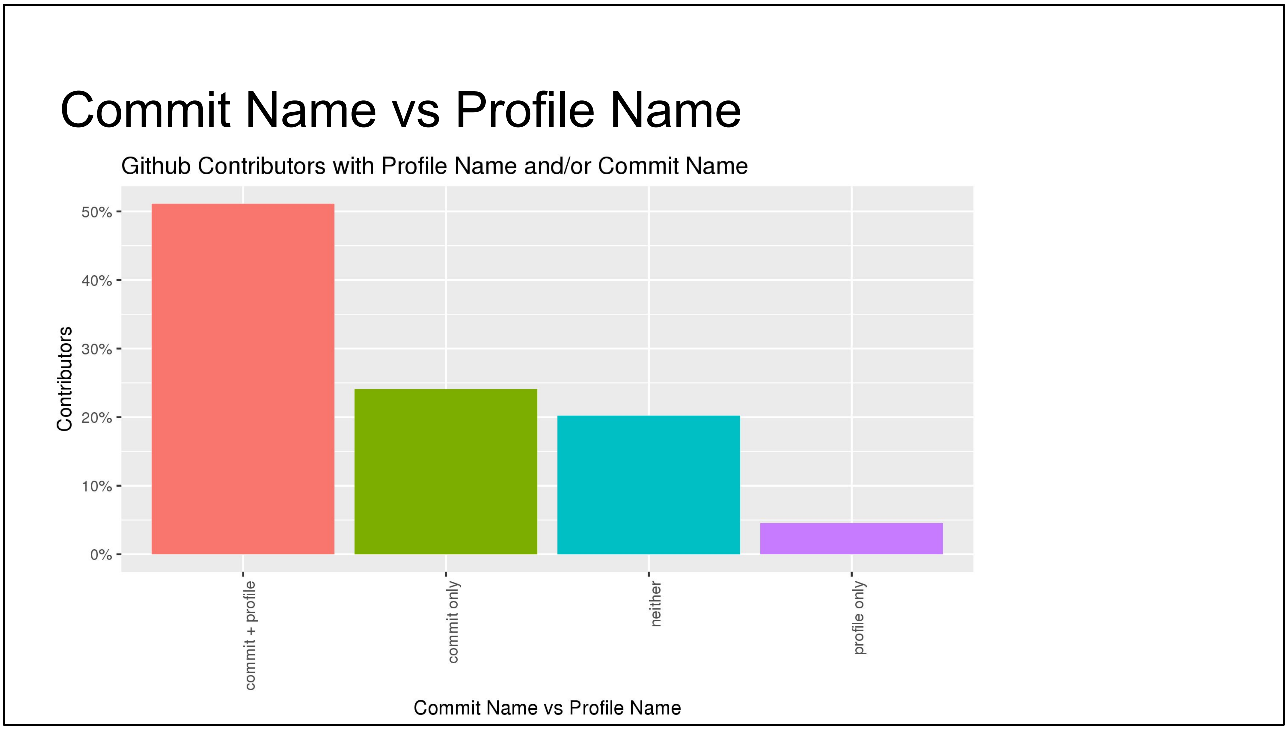
Company vs Name - Last Updated

We've already established that most of the Github profiles are fairly up to date, however it's worth looking at that distribution in terms of company vs name. We see a fairly similar distribution suggesting that the profiles are fairly up to date and neither parameter, name nor company, skews either way.



Names in Commits

Almost 70% of contributors had a name in their commit history.

