



content.

BIG CODE DATASET

- ‡ 182,014 GitHub top-starred repos
- ‡ 248,043 Siva files
- ‡ 3 TB of Git repositories
- ‡ 54,5 million files in HEAD
- ‡ 15,941 million lines of code in HEAD
- ‡ 455 distinct languages
- ‡ 289 distinct licenses

index file.

COLUMN NAME

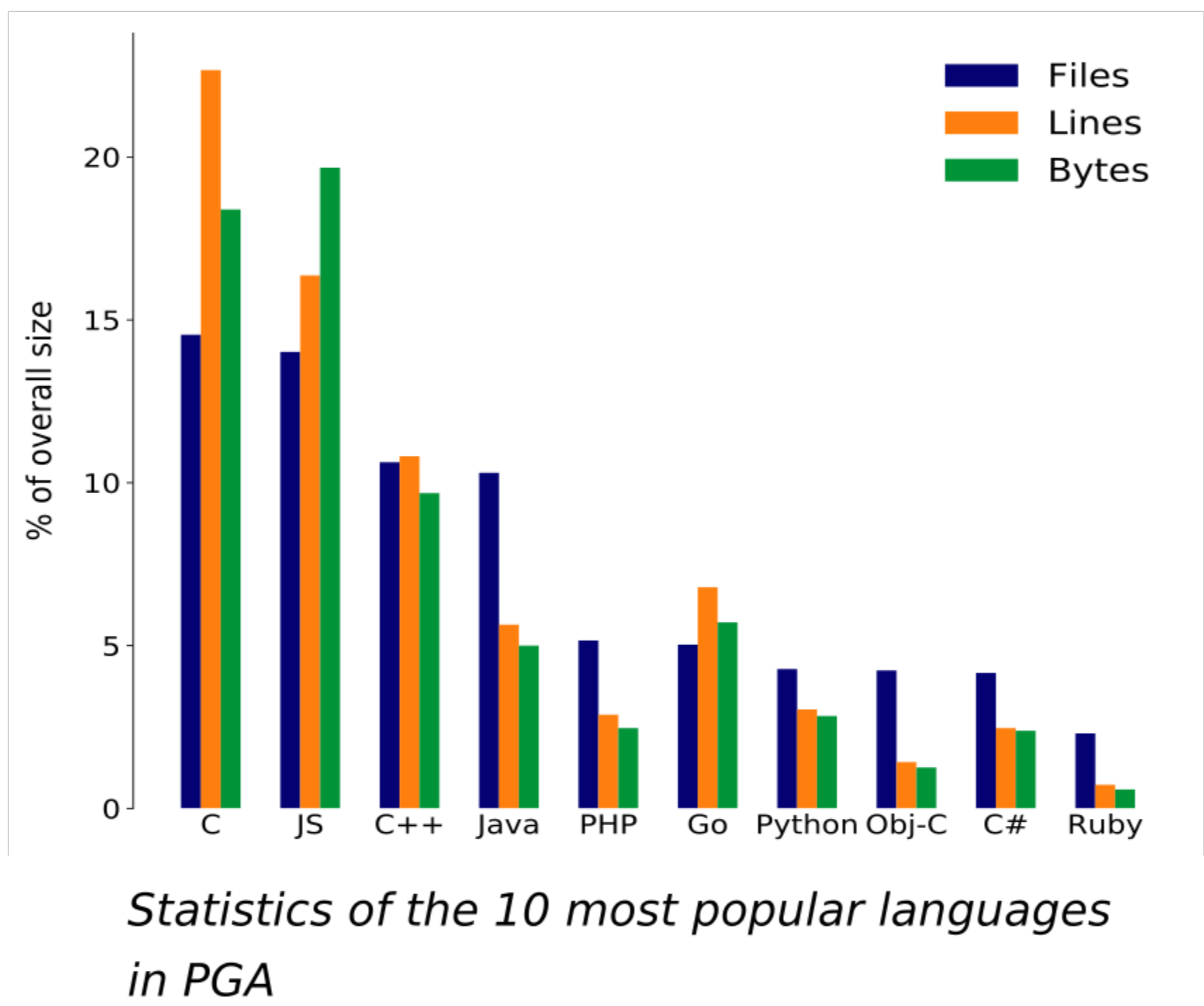
url
siva_filenames
file_count
langs
langs_{byte,lines,files}
commits_count
branches_count
fork_count
{empty,code,comment}_lines_count
license

