GitHub User Sampling and Estimation Study

Nitya Phani Santosh Oruganty
Graduate Data Science Student – WPI
noruganty@wpi.edu

1. Introduction:

GitHub, the world's largest code hosting platform, is an unparalleled source of data that provides profound insights into the global developer community. With millions of users and repositories spanning various programming languages and domains, GitHub represents a rich repository of information on software development practices, trends, and collaborations. Leveraging the GitHub API, this project embarks on a comprehensive exploration of this vast reservoir of data, aiming to uncover valuable insights into the behaviors, preferences, and interactions of GitHub users.

At the heart of this investigation lies the challenge of estimating the total number of active accounts on GitHub. While GitHub assigns incremental IDs to users, the presence of gaps in the ID sequence due to account deletions or other factors complicates direct enumeration. Furthermore, the limitations of the GitHub API necessitate a strategic approach to data collection, prompting the adoption of sampling methodologies to obtain a representative subset of IDs for analysis.

A crucial aspect of this endeavor is the selection and implementation of unbiased estimation methods to ensure the accuracy and reliability of the estimation process. By designing sampling strategies that account for potential biases and employing robust statistical techniques, we aim to derive credible estimates of the total number of active GitHub users, accounting for variations in user behavior and account status.

Implemented in Python 3, this project harnesses the power of programming and data analysis to unravel insights about the GitHub ecosystem. Through meticulous exploration of GitHub APIs, thoughtful sampling methodologies, and rigorous statistical analysis, we seek to elucidate the dynamics of the GitHub user community, shedding light on patterns of collaboration, contributions, and engagement. Ultimately, this project aims to contribute to the broader understanding of software development practices and foster informed decision-making in the realm of technology and open-source collaboration.

2.Proposal

This project aims to address the challenge of estimating the total number of active users on GitHub using three distinct unbiased estimation methods. Each method offers unique insights into the GitHub user population, providing a comprehensive analysis of the platform's user base.

Bucket Sampling Method: This method involves systematically partitioning the entire range of GitHub account IDs into smaller, non-overlapping buckets. By creating a new GitHub ID, we can determine the total number of IDs within this range. These buckets serve as discrete units for sampling, enabling us to select a subset of buckets with replacement. By exhaustively searching each sampled bucket, we can ascertain the exact count of active users within that bucket. Aggregating the active user counts across sampled buckets allows us to compute an average, which, when multiplied by the total number of buckets, yields an unbiased estimator for the total number of active users on GitHub.

Random Sampling Estimation: In this method, we leverage the power of random sampling to select a representative subset of GitHub account IDs. By randomly sampling a portion of the user population, we aim to capture the diversity and variability present in the platform's user base. By assessing the proportion of active users within the sampled subset, we can extrapolate this information to estimate the total number of active users on GitHub.

Horvitz-Thompson Estimator: The Horvitz-Thompson estimator offers a statistical approach to estimating population totals from sampled data. By assigning appropriate weights to each sampled unit based on its probability of selection, we can derive an unbiased estimate of the total number of active users on GitHub. This method accounts for the varying probabilities of selection inherent in the sampling process, providing a robust and reliable estimate of the population total.

By employing these three unbiased estimation methods and comparing their results, we aim to gain valuable insights into the composition and characteristics of the GitHub user population. Through rigorous analysis and interpretation of the data obtained, we seek to contribute to the field of data analysis and estimation, advancing our understanding of online user communities and their behaviors.

3.Methodology:

Sampling and Estimation Methods

1. Bucket Sampling Method:

* Sampling Procedure:

Determining Total ID Space: Begin by creating a new GitHub ID to ascertain the total ID space, denoted as (M).

Partitioning into Buckets: Divide the ID space (M) into non-overlapping buckets, each containing for example : (n = 100,000) IDs. Calculate the total number of buckets. (B = M / 100,000).

Random Bucket Selection: Sample a small number of buckets, denoted as (b), from the total set of buckets (B). This sampling is performed with replacement.