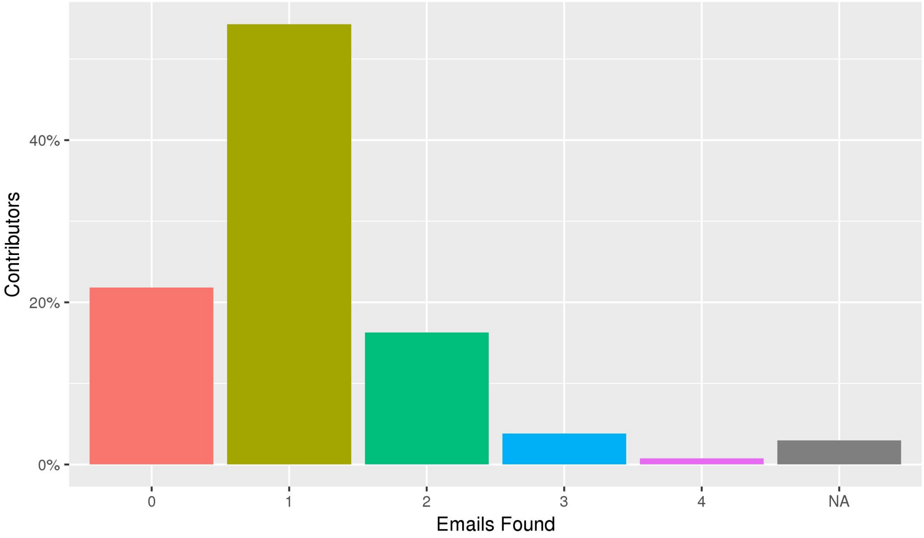


Commit Name vs Profile Name

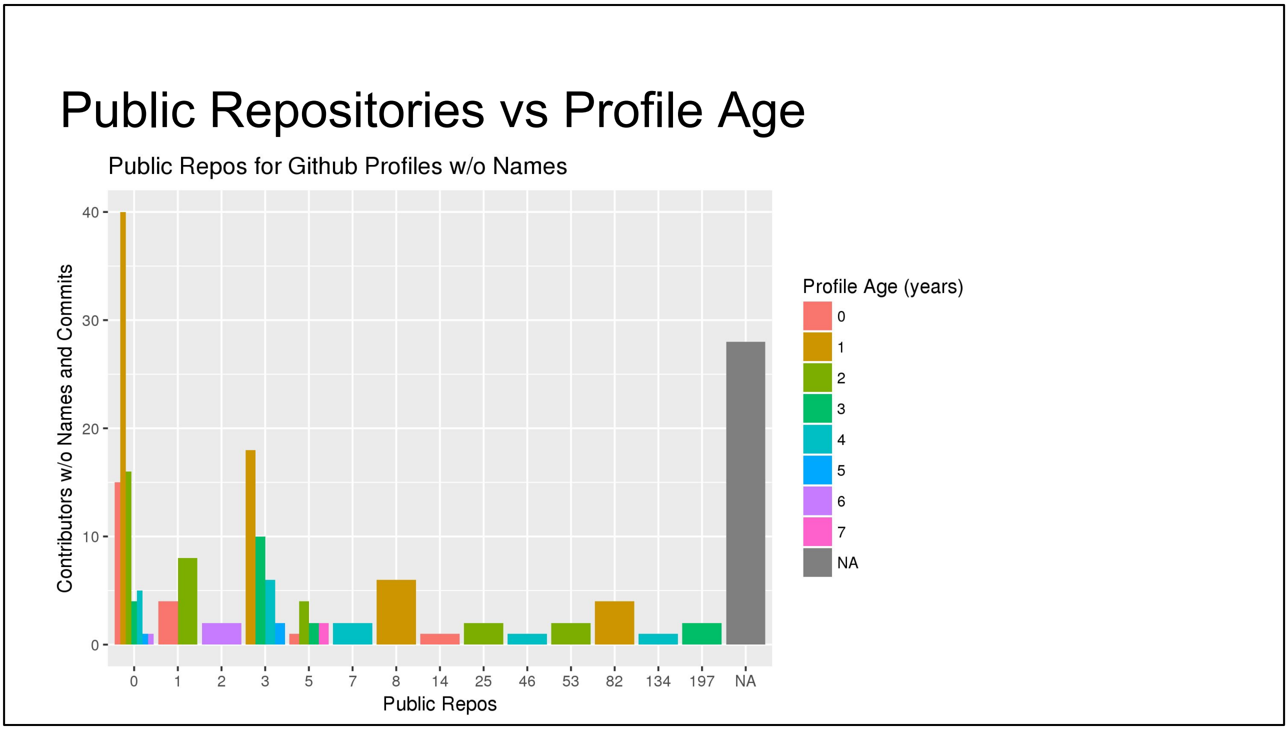
Github Contributors with Profile Name and/or Commit Name

We were able to get names for an additional 25% of contributors that had not provided a name in their profile. By extracting identifying information from commit histories, we are able to potentially identify 80% of the most active mxnet contributors.

