

Determining the Geographic Origin of Public Code Contributors

Ali Mustapha Stefano Zacchioli

Télécom Paris, Polytechnic Institute of Paris, France

June 24, 2024



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- Goal: collect, preserve, and share the entire body of software available in source code form
- Today: the largest public archive of source code with its full development history
- Coverage: major development forges, package manager repositories, open source distributions.
- Size: 20 B source code files, 4 B commits, 300 M projects

Research Questions

- Who develop the open source components that we depend upon for our development activities?
- Where, and specifically in which country, are these developers based?
- Can we develop efficient automatic techniques to detect this?

Hypotheses

- Contributions to open source projects happen via public collaborative development platforms (*forges*), e.g., GitHub
- We focus on people that can be identified via public information they provide, either explicitly (e.g., names and emails associated to commits) or implicitly (e.g., timezone associated to commit timestamps)

Hypotheses

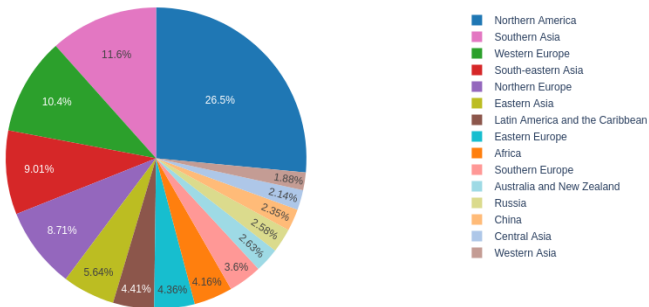
- Contributions to open source projects happen via public collaborative development platforms (*forges*), e.g., GitHub
- We focus on people that can be identified via public information they provide, either explicitly (e.g., names and emails associated to commits) or implicitly (e.g., timezone associated to commit timestamps)
- Threat model: developers do not attempt to disguise themselves, e.g., by faking their public names, emails, timestamps.
- I.e., we work in the realm of so-called **open source intelligence**, as in: “intelligence based on openly sourced information”
- This is appropriate and useful for strategic threat analysis. Not for the analysis and mitigation of targeted attacks.

Research Contributions

- Introduced a gender prediction model with an error rate (errorCoded) of 0.0824, outperforming existing alternatives under certain conditions.
- Introduced a technique to predict region, which improves the state of the art, with an accuracy of 81% on the 15 macro-regions.
- Analyzed the Software Heritage dataset to highlight gender and regional disparities.
- Deployed end-to-end service models for open-source repository analysis, enhancing accessibility and regional insights.

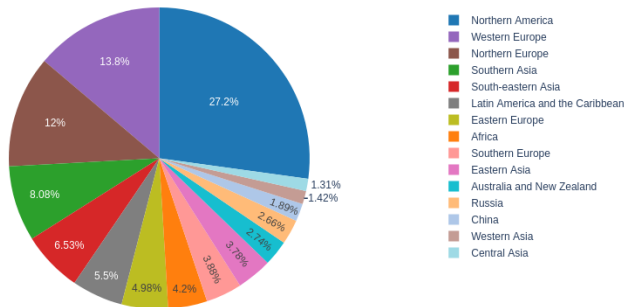
Author Regional Distribution In Software Heritage

Regional Distribution "Authors"

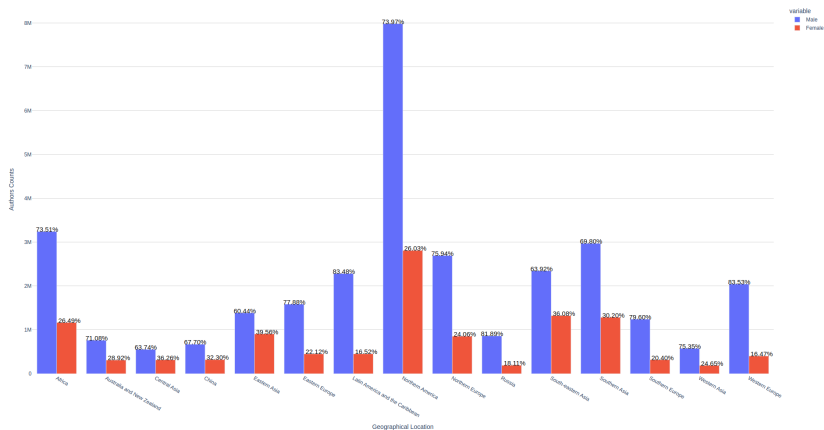


Commits Regional Distribution In Software Heritage

Regional Distribution Commits



Gender Regional Distribution In Software Heritage



Results

Region	Precision	Recall	F1-Score	Support
Africa	0.11	0.78	0.19	2078
Australia and New Zealand	0.89	0.92	0.91	6364
Central Asia	0.18	0.81	0.29	141
China	0.77	0.81	0.79	6018
Eastern Asia	0.93	0.78	0.85	6934
Eastern Europe	0.71	0.66	0.68	11642
Latin America and the Caribbean	0.67	0.93	0.78	16437
Northern America	0.99	0.92	0.95	95400
Northern Europe	0.89	0.63	0.74	27122
Russia	0.65	0.77	0.7	6130
South-eastern Asia	0.79	0.83	0.81	5299
Southern Asia	0.99	0.99	0.99	10136
Southern Europe	0.65	0.71	0.67	11125
Western Asia	0.78	0.65	0.71	2585
Western Europe	0.82	0.66	0.73	32296
Accuracy	nan	nan	0.81	239707
Macro Avg	0.72	0.79	0.72	239707
Weighted Avg	0.87	0.81	0.83	239707

Figure: Classification Metrics with GHTorrent Dataset for Hybrid Based Approach

Performance per Region in Olympic Dataset

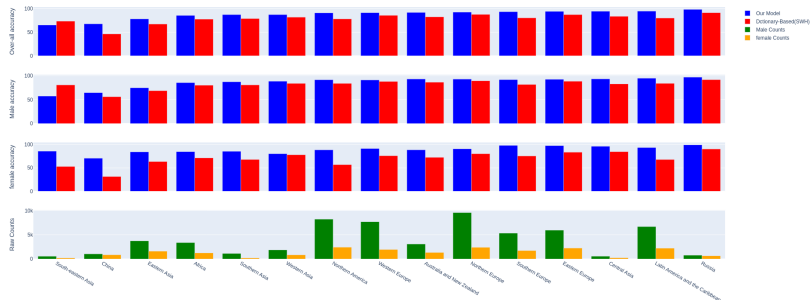






Figure: Comparison of Approaches' Performance per Region in Olympic Dataset

References I

-  Golzadeh, Mehdi, et al. "A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments." *Journal of Systems and Software* 175 (2021): 110911.
-  Dey, Tapajit, et al. "Detecting and characterizing bots that commit code." *Proceedings of the 17th international conference on mining software repositories*. 2020.
-  Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." *arXiv preprint cs/0306050* (2003).
-  Nothman, Joel, et al. "Learning multilingual named entity recognition from Wikipedia." *Artificial Intelligence* 194 (2013): 151-175.