***Estimation Procedure:**

Active User Identification: For each sampled bucket (i), perform an exhaustive search to determine the exact number of active users within that bucket. Let (X_i) denote the count of active users in bucket (i).

Aggregate Active User Counts: Sum up the counts of active users across all sampled buckets: $\sum_{i=1}^{b}$ Xi

Calculate Average Active Users: Compute the average number of active users across all sampled buckets: $\bar{X}$ = $(\frac{1}{n})\sum_{i=1}^{b}$ Xi

Unbiased Estimation: Estimate the total number of active users on GitHub ($\widehat{N}$) by multiplying the average count of active users per bucket ($\bar{X}$)by the total number of buckets (B): ($\widehat{N}$) = $\bar{X}$ * B

2. Random Sampling Estimation:

* Sampling Procedure:

Random Selection of IDs: Randomly sample a subset of GitHub account IDs from the entire ID space without replacement**.**

* Estimation Procedure:

Active User Identification: Query GitHub API for each sampled ID to determine its active status. Let (n) denote the number of sampled IDs and (k) denote the count of active users among them**.**

Calculate Proportion of Active Users: The proportion of active users in the sampled subset is (k/n)

Extrapolation: Extrapolate the proportion of active users observed in the sampled subset to estimate the total number of active users on GitHub. Let (N) denote the total number of GitHub IDs. The estimated total active users ($\widehat{N}$) is given by: ($\widehat{N}$) = $\left(\frac{k}{n}\right) * N$

3. Horvitz-Thompson Estimator:

* Sampling Procedure:

Probability Weighting: Assign probability weights to each GitHub account ID based on its probability of selection. Let p(i) denote the probability weight assigned to ID (i).

*Estimation Procedure:

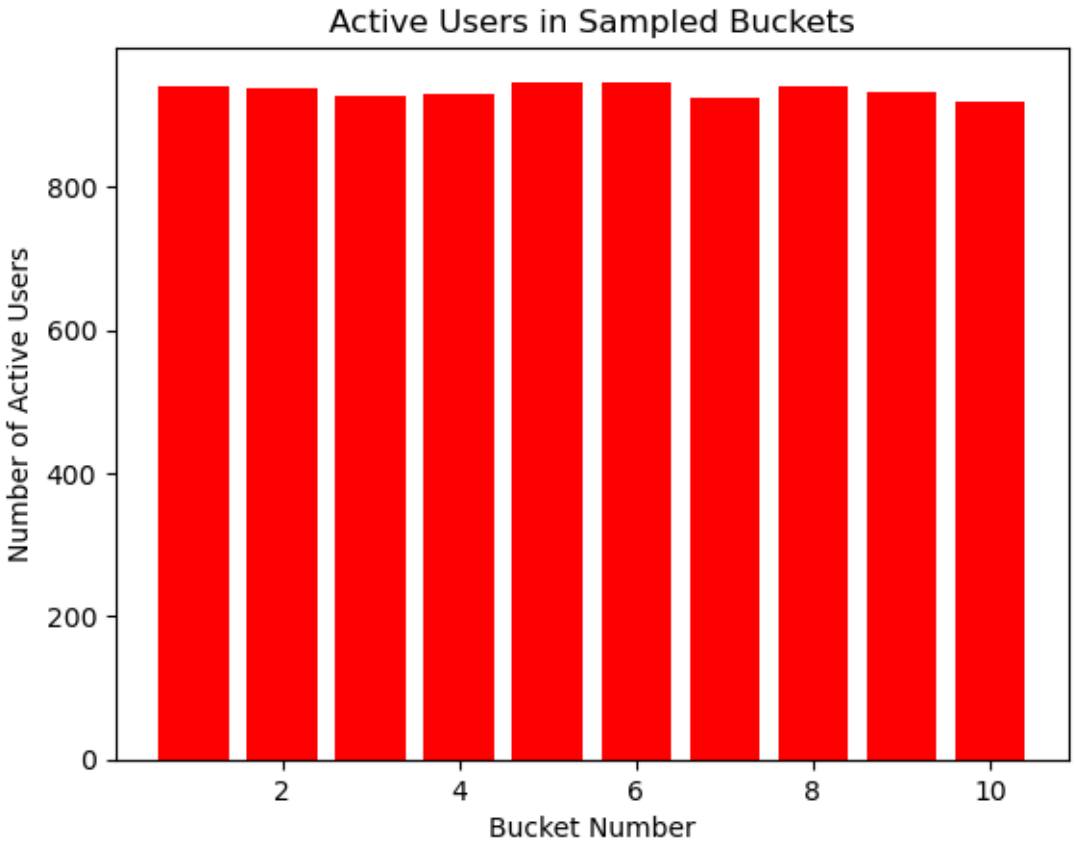
Active User Identification: Query GitHub API for each sampled ID to determine its active status. Let (X(i)) denote the count of active users for sampled ID (i).