Emails in Commits


Emails Found Per Contributor

Around 20% of contributors with active profiles had no email addresses and 5% had deleted profiles that no longer exist. The majority of contributors only had one email address.

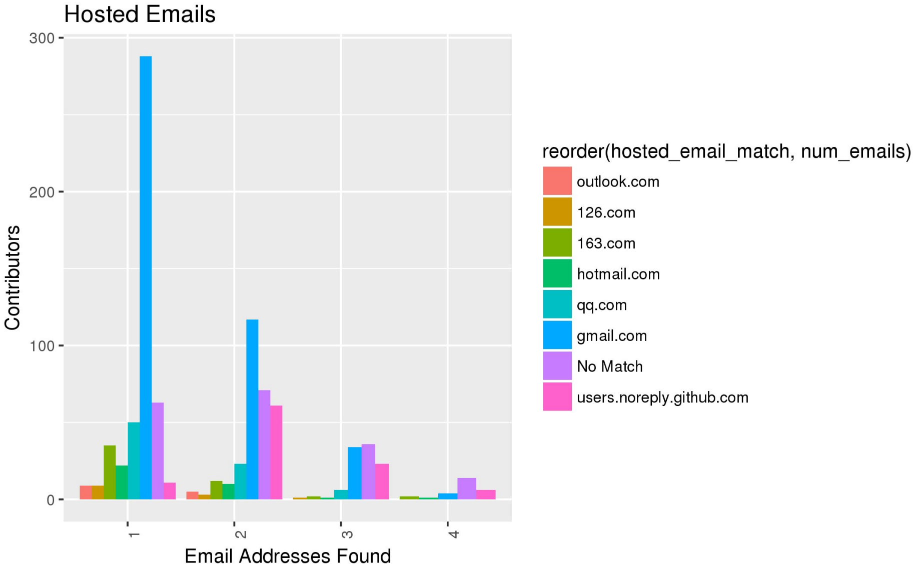


Public Repositories vs Profile Age

The number of public repos for unidentified contributors was examined and no correlation between profile age and number of profiles is indicated. Around 20% of contributors not identified via commit had a high number of public repos and a manual verification showed that information was available in commit histories depending on how the repositories were sorted in the API request. The script should probably be modified to sort the repos differently. In addition, looking for certain types of events linked to commits in the users' public event stream and extracting the repo name could be another method worth exploring.

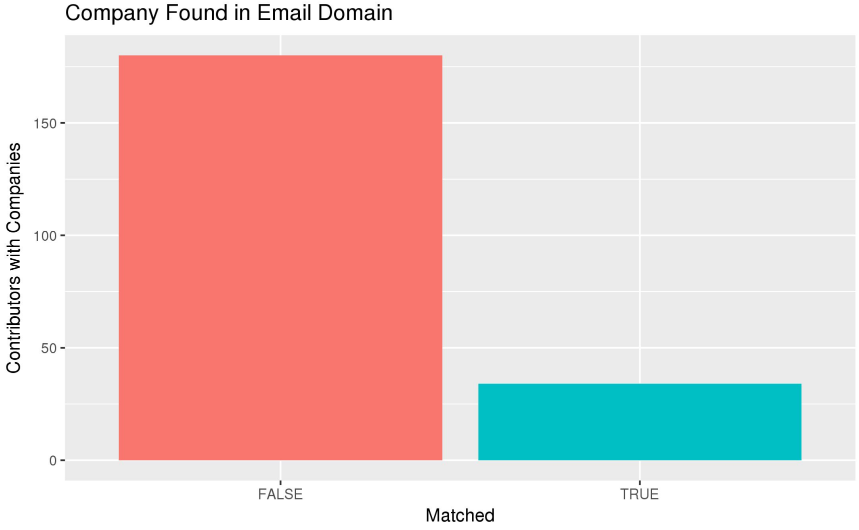
The majority of profiles that could not be identified through commit history only zero or just a small number of repositories. Future analysis should consider their event activity in the project that identified them as one of the most active contributors. It's possible refining that metric will reduce these.

