university of groningen

faculty of science and engineering

Location inference using GitHub profiles

Andreea-Cristina Zelko

university of groningen

faculty of science and engineering

University of Groningen

Location inference using GitHub profiles

Master's Internship Report

To fulfil the requirements for the Research Internship course
for the Master's degree in Computing Science at the University of Groningen
under the supervision of

Dr. A. Rastogi (Software Engineering, University of Groningen)
and

Andreea-Cristina Zelko (s4311833)

June 24, 2024

Contents

Abstract

One interesting focus within the software analytics research domain is understanding the impact of international collaboration on the productivity of a development team. In order for such research to be carried out, researchers need access to the locations of the team members. Often, this information is only shared by a small number of users, meaning that a lot of users can not be used in research due to insufficient data.
The goal of this research project is to explore methods of obtaining locations which can be associated with the GitHub profile of a user. Being able to infer the location of users using only their profile would enable researchers to expand their candidate pool. This would result in more generalizable and more accurate findings.
The name, bio and email present in the GitHub profile of a person are used to infer locations. The results of the inference show that names can easily be used to obtain a nationality, while the bio and email are harder to use since they are not always provided. It is concluded that inference methods are best used in combination with each other, so that multiple information sources get analysed.

1 Introduction

Software analytics is the practice of collecting, analyzing, and interpreting data related to software development. This includes data about the developers of the software, since insights into the people working on a project can often lead to the most interesting discoveries.

One problem that researchers often run into is the lack of high-quality big datasets that have all the necessary information for a specific study. For example, if one wanted to carry out research about the productivity of international versus national teams, one would first need to know the locations of the developers in the team. For context, an international team is one in which the team members do not all share the same country. So in this case, researchers would first need to collect information about multiple teams for which the location of all the developers is known. However, the issue arises that not a lot of developers actually include their location in their profile. For example, one article published in 2018 found that out of the 15 million users gathered for the research, only 2.3 million had set a location [1]. This represents about 15% of all the users. While the number of GitHub users is ever increasing, this percentage is likely to stay the same. In 2023, GitHub announced they reached the milestone of 100 million users [2].

The goal of this research project is to solve the exact problem posed by the given example. Specifically, the aim is to identify ways in which the location of GitHub users can be inferred, so that future researchers are not limited to around 15% of all users.

When speaking of GitHub developer location, it is important to acknowledge a few things.
First, the location set by a user in their profile is not verified in any way. The field is simply a text input, so it is possible for one to set their location to 'Earth' or 'the Internet'. Such inputs are false positives, because it appears as though the user has set a location, but the information in it is not actually useful to any research.

Second, a person actually has multiple locations that can be associated with them. This is because humans travel often. Many people do not work or study from home for example. If a user includes a location on their profile, it could represent any one of said possible locations. This list of locations includes, but is not limited to:

- Home location
- Work location
- Nationality

Additionally, a difference must be made between perceived location and true location. For example, the possible circadian rhythm of a person might lead to the inference of a perceived location [3]. This means that the person's behaviours match them to a specific time zone. However, it is possible that the person simply works at unusual times, which leads to the result of a perceived location. A true location would be one that positions the person on a map while leaving room for no other alternatives.

When inference methods are used on a user, it is important to keep track of which location is being inferred. For example, a method based on the time of activity of a person can be used to infer their circadian rhythm [3]. This would lead to finding the time-zone of the user and therefore their possible home or work location. By contrast, a method that focuses on the name of the person can only infer the probable nationality of the user.

The structure of this paper is as follows. The next section, Section 2, will present the state of the art with regards to location inference. Section 3 will then give details about how the dataset used for this research was gathered, as well as explore the dataset. Afterwards, Section 4 will dive into the methods and implementations used. Section 5 will present the results which were obtained. Next, Section 6 and Section 7 will cover the discussion of the results and the threats to validity of the research. Lastly, Section 8 will close off this paper by presenting the conclusions.