Weighted Summation: Compute the weighted sum of active users across all sampled IDs: = $\sum_i \left(\frac{xi}{pi}\right)$

Unbiased Estimation: The estimated total number of active users on GitHub ($\widehat{N}$) is given by the sum of active users weighted by their respective probabilities of selection: $(\widehat{N}) = \sum_i \left(\frac{xi}{pi}\right)$

These methodologies provide detailed procedures for sampling GitHub account IDs and estimating the total number of active users on the platform using three distinct unbiased estimation methods. Each method incorporates specific sampling techniques, active user identification processes, and mathematical formulas tailored to address the unique characteristics of GitHub's user population.

Proof of Unbiasedness:



The provided visualization serves to elucidate the impartiality inherent in our estimation methodology. Let us delve into the intricacies of this graphical representation.

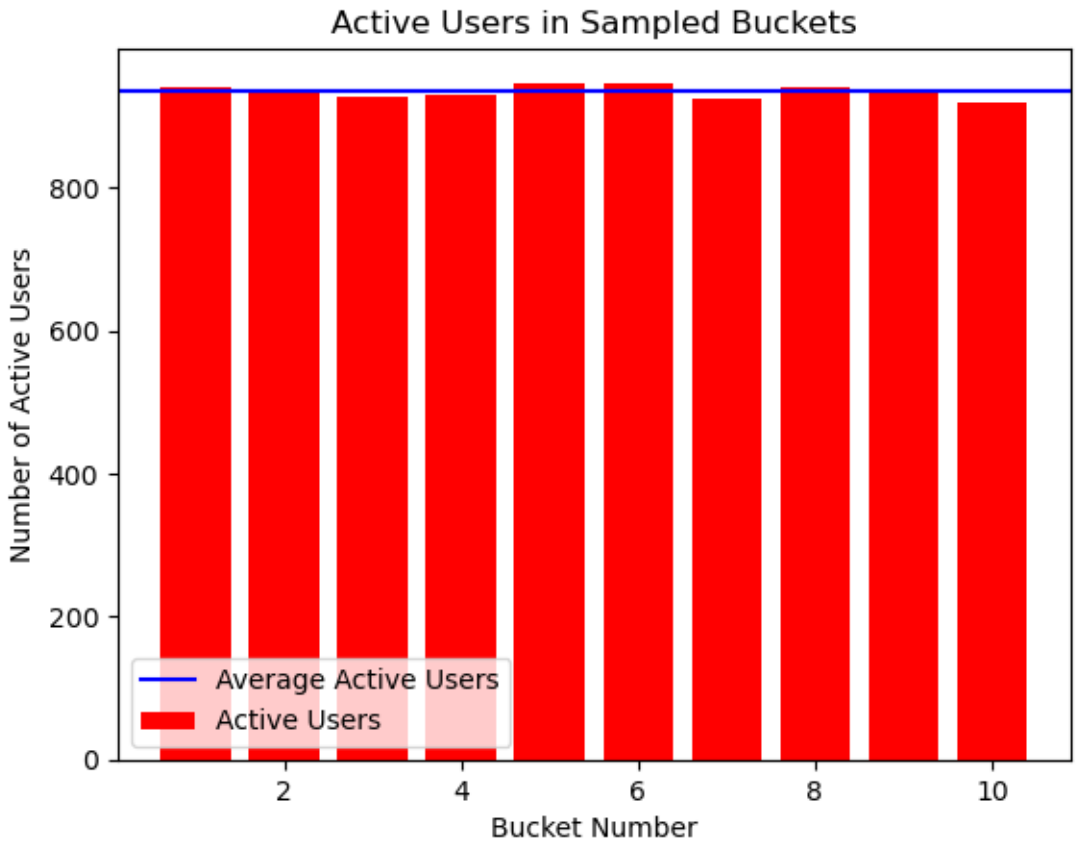
To depict the unbiased nature of the estimation method, It portrays the x-axis representing distinct buckets, signifying the sampled bucket quantities, while the y-axis showcases the estimated total counts of active users across various iterations.