Common Email Domains

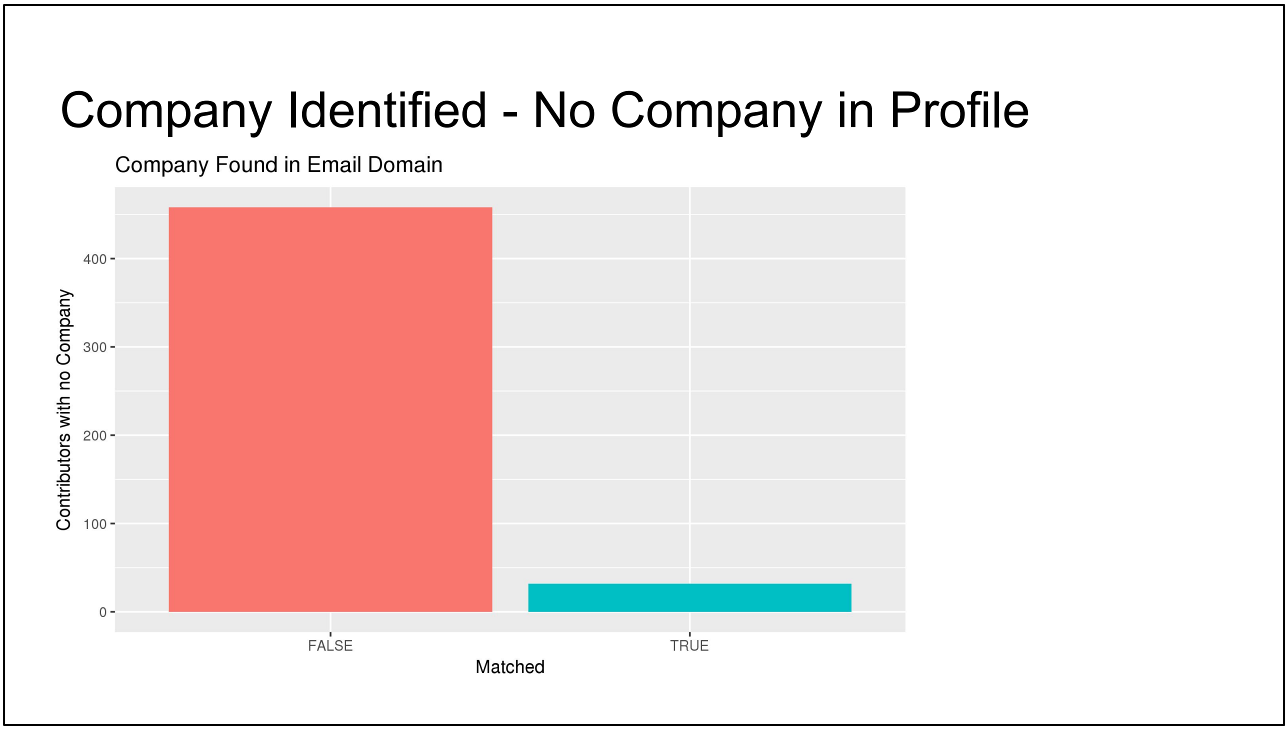


Domains were extracted from email addresses to compare with company names. To predict the usefulness of these addresses in making company matches, we can identify common domains for hosted email services. We find that the majority of contributors have a hosted email address (eg, gmail.com, hotmail.com) and only one email address. Contributors with more than one email address have a higher chance of having a non-hosted email address. The majority of contributors use gmail.com addresses.

