Geographic Diversity in Public Code Contributions

An Exploratory Large-Scale Study Over 50 Years

Davide Rossi - University of Bologna, Italy
Stefano Zacchiroli - Télécom Paris, Polytechnic Institute of Paris, France

Contribution

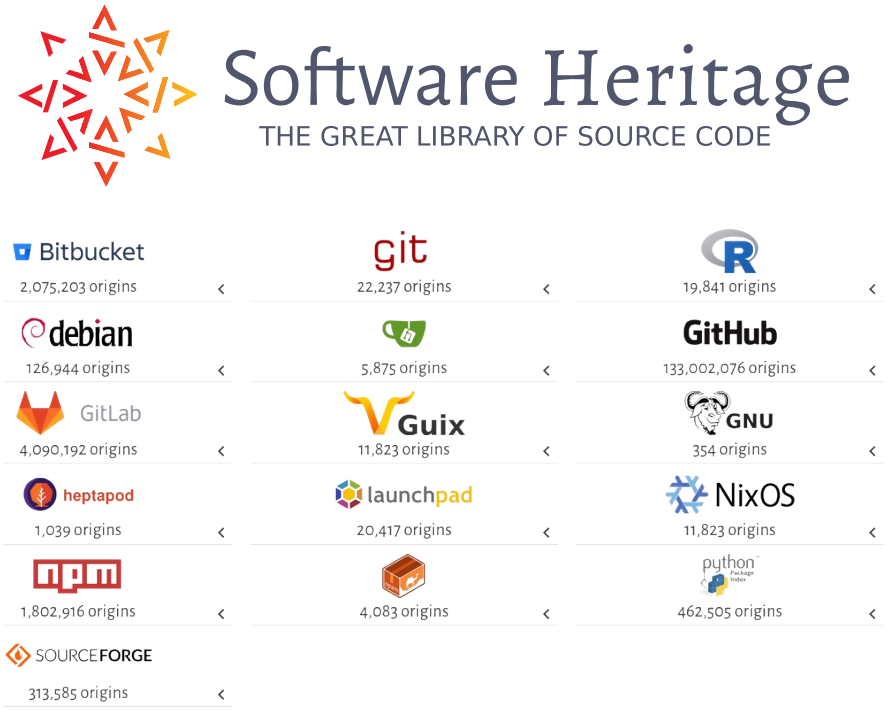
General topic: Diversity in contributions to Free/Open Source Software

- Significant body of research on **gender diversity**
- Less so on **geographic diversity**, constituted for the most part by:
  1) Survey-based studies, and/or
  2) Point-in-time studies

Contributions: we conduct the first large-scale longitudinal study of the geographic origin of contributors to public code over 50 years.

Research question: From which world regions do authors of publicly available commits come from and how has it changed over the past 50 years?

Dataset

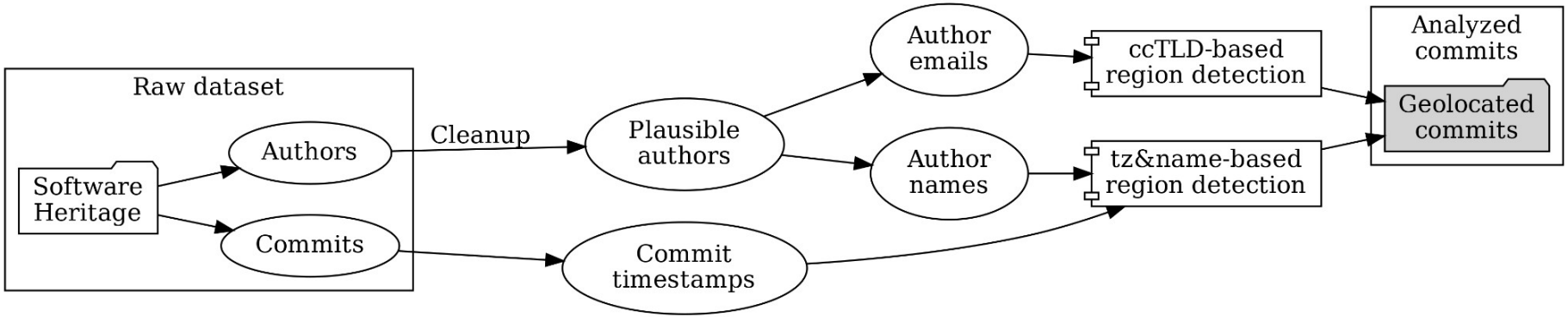


- 160 million public projects
- Revisions (commits) and their authors from 1970-01-01 to 2021-07-07 (50 years)
- 2.2 billion commits in total
- 43 million (unique) authors

World regions

- Granularity: world partition of 12 macro regions
- Based on the United Nations geoscheme, with adaptations to avoid under-represented regions and over-represented countries (to avoid they dominate a region)