DESCRIPTION

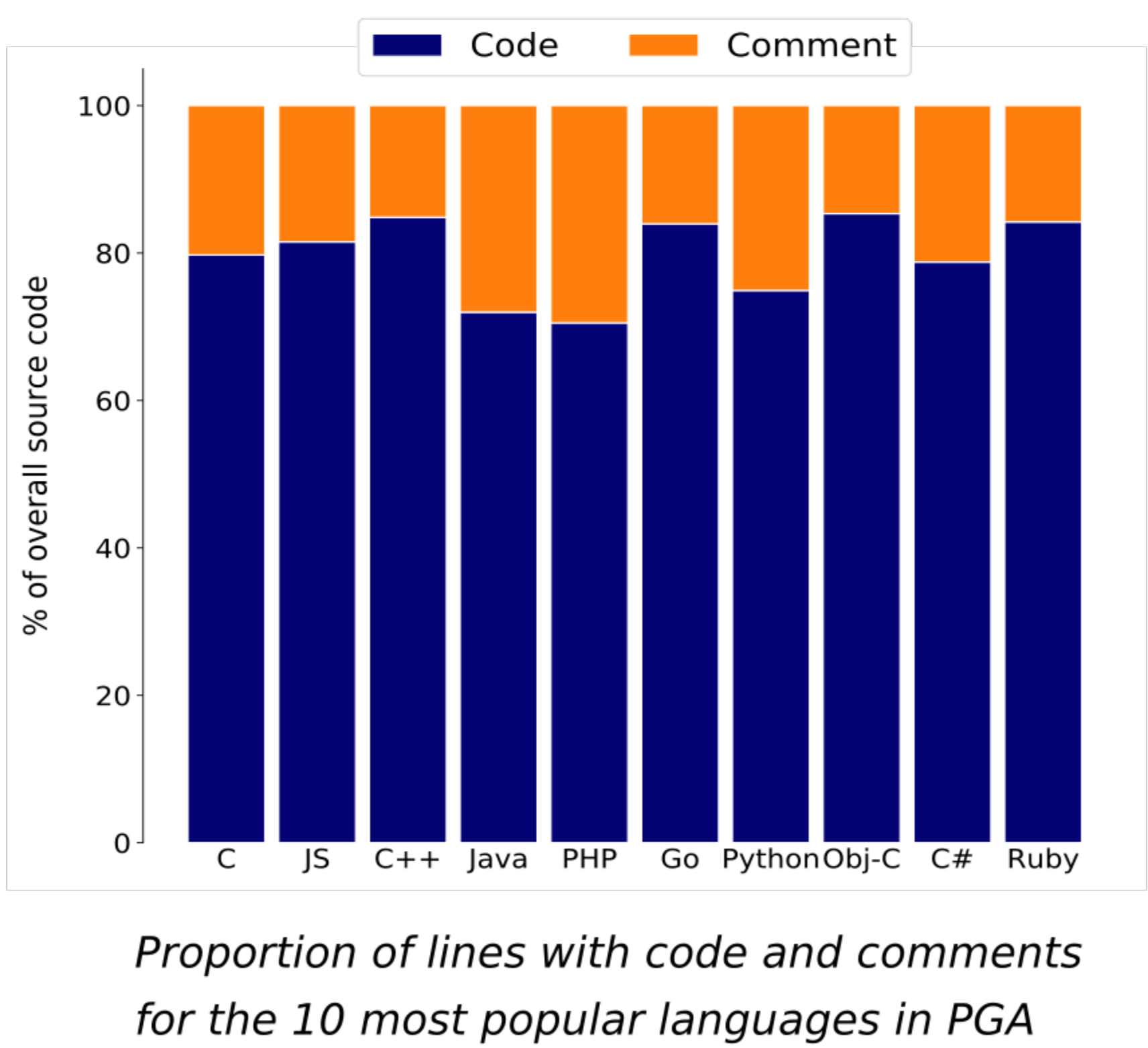
URL of the GitHub repository.
Siva files which contain parts of that repository.
Number of files in default HEAD reference.
Languages encountered in default HEAD.
Byte, line, file counts per each language, in the same order as langs.
Number of unique commits in the Siva files which refer to that repo.
Number of references, tags excluded.
Number of remotes in the referring Siva files.
Number of empty, code, commented lines in the default HEAD
License names and corresponding confidences.

description.

LANGUAGES



CODE VS. COMMENT



usage.