Companies Identified - Company in Profile



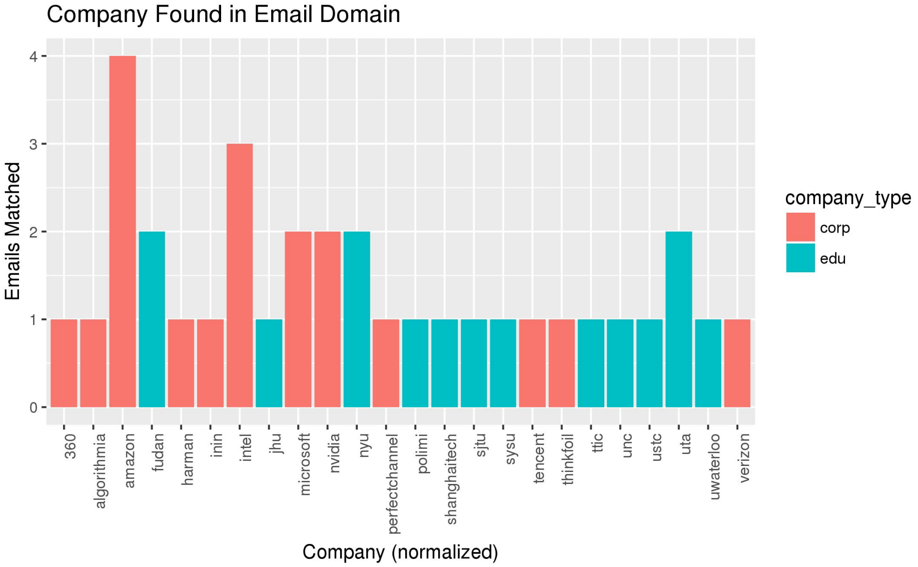
Can we identify the company from the email domain name? For contributors that have both email addresses in commits and a company name, how well do they match? How does the frequency of company names in profiles compare to the frequency of company-identifiable email addresses? Given that most users are using gmail accounts and only have one email address, we should expect this to be pretty low. Less than 20% of normalized company names were found in the email domain.



Company Identified - No Company in Profile

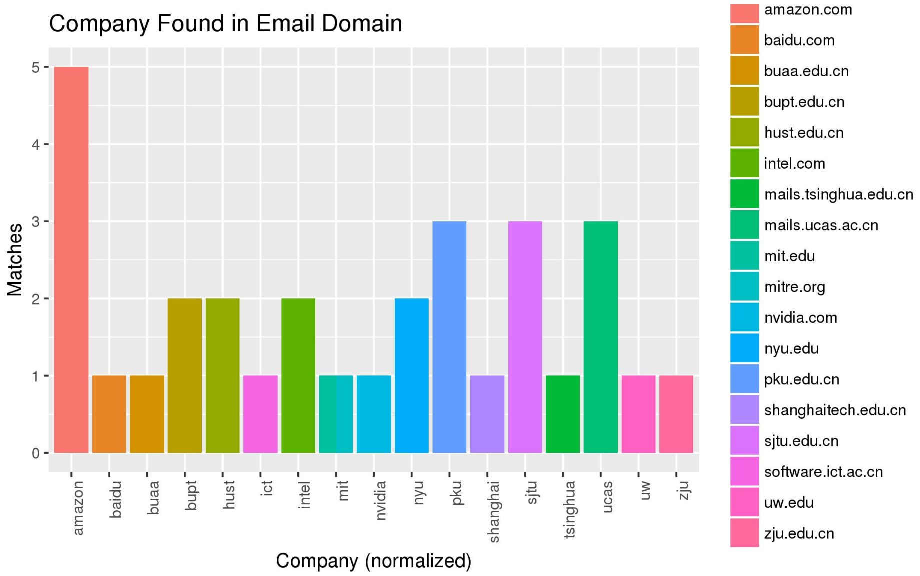
The above analysis suggests we may be able to find additional company matches by checking email domains against a list of normalized company names. Less than 10% of the Github profiles without companies were matched. The method used for this was very simple, the normalized company names found above were checked against the email domains. This depends on the Company name normalization matching their email domain name, and the Github user having committed under their work email address. This could pick up old employers or there could be a false match if the company name is an acronym or very short.