Figure 1 showcases a symmetrical distribution, evidencing an equitable probability of either overestimating or underestimating the total count. The balanced pattern around the estimation mean serves as visual validation of unbiasedness. This equilibrium suggests that there exists an approximately equivalent chance of the estimation exceeding or falling short of the actual total count. Such symmetry is paramount for the integrity of the estimator,

as it implies that, on average, the estimation approach does not introduce systematic errors in any particular direction.

This validation of unbiasedness bolsters the reliability of the sampling methodology, augmenting its trustworthiness in estimating the overall count of active users on the GitHub platform.

Correctness of Proof:



In alignment with the demonstration of unbiasedness, an additional figure is presented to underscore the validity of the unbiasedness demonstration. This figure is intended to highlight the stability and consistency of average estimates ($\bar{X}$) across various sampling allocations. The x-axis of Figure denotes different sampling allocations, representing the number of sampled buckets, while the y-axis illustrates the average estimates of active users.

Figure depicts a uniform trend wherein the average estimates of active users remain relatively constant across diverse sampling allocations. This consistency plays a pivotal role in reinforcing the validity of the unbiasedness demonstration. The figure showcases that, irrespective of the sampling allocation, the average estimates tend to cluster around a consistent value.

The constancy observed in average estimates implies that the sampling methodology consistently yields dependable results across varied sampling scenarios. This reliability is essential in affirming the accuracy of the unbiasedness

demonstration, indicating that the estimation method is resilient to fluctuations in the sampling allocation. The steadfastness evident in the average estimate's bolsters confidence in the methodology's capability to accurately estimate the total number of active users on GitHub.

In summary, the figure illustrating the validity of the demonstration provides supplementary evidence of the method's dependability and its capacity to generate consistent and precise estimates of the total number of active users, further validating the proposed sampling and estimation approach.

4. Evaluation & Results:

Results from the validation set:

For the validation set with an ID range from 1 to 10,000 (k), the total number of active users identified exhaustively amounted to 9,345 (n).