EXTRACT IDENTIFIERS

```
from sourced.engine import Engine
engine = Engine(spark, '/path/to/siva', 'siva')

engine.repositories.references.head_ref \
    .commits.tree_entries.blobs \
    .classify_languages() \
    .filter('lang = "Python"') \
    .extract_uasts() \
    .query_uast('/[*[@roleIdentifier]') \
    .extract_tokens('result', 'tokens') \
    .select('blob_id', 'path', 'tokens')
```

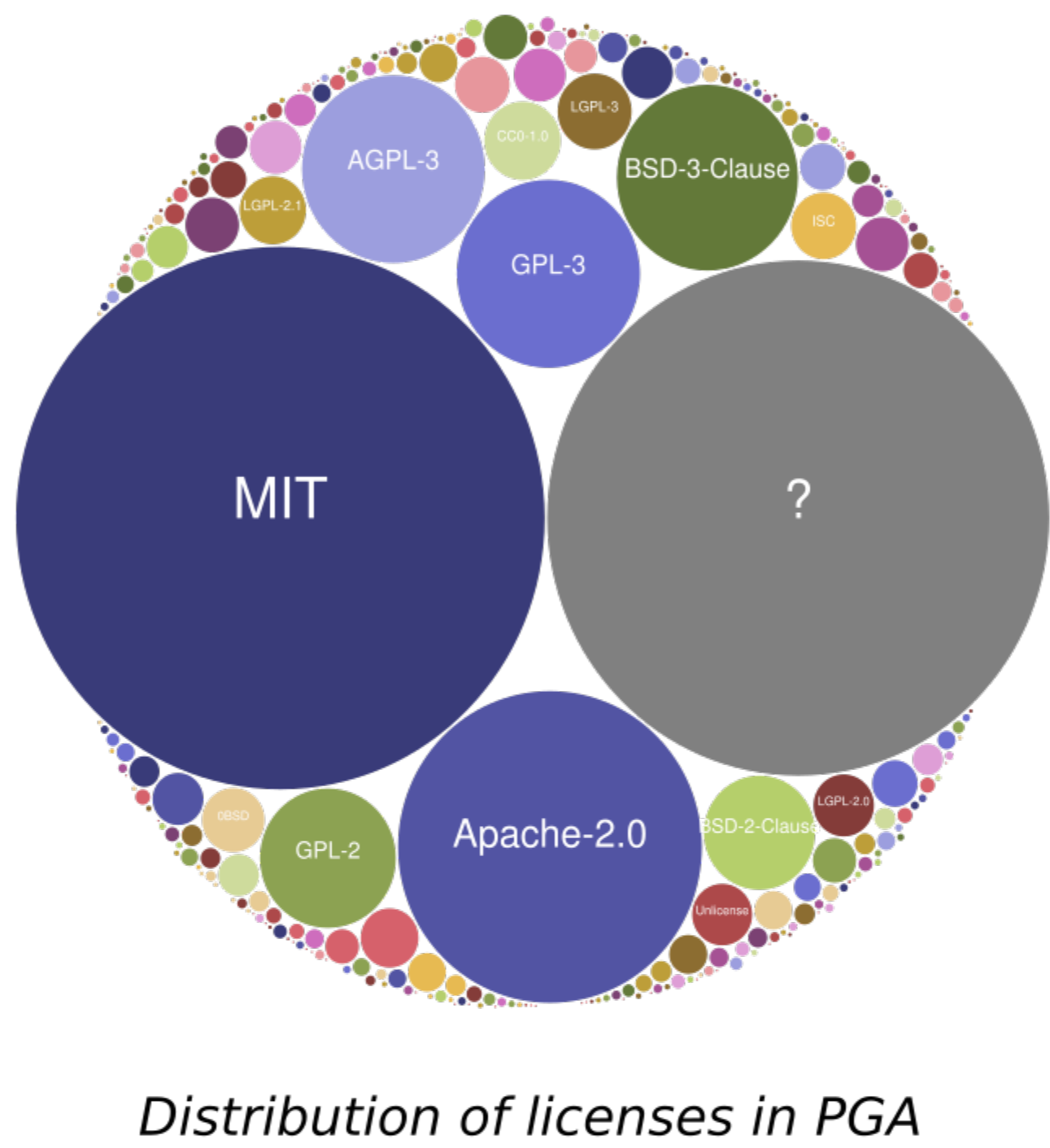
BRANCHES AND COMMITS

Repository	Branches
1. JetBrains/intellij-plugins	59,321
2. openstack/cinder	48,515
3. servo/servo	43,022
4. google/angle	38,327