Matched Companies - Company in Profile



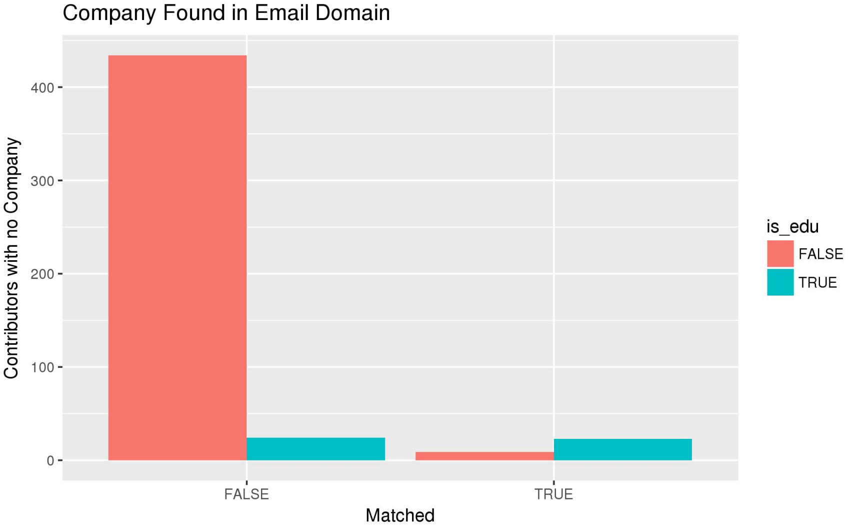
Universities and Corporations showed an equal frequency of domain matches.

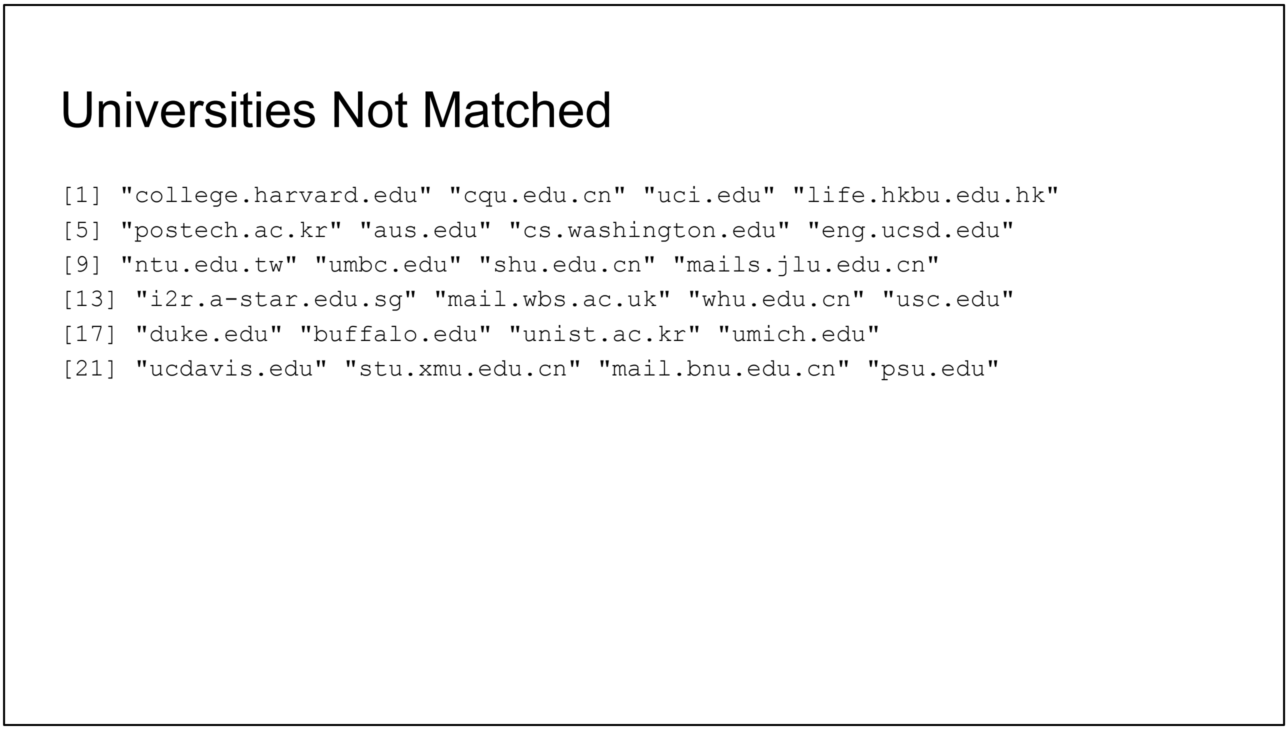
Matched Companies - No Company in Profile