2 State of the Art

Research on the topic of location inference has been going on for more than a decade, with one of the most cited paper being from 2010 [4]. The main reason behind such studies is the aid they provide for future research. Many other domains can benefit from being able to associate a location with the subject of interest. Such research can range from event tracking [5] to global research [6], where observations are made about how different nationalities or groups behave. However, a big portion of these studies are done on already existing data which might not contain the precise location of the subject, but might very well contain other sources which can give away the location information. So location inference research aims to develop ever improving techniques for inferring precise location information about the subject of interest. This is usually done by using auxiliary information about the user and their network to reach an educated guess. In the following parts of this section, current approaches to location inference, as well as their applications will be discussed.

2.1 Inference on Social Media

It seems as though the focus of research in this domain is on inferring the location of social media users primarily. As stated before, the goal of such research is to aid in the monitoring of important events. For example, researchers have found that it is possible to track the spread of diseases using tweets that mention the ongoing epidemic [7]. Similarly, researchers were able to monitor and alert about earthquakes by observing the number of tweets related to seismic events [5]. In fact, there have been so many different techniques and methods invented in recent years that researchers have started comparing them in surveys [8, 9].

Considering the recent leaps in the domain of Large Language Models (LLMs), researchers have started taking advantage of the strong language processing abilities they present. As such, Staab *et al.* were able to use an LLM to infer a wide range of personal attributes (e.g., location, income, sex) using only the Reddit posts of users [10]. The results they obtained have an impressive 95% accuracy. However, this research is once again focused on the social media posts of the user, which can clearly divulge a surprising amount of information about the writer.

2.2 Location inference on GitHub

The challenge of collecting data from collaboration platforms such as GitHub is the limited amount of publicly available information or the lack of structure in the information that is available. For example, the location field of a GitHub profile is not checked so that the input can be irregular or just false, such as ”the Internet” or ”Mars”. This has led researchers to rely on location inference techniques for their studies.

One such relatively simple approach used by Xia *et al.* involves looking at the times when the user is active on the platform [11]. This allows the observation of the circadian rhythm of the subject. From this, the assumption that work is carried out between 10:00 and 18:00 can be used to obtain the user’s time zone. However, this method still does not find the country or city of the user. By contrast, location inference techniques that use social media, like the ones mentioned in the previous section, can obtain coordinate-level precision.

2.3 Search methodology

Table 1 details the search engines used, as well as the search terms used to obtain some of the research papers discussed above. Other articles were found by following the references listed in the found papers, through backwards snowballing. The search was stopped when a general understanding of currently available approaches and possible use cases was reached.

Table 1: Search engines and search terms.

Search Engine	Search Term
Google Scholar	“LLMs for location inference”
Google Scholar	“Location inference GitHub”
Google Scholar	“Survey of location inference”
RUG SmartCat	“Location inference GitHub”

3 Dataset

This section will elaborate upon the steps taken to obtain the dataset that is used for this research project.

The first step was to obtain a list of repositories from which to collect users. For this task, the ‘On the Shoulders of Giants’ dataset was used [12]. The dataset contained a total of around 11,000 projects of different sizes.

For each project in the dataset, the name of the repository and the name of the owner of the repository were used to place a request to the GitHub REST API. The purpose of these requests was to obtain the list of all the contributors to each respective repository. Each such list contains the GitHub usernames of all the contributors to a repository, as well as each user’s respective number of contributions. A contribution can be anything from pull request to a simple comment on an issue. For each repository, the list of contributors can be quite long. It can contain very active users with thousands of contributions as well as users that have only ever made only one contribution. It is important to note that among the contributors of a project it is also likely and expected to find a few bots as well. These will be filtered out later.

Out of the 11,230 repositories that were requested through the API, 409 returned a ‘404: Not found’. This means that the repositories were either deleted or made private after the ‘On the Shoulders of Giants’ dataset was gathered. This left a total of 10,821 repositories which did return a list of contributors. So of the contributors participated in multiple projects, so there were some duplicates in the initial full list. In total, 355,637 unique contributors were gathered.