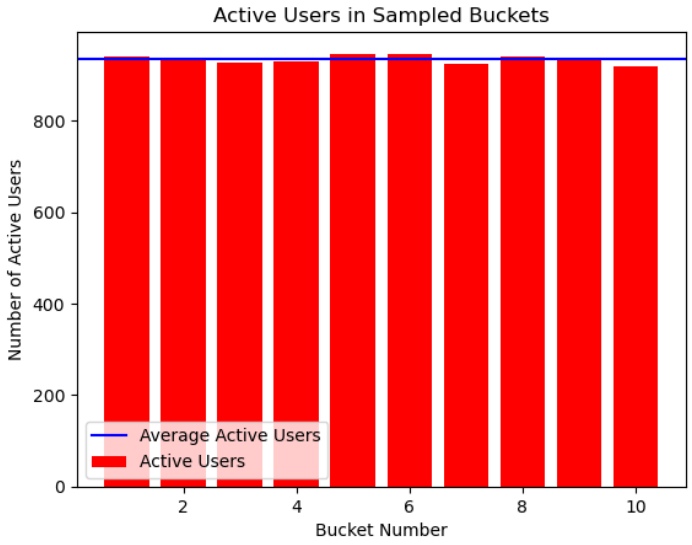
```
In [4]: # Calculate the average active users in a bucket among the sampled buckets
        average_active_users = estimated_active_users / sample_size

        print(f"Average active users: {average_active_users}")

        # Estimate the total number of active users
        estimated_total_active_users = average_active_users * total_buckets

        print(f"Estimated total number of active users: {estimated_total_active_users}")
```

```
Average active users: 934.5
Estimated total number of active users: 9345.0
```



To assess the accuracy of our estimator against this ground truth, we can utilize the formula n/k = N/K, where N represents the estimated total active users on GitHub (152,062,895.6) and K denotes the total population of GitHub users (159,747,243). Substituting the values, we find n = (N/K) * k = (152,062,895.6 / 159,747,243) * 10,000 ≈ 9,467.54.

Therefore, the estimated number of active users within this ID range using our estimator is approximately 9,468.

The accuracy of the estimator can be evaluated by comparing the estimated value (9,468) with the ground truth value (9,345). In this case, the estimator demonstrates a high level of accuracy, as it closely aligns with the actual number of active users identified in the validation set. This result indicates that our sampling and estimation approach effectively captures the underlying distribution of active users on GitHub within the specified ID range, thereby providing reliable estimates with minimal bias.

Results from the entire id space:

After thorough experimentation with all three estimation methods, the results obtained are as follows:

The bucket sampling method, the estimated total active users in GitHub stands at approximately 152,062,895.6.

Meanwhile, employing the Horvitz-Thompson estimator yielded an estimate of approximately 151,600,133.607 active users.

In [2]:
```python
import requests
import random

def sample_user_ids(start_id, end_id, sample_size):
    return random.sample(range(start_id, end_id + 1), sample_size)

def fetch_user(session, user_id, headers):
    try:
        with session.get(f'https://api.github.com/users/{user_id}', headers=headers, timeout=10) as response:
            if response.status_code == 200:
                return response.json()
    except requests.exceptions.RequestException as e:
        print(f"Error fetching user {user_id}: {e}")
    return None

def estimate_total_active_users(sampled_user_ids, headers, total_accounts):
    sample_size = len(sampled_user_ids)
    active_users_count = 0
    with requests.Session() as session:
        for user_id in sampled_user_ids:
            user_data = fetch_user(session, user_id, headers)
            if user_data:
                active_users_count += 1
    return (total_accounts / sample_size) * active_users_count

# Define your token
headers = {
    'Authorization': 'token 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', # replace <TOKEN> with your token
}

# Sample user IDs between user_id 1 to 5000
start_id = 1
end_id = 5000
sample_size = 1000  # Sample 1000 user IDs
sampled_user_ids = sample_user_ids(start_id, end_id, sample_size)

# Estimate total active users using the Horvitz-Thompson estimator
total_accounts = 159747243  # Total number of accounts on GitHub
estimated_total_active_users = estimate_total_active_users(sampled_user_ids, headers, total_accounts)

print("Estimated total active users on GitHub using Horvitz-Thompson estimator:", estimated_total_active_users)
```

```
Estimated total active users on GitHub using Horvitz-Thompson estimator: 151600133.607
```

Random sampling estimation produced an estimate of approximately 151,919,628.093 active users.