The highest single number of matches came from Amazon, but Amazon is the most represented company in this project, so that may be reflecting that skew. Further analysis is needed on other projects to see if something similar is reflected. In this sample most of the matches actually came from universities so it could be a reasonable way of identifying contributors affiliated with education (either past or present).

Universities - Matched and Not Matched

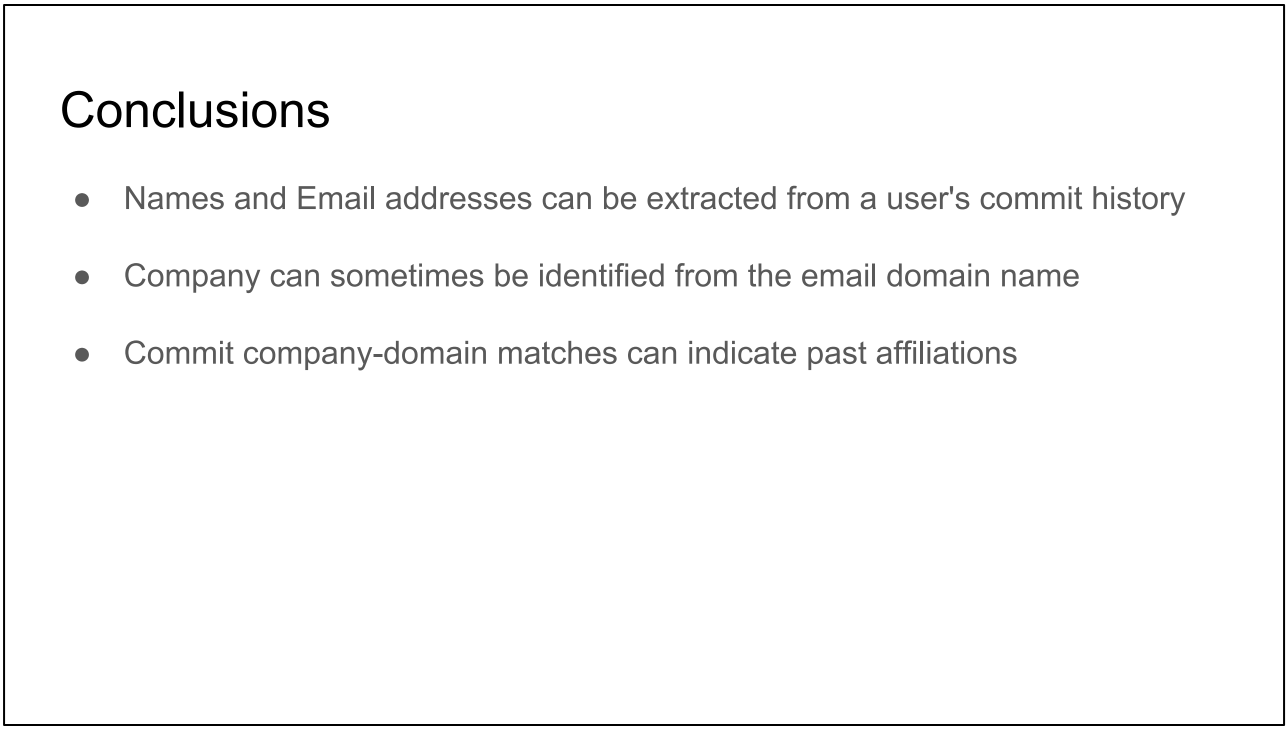




Universities Not Matched

```
[1] "college.harvard.edu" "cqu.edu.cn" "uci.edu" "life.hkbu.edu.hk"
[5] "postech.ac.kr" "aus.edu" "cs.washington.edu" "eng.ucsd.edu"
[9] "ntu.edu.tw" "umbc.edu" "shu.edu.cn" "mails.jlu.edu.cn"
[13] "i2r.a-star.edu.sg" "mail.wbs.ac.uk" "whu.edu.cn" "usc.edu"
[17] "duke.edu" "buffalo.edu" "unist.ac.kr" "umich.edu"
[21] "ucdavis.edu" "stu.xmu.edu.cn" "mail.bnu.edu.cn" "psu.edu"
```