At this point in the process, only the usernames of all the contributors have been obtained. In order to gather more data about each user, the GitHub REST API was used again. This time, each request was made with the goal of obtaining the profile information of the respective user. This step was by far the most time consuming due to the rate limit of the API. The rate limit is 5,000 requests per hour, so in order to obtain the profile of each contributor, more than 70 hours were needed.

Once all the data gathering was complete, a dataset containing the profile information of more than 355,000 GitHub users was created. This dataset was then used for the experiments carried out during this project.

3.1 Dataset exploration

In order to give context about the experiments detailed in the rest of this paper, as we as providing insights into GitHub user profiles, a dataset exploration was carried out. The rest of this section will be dedicated to discussing it.

The UpSet plot presented in Figure 1 shows the intersections and relationships between different sets created by filtering the full set of list of user profiles. The filters used were whether or not the user type was that of a bot, whether the profile has a bio, whether the profile had an email, whether

the profile had a location, and whether the profile had a name.

3.2 Bot profiles

Once all the profiles were gathered, it was immediately apparent that certain filters will be of interest. Namely, it was important to filter out bot profiles for all experiments that were carried out. Luckily, as can be observed in Figure 1, none of the bot profiles had a name, bio, email or location. This meant that removing the bots from the list of profiles would not shorten the list of profiles that could be used for the experiments. It should also be noted that out of 355,564 profiles, only 155 were bots. Therefore, 99.9% of the profiles were user profiles that could be used for the experiments.

3.3 Profiles with a name

Since one the experiments will focus on using names to infer the nationality of a user, it is important to know how many users set a name in their profile. As presented in Figure 1, there are 289,659 profiles that contain a name. This represent about 81.5% of all users. This is a fortunately high percentage of users, as it means that the information gained from the nationality inference experiment can be used on many users in the future.

3.4 Profiles with a bio

The second experiment requires the bio of a user profile. Figure 1 shows that few users actually enter a bio, namely only 113,853 profiles had a bio. This represents 32% of all user profiles in the dataset. Looking at the intersection which represents profiles with both a name and a bio, there are 109,172 entries. This subgroup comprises 37.7% of the named profiles, and 95.9% of the profiles with a bio.

3.5 Profiles with an email

The third experiment makes use of the user's email. The dataset contains 116,822 email entries, which can be seen in Figure 1. This represents 32.9% of all users. The percentages are quite similar to those of user bios. There are 112,724 profiles that have both a name and an email. These make up 38.9% of names profiles and 96.5% of profiles with an email. However, there does not seem to be any relation between users that set an email and users that set a bio, because only 45% of profile with an email have a bio as well.

3.6 Profiles with a location

This set of user profiles is not directly related to any of the experiments. However, it does define the reason behind the need for location inference. Out of the 355,564 user profiles in the dataset, only 202,253, about 56.9%, contained a self-reported location. This highlights that any software analytics attempts which do not use location inference would be limited to a much smaller dataset. Still, this number is much higher than the one presented in Section 1. The reason behind this uncertain, but it is likely due to some bias during the dataset gathering phase. For example, maybe the projects taken from the 'On the shoulders of Giants' dataset were biased towards active developers, which might be more likely to include a location in their profile. Additionally, this higher percentage was obtained without filtering for unhelpful locations such as 'Earth' or 'the Internet'. So the percentage of actually informative locations is lower, but harder to obtain.

3.7 Profiles with all attributes

In an ideal situation, location inference would be carried out using all possible approaches for each user. This would ensure that the inferred location is more accurate, since multiple vectors would point in the same direction. However, as Figure 1 shows, only 51,519 profiles contained a name, a bio and an email. This is only 14.5% of all profiles, so combining the methods explored in this paper will only be possible for a small selection of profiles. While it is unfortunate that the combination will not work for most users, it does mean