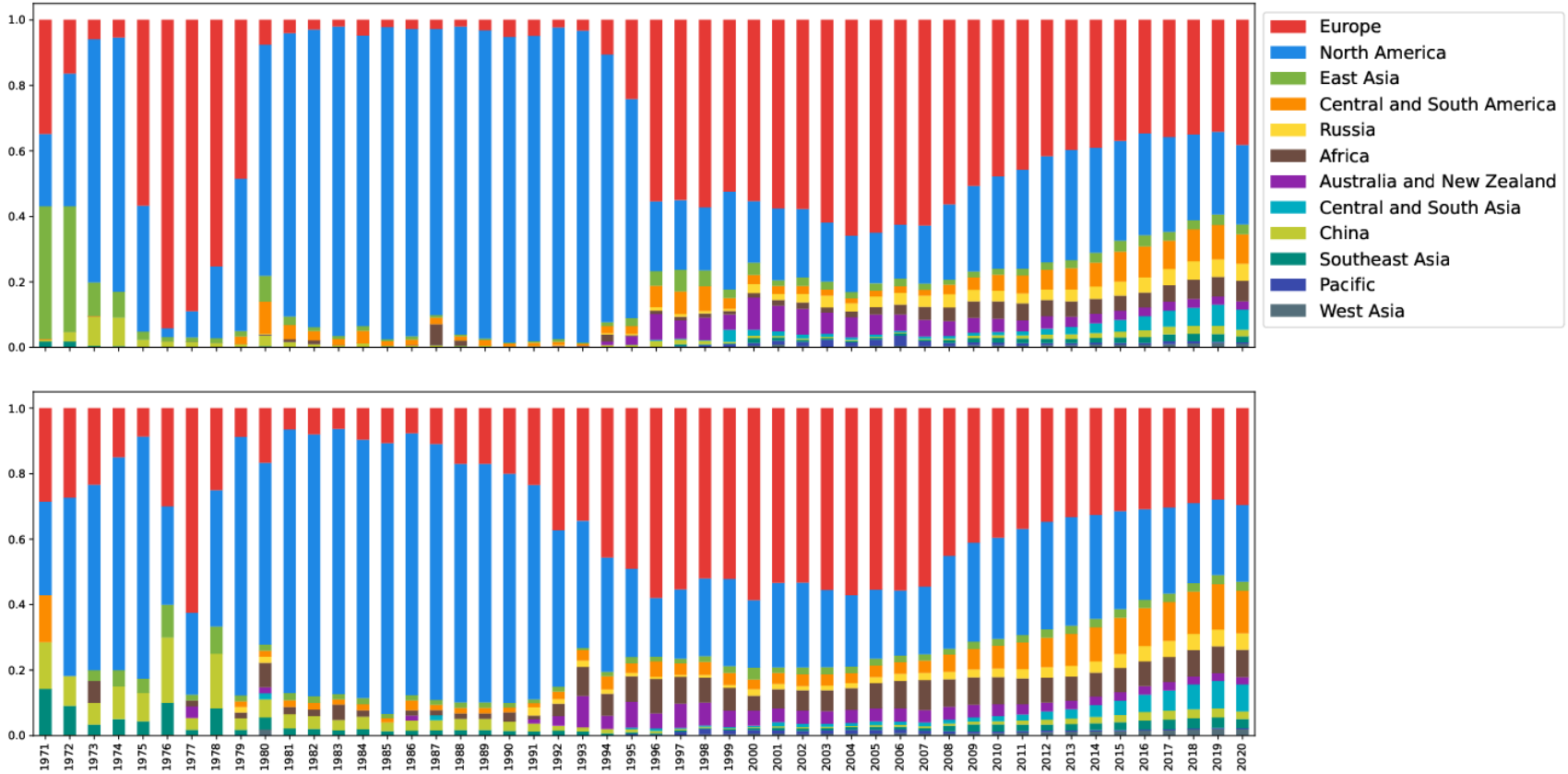
Data processing pipeline



- Two different **geolocation approaches** with complementary pros/cons:
  - ccTLD: Country-level **top-level domain in emails** (e.g., .fr, .it)
  - tz&name-based: **commit timezone offset**, compared against the countries that were in that timezone (at that time), and using **author names** to assign them to the most likely country among those (based on name frequencies)

(For details about the methodology see: Davide Rossi, Stefano Zacchiroli. *Worldwide Gender Differences in Public Code Contributions* at ICSE later this week.)

Raw results (commit ratios above, authors below)



Takeaways

- **Geographic diversity in public code is increasing over time**
  - North America and Europe alternated their dominance until mid 90s, then most other regions started growing stably overtime (albeit slowly)
  - Sudden drop of North America commit ratio from 1995 on, due to the disappearance of massive contributions from “@ucbvax.Berkeley.EDU” authors (UNIX wars, Computer Systems Research Group disbanded that year)
- We also found **traces of wide-social phenomenons** such as colonialism (e.g., via the adoption of popular European names in the Africa region) and immigration (e.g., via the increased popularity of popular Central/South American names in North America) → see paper for details

Threats and future work

- Construct validity
  - Accuracy/scale trade-off
  - Zone detection is based on heuristics
- External validity
  - The study dataset is the largest *approximation* of contributions to public code that is readily and publicly available for analysis
- Future work
  - Multi-method approaches should be applied to dig into phenomena affecting specific world-regions

Learn more

- Davide Rossi, Stefano Zacchiroli. *Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years*.
  - DOI: http://dx.doi.org/10.1145/3524842.3528471
  - Preprint: https://hal.archives-ouvertes.fr/hal-03622621v1
  - Talk to me at coffee break

- Related work later this week at ICSE: Davide Rossi, Stefano Zacchiroli. *Worldwide Gender Differences in Public Code Contributions (and How They Have Been Affected by the COVID-19 Pandemic).*

Appendix

Geolocation — details

- ccTLD-based detection
  - Straightforward
  - Problem: the adoption of national domains varies across zones
- tz&name-based detection
  - Likelihood measure of an author’s name to belong to a country
  - Limited to nations/territories in the commit’s time offset (at the time the commit was performed)
  - A score which is a sum of products name_frequency*population in that given offset for all candidate nations, aggregated by zone
  - Problem: time offset are not always reliable
- For this study: ccTLD if offset is zero, tz&name otherwise