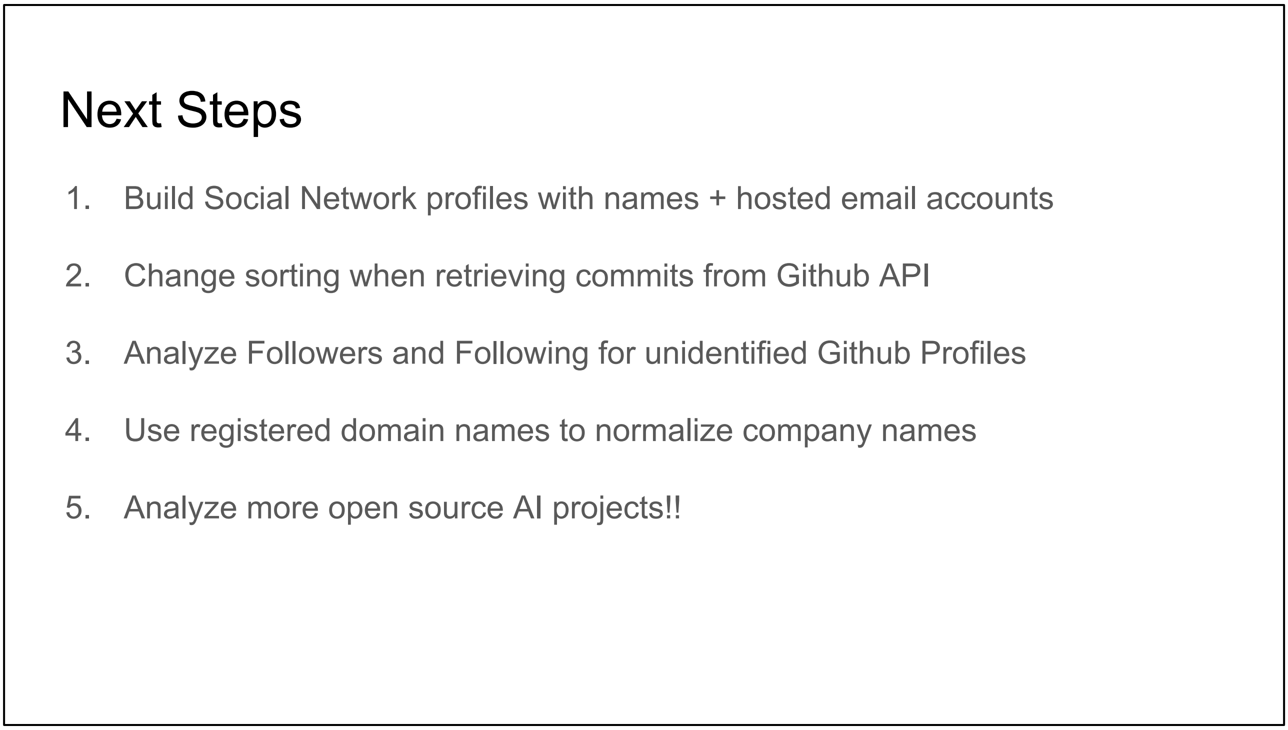
Can we think of a way to match these? Sure can!



Conclusions

- Names and Email addresses can be extracted from a user's commit history
- Company can sometimes be identified from the email domain name
- Commit company-domain matches can indicate past affiliations

We can’t make bolder statements about gmail use and university identification because this is only one project. We will need to evaluate others before drawing larger conclusions.



Next Steps

1. Build Social Network profiles with names + hosted email accounts
2. Change sorting when retrieving commits from Github API
3. Analyze Followers and Following for unidentified Github Profiles
4. Use registered domain names to normalize company names
5. Analyze more open source AI projects!!