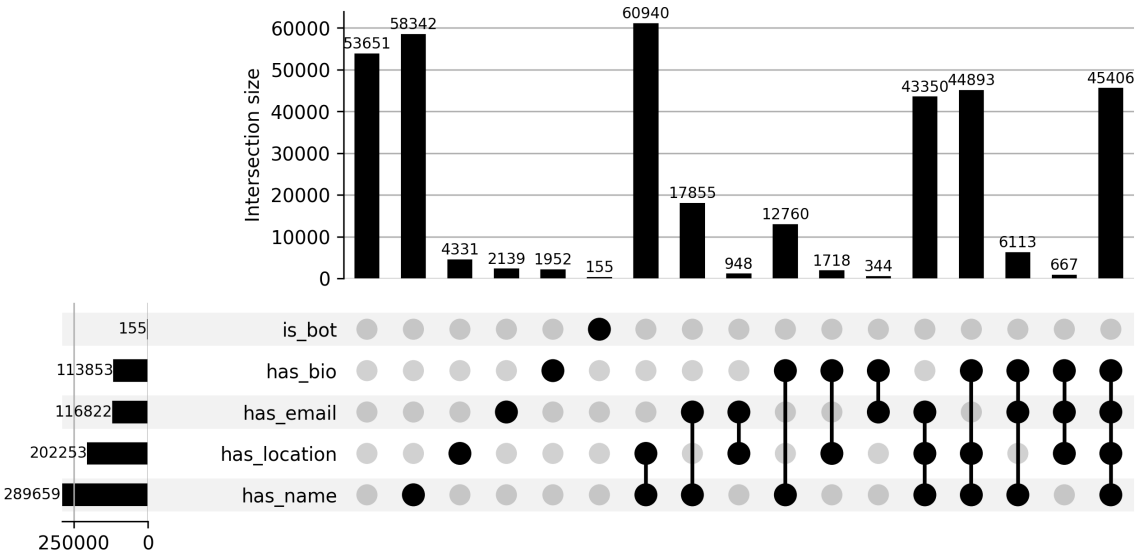

Figure 1: Plot showing the number of entries in each subset, as well as all combinations of intersections between the subsets.

that certain users will have an especially accurate inference.

4 Implementation

The goal of this research project is to determine which parts of the GitHub profile of a user could be used to find a location or set of locations that could be associated with the user. The following parts of a profile were used in the exploration:

- the name of the user
- the bio of the user
- the email of the user

These profile details were chosen because they are easy to extract. It is therefore expected that combining all three inference methods will be easily achieved afterwards.

The following parts of this section will elaborate on how each method was constructed.

4.1 Name

The name of a person can be associated with a location. In most cases, this can represent the nationality of the person. However, it is important to keep in mind that this is not always the case. For example, it is possible to be given a name popular at the time of birth that has not relation to the actual nationality(e.g. an American baby named 'Sasuke' after the name of character in a popular Japanese cartoon). It is also possible that the first and last names of a person indicate different nationalities (e.g. 'Andreea' indicates Romanian while 'Zelko' indicates Hungarian).

Still, for a large majority of people, their name and nationality are correlated. This represent that basis of this exploration. The experiment entails using the name provided by the user and passing this information to the `name2nat` python library. The library will then return the top 3 nationalities associated with the name, as well as the confidence for each guess. Table 2 shows some examples of what the library is capable of.

Table 2 also showcases the limitations of the chosen library. When the first name and last name do not really have a matching nationality, the package returns three guesses that each have quite low confidence. However, when the names agree, the package is able to return quite confident guesses. Additionally, if the last name is placed before the first name, the nationality confidence lower drastically. For this specific package, a confidence of

Table 2: Examples of nationalities inferred from names.

Full name	Inferred nationalities	Confidence
John Doe	American	1.0
Andreea Zelko	German	0.23
	Romanian	0.16
	American	0.13
Andreea Cristina Zelko	Romanian	0.34
	American	0.18
	German	0.11
Matej Kucera	Slovak	0.70
	Czech	0.23
	American	0.02
Kucera Matej	American	0.31
	Argentine	0.10
	Albanian	0.07
Joost de Haan	Dutch	0.98
	Belgian	0.01
	French	0.001

1.0 means that the given name has a Wikipedia page that mentions its nationality. This is because the model used by the package was trained on a Wikipedia dataset. More details about the package can be found on its PyPI page.