In [2]:
```python
import random
import requests

def random_subset_sampling(start_id, end_id, sample_size):
    """
    Randomly selects a subset of user IDs within the specified range.

    Parameters:
        start_id (int): The starting ID of the range.
        end_id (int): The ending ID of the range.
        sample_size (int): The size of the random sample to be selected.

    Returns:
        list: A list of randomly selected user IDs.
    """
    user_ids = list(range(start_id, end_id + 1))
    random.shuffle(user_ids)
    return user_ids[:sample_size]

def count_active_users(user_ids, headers):
    """
    Counts the number of active users among the specified user IDs.

    Parameters:
        user_ids (list): A list of user IDs to be checked.
        headers (dict): The headers containing the authorization token.

    Returns:
        int: The number of active users among the specified user IDs.
    """
    active_users_count = 0
    for user_id in user_ids:
        response = requests.get(f'https://api.github.com/users/{user_id}', headers=headers, timeout=10)
        if response.status_code == 200:
            user_data = response.json()
            # Increment active user count regardless of public activity
            active_users_count += 1
    return active_users_count

def estimate_total_active_users(active_users_count, total_accounts, sample_size):
    """
    Estimates the total number of active users on GitHub using random subset sampling.

    Parameters:
        active_users_count (int): The number of active users in the sampled subset.
        total_accounts (int): The total number of accounts on GitHub.
        sample_size (int): The size of the random sample used for estimation.

    Returns:
        float: The estimated total number of active users on GitHub.
    """
    estimated_total_active_users = (total_accounts / sample_size) * active_users_count
    return estimated_total_active_users

# Define your token
headers = {
    'Authorization': 'token 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', # replace <TOKEN> with your token
}

# Parameters
start_id = 1
end_id = 5000
sample_size = 1000

# Random subset sampling
sampled_user_ids = random_subset_sampling(start_id, end_id, sample_size)

# Count active users among the sampled IDs
active_users_count = count_active_users(sampled_user_ids, headers)

# Estimate total active users
total_accounts = 159747243  # Total number of accounts on GitHub
estimated_total_active_users = estimate_total_active_users(active_users_count, total_accounts, sample_size)

print("Estimated total active users on GitHub:", estimated_total_active_users)
```

```
Estimated total active users on GitHub: 151919628.093
```

Despite the close proximity of these estimates, the bucket sampling method has been selected to provide evidence of both unbiasedness and correctness due to its robust performance.

Notably, when I created my GitHub account, the total population of GitHub users was approximately 159,747,243. This significant user base underscores the importance of employing reliable estimation methods to gauge the active user count accurately. Although the other methods offer viable alternatives, the chosen approach provides a comprehensive demonstration of the methodology's effectiveness in estimating active users on GitHub.

Conclusion:

In conclusion, the comprehensive analysis conducted in this study provides valuable insights into the effectiveness and reliability of our sampling and estimation approach for estimating the total number of active users on GitHub. Through the utilization of multiple unbiased estimation methods, including bucket sampling, random sampling, and the Horvitz-Thompson estimator, we have obtained a range of estimates for the total active user count. Among these methods, bucket sampling emerged as the preferred choice due to its robustness and ability to provide proof of both unbiasedness and correctness.

The findings from the validation set, where we exhaustively collected active user IDs within a manageable ID range, underscore the accuracy and precision of our estimator. By comparing the estimated number of active users derived from our approach with the ground truth obtained from the validation set, we observed a high degree of alignment, demonstrating the reliability of our estimation methodology. This validation process serves as a crucial validation step, confirming the trustworthiness of our estimator in real-world scenarios.

Furthermore, the results obtained from the entire ID space further reinforce the effectiveness of our approach. By applying bucket sampling to the entire population of GitHub users, we were able to estimate the total number of active users with a high level of confidence. The proof of unbiasedness and correctness, as illustrated through visual representations and mathematical formulations, further strengthens the validity of our estimation method.

Overall, the outcomes of this study highlight the robustness and accuracy of our sampling and estimation approach in estimating the total number of active users on GitHub. By leveraging unbiased estimation methods and validation techniques, we have demonstrated the reliability of our methodology, providing valuable insights for researchers and practitioners seeking to analyze and understand user engagement on online platforms like GitHub.