4.2 Bio

The Bio of a GitHub user profile represents a simple text field in which the user can write pretty much anything they wish. Most users fill it in with some descriptions of themselves and the things they are passionate about. There are also some users that choose to elaborate on their locations. For example, the Bio of one user might be '*Sydney to Seattle. Did Xbox. Now at Code.org. I like to make things.*'.

The goal of this second experiment is to parse the Bio of a user and search for terms which represent a location (e.g. a city, a country, etc.). Such terms are generally referred to as Geopolitical Entities (GPEs). If any such entities are found, some more questions can be answered:

- What is the percentage of users that include GPEs in their Bio?
- What is the average number of GPEs mentioned in a Bio?
- Do the GPEs in a Bio point to the same location as the self-reported location of the user?

As such, this experiment represents a Natural Language Processing task. Answering the questions above will provide insight into the behaviour of users, and can help guide other future research.

4.3 Email

Most GitHub users also provide an email address which is returned by the API when querying profiles. Naturally, a lot of users employ popular email providers for their own needs. The email structure '*firstname.lastname@gmail.com*' is very common, but does not provide a lot of insight into the location of its user.

However, there are also a few users which purchase their own domain or use a purchased domain. Examples of such emails are '*mail@user.net*' and '*user@apache.org*'. In these cases, the use of a domain different from the popular options (i.e. Gmail, Hotmail, Yahoo, etc) provides an opportunity for location inference.

The goal of this third experiment is to explore the possible insights provided by the email of a user. The following questions should be answered:

- What percentage of users have an email address with a custom domain?
- What location information can be inferred from a custom domain?

5 Results

Since the experiments were designed and separated into three categories, this section will be split up into three subsections. Each subsection will provide details about the process of obtaining the results, as well as discuss the results and their implications and future impact.

5.1 Name

While location inference from only a users name is not directly possible, it is usually possible to find a nationality that can be associated with the given name. On it's own, the nationality of a user does not directly correlate with their current location. However, a nationality can provide insight into the culture and possible traditions of a person. For future software analytics research, knowing the culture and traditions of a person can provide valuable insights.

The Python package used to infer the nationality of a name also provides the confidence of the inference. For the purposes of this experiment, a threshold of 70% confidence was decided. Thus, a confident guess is one that has a confidence score

higher than 70%.

Nationality inference was successfully applied to 289,657 users. Two of the 289,659 profiles contained unusable names. For each user, it was possible to obtain between one to three nationalities. The number of users which received a confident guess was 94,599. This represents about 32.7% of the users for whom it was possible to infer a nationality.

The obtained nationalities were also compared against the self-reported location of each user. This comparison should not be interpreted as a test of accuracy, since it is not adequate to compare two unrelated locations associated with one user. The comparison was simply carried out for exploratory purposes. As per Figure 1 the subset made up of profiles which have both a name and a location contains 194,589 entries. Out of these, it was found that for 107,403 of users, the inferred nationality and the self-reported location pointed to the same country. This amounts to 55.2%.

5.2 Bio

Inferring location from a user's bio proved to be rather trickier. The difficulty lies in the fact that not many users set a bio in the first place. As discussed in subsection 3.1, only 32% of profiles contain a bio.

While parsing the available bios, it was discovered that only 3,480 of them contained geopolitical entities. This represents only 3.0% of profiles with a bio, and only 1.0% of all profiles.

However, for this experiment, it is interesting to look at how the bio location compares with the self-reported location. First, it only makes sense to look at the subgroup of profiles for which we were able to infer a location from the bio and which also have a self-reported location. This subgroup contains 2981 profiles. Out of this subgroup, a total of 2620 profiles had a matching bio location and self-reported location. This amounts to 87.9%. Interestingly, there were also 499 profiles for which location inference from bio was successful even though the profile did not contain a self-reported location. This highlight that some users do not mention a location directly in their profile, but are willing to write a location in their bio.

5.3 Email

This third experiment revolves around obtaining a location from the email of a user. Initially, it was intended to use the country code top-level domains (ccTLDs) present in email (e.g. 'student@rug.nl') and extract the country from that. However, this approach would leave out all generic top-level domains (gTLDs) such as '.com', '.org' or '.dev'. So a new approach was devised. This new approach used the whole domain extracted from the email address and passed that to a WhoIs service. This service would then return information about the person who registered the domain. This information would often include a country which could be used as the inference result.

Using this method, 30,771 countries were successfully inferred from emails. This amounts to 26.3% of profiles that contain an email and 8.6% of all available profiles. However, a lot of the emails used by users are hosted by generic email services like Gmail, Yahoo, Hotmail and Outlook. Inferring the country of these services does not provide much informational about the user. Out of all the profiles with an email, 50,035, which is 44.4%, did not belong to a popular mailing service. This minority group of emails is the one that provides actual insight into the location of the user.

5.3.1 Personal vs Company Email

There is a distinction which can be made between personal and company emails. While both kinds of emails might not use a generic mailing domain, the country inferred from a personal email will be more closely tied to the user compared to the country inferred from their company email. Consider, for example, a developer with the email domain

'python.org'. This is a custom domain, but does not offer a lot of insights into the location of the developers, because there is no location strongly associated with the Python programming language.

Unfortunately, distinguishing between personal and company emails is quite tricky. While it is easy for a human to tell that 'andreeazelko.dev' is likely a personal domain while 'apache.org' is probably a company domain, it is very difficult to set and implement a clear distinction. Therefore, a creative solution was found. For the sake of this experiment, if the second-level domain on a URL had a page on Wikipedia, then it was considered that the domain belongs to a company. Inversely, if the entity did not have a page on Wikipedia, then it likely represented a personal domain. There are of course still some gray areas. For example, it is possible that a company is so small that it does not have a Wikipedia page yet. In this case, it is considered that the small scale of the company does not detract from the accuracy of the inferred country.

There is one more approach which could be considered. Once a location is strongly associated with a certain domain, it becomes possible to assign that location to all users of the domain. An example of this would be the domain of a university (e.g. 'rug.nl'). A university is located at a point on a map, and all users with an email domain belonging to the university can be associated with that location.

Using the above described approach, it was found that 25,551 of the emails, 21.9%, used personal domains. Out of these, it was possible to infer a country for 16,713, or 65.4%.

5.4 Combination of methods

The reason behind exploring multiple inference methods is the possibility of combining them. The goal here is to see if the results of the methods explored in this paper point in the same direction. It is also interesting to explore if implementing multiple inference approaches can benefit researchers.

It was found that combining all three approaches and counting the number of users for which it was possible to infer location through any means results in 290700 users with an inferred location. This means that location can be successfully inferred for 97.7% of users.

Additionally, there were 428 users for which it was possible to infer location using all three methods. For this subset, it was found that all three approaches pointed to the same location for 139 users.

6 Discussion

Looking at the results that were obtained, there are a few observations that should be stated.

First, while nationality can not be directly correlated to the actual location of a person, it can still offer important insights into the culture of the user. Considering this fact, it is fortunate that nationality inference could be applied to the majority of users. The inference method used was also not very computationally intensive, so it could be easy implemented alongside other inference methods.

Second, emails could be used for location inference for only a small subgroup of the users. However, the location inferred from the email might be the most accurate and detailed of all the methods tried during this project. For some users, it was even possible to obtain a city, or full address from the WhoIs service used. This was only possible for users which provided that information when they registered the domain. On top of this, knowing whether the person is associated with a relatively big company can also provide useful insights.

Overall, all of the methods tried could be used together to obtain a list of countries which should be associated with a single user. It is important to keep in mind that few people only belong to one location. Taking an international student as an example, they can be associated with both their

home country and their study country. The student does not lose their association with their home country the moment they move out, but they also can not completely ignore the influence of their new location. Therefore, it is reasonable to construct a list of locations for each users, instead of restricting future research to only one location.

As part of this discussion it is also important to mention the ethics involved in trying to infer the location of a user who did not provide it directly. As stated by all other research papers with similar goals, the aim is to provide researchers with more tools. This future research is aimed at towards societal benefits similar to goals like predicting earthquakes or pandemics. Additionally, while the users did not provide a publicly available location, all the information used to infer the location was publicly available. Therefore, it can be stated that no privacy was breached using the methods explored.

7 Threats to Validity

This research could suffer from multiple threats to validity due to its reliability on external packages.

First, the fact that the package used to infer nationality based on the name of a user uses a model trained on Wikipedia data can introduce some bias. Specifically, for people that come from English speaking countries might get grouped as American. To give an example, as can be seen in Table 2, a British person named *John Doe* would be misidentified as American. Additionally, it is also simply possible that the name of a person does not match with their actual nationality. This fact has to be simply accepted by the users of this method, as it can not simply be remedied.

Second, it is also possible that the WhoIs service provided by the `whois` python package fails to provide the most recent or most accurate information about the given domain. One improvement that could be considered in order to try to mitigate this issue would be employing multiple WhoIs services. The results returned by these could then be compared and compiled into a more accurate location.

8 Conclusion

The goal of this research project was to identify a few ways in which the GitHub profile of a user could be used to infer a list of locations. The locations in the list can represents different aspects of a person, so it is expected that they do not all point to the same location.

This paper covers the implementation, results and discussion for three inference methods. The name, bio and email of a user were used to extract location information from a profile. The results show that the nationality inferred from a name can be successfully applied to the majority of users, while the other two methods can only be applied to a smaller subset of profiles.

The conclusion of this research is that multiple methods can be combined in order to extract as much information as possible from a GitHub profile. The methods presented in this paper do not contradict each other and would work well in combination with each other.

Bibliography

[1] B. Frederickson, "Where do the world's software developers live?," 2018. Accessed: 2024-05-18.

[2] T. Dohmke, "100 million developers and counting," 2023. Accessed: 2024-05-18.

[3] J. M. Gonzalez-Barahona, G. Robles, and D. Izquierdo-Cortazar, "Determining the geographical distribution of a community by means of a time-zone analysis," in *Proceedings of the 12th International Symposium on Open Collaboration*, OpenSym '16, (New York, NY, USA), Association for Computing Machinery, 2016.

[4] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), p. 61–70, Association for Computing Machinery, 2010.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), p. 851–860, Association for Computing Machinery, 2010.

[6] J. Wachs, M. Nitecki, W. Schueller, and A. Polleres, "The geography of open source software: Evidence from github," *Technological Forecasting and Social Change*, vol. 176, p. 121478, Mar. 2022.

[7] J. Gomide, A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira, "Dengue surveillance based on a computational model of spatio-temporal locality of twitter," in *Proceedings of the 3rd International Web Science Conference*, WebSci '11, ACM, June 2011.

[8] O. Ajao, J. Hong, and W. Liu, "A survey of location inference techniques on twitter," *Journal of Information Science*, vol. 41, p. 855–864, Nov. 2015.

[9] R. Lamsal, A. Harwood, and M. R. Read, "Addressing the location a/b problem on twitter: the next generation location inference research," in *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising*, LocalRec '22, (New York, NY, USA), Association for Computing Machinery, 2022.

[10] R. Staab, M. Vero, M. Balunović, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," 2023.

[11] X. Xia, Z. Weng, W. Wang, and S. Zhao, "Exploring activity and contributors on github: Who, what, when, and where," in *2022 29th Asia-Pacific Software Engineering Conference (APSEC)*, IEEE, Dec. 2022.

[12] X. Zhang, A. Rastogi, and Y. Yu, "On the shoulders of giants: A new dataset for pull-based development research," in *Proceedings of the 17th International Conference on Mining Software Repositories*, MSR '20, (New York, NY, USA), p. 543–547, Association for Computing Machinery, 2020.