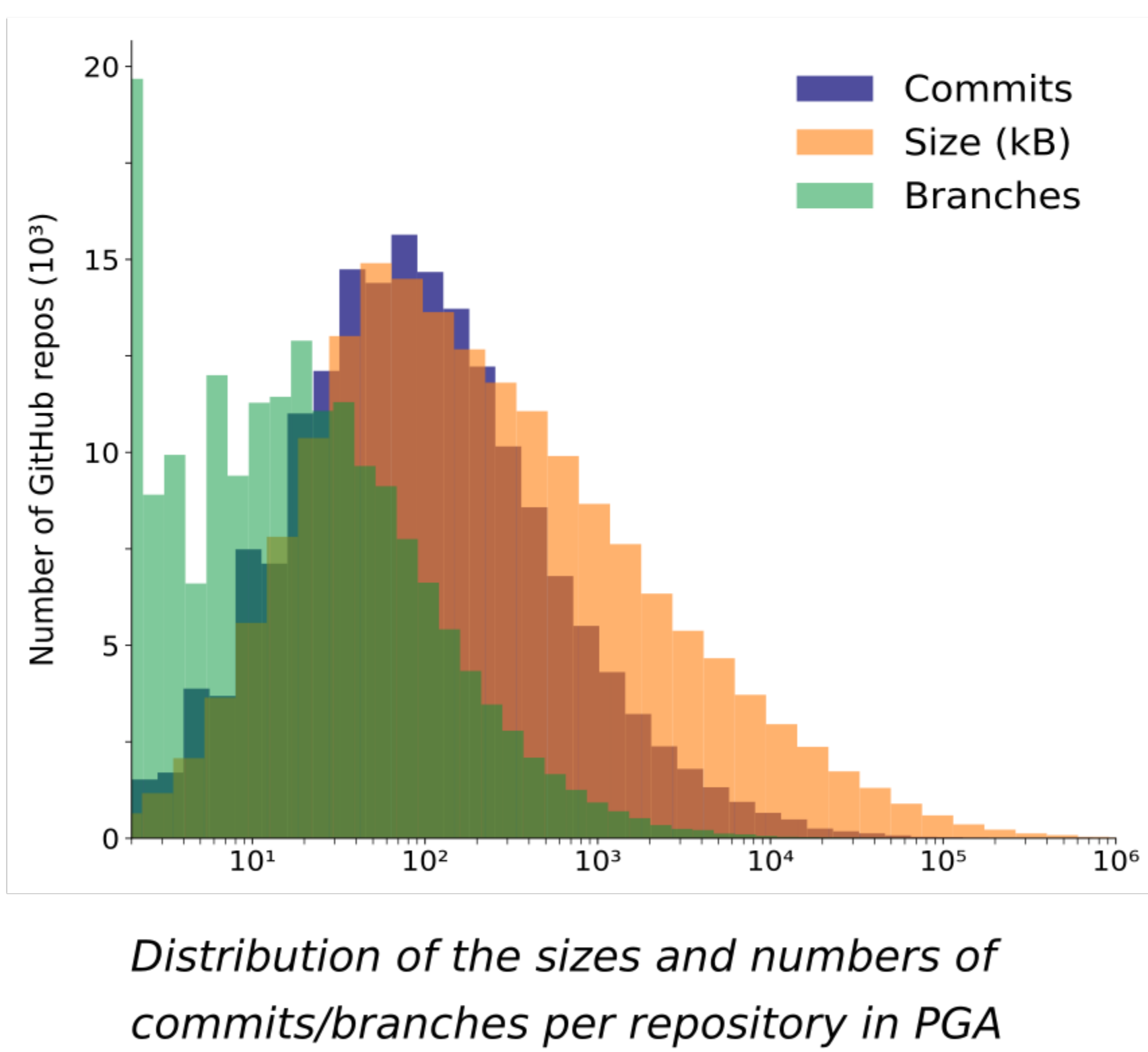
Repository	Commits
1. OpenChannelSSD/linux	735,972
2. google/ktsan	535,479
3. mirrors/chromium	260,366
4. Azure/azure-content	243,185

description.

LICENSES



BRANCHES AND COMMITS



pga tool.

LIST THE REPOSITORIES IN THE INDEX

- ‡ pga list --format csv to print CSV rows with all the details - supports also JSON
- ‡ pga list --lang java,go to list repos with at least some code in those two languages.
- ‡ pga list --url regexp to list repos for which the url matches the given regexp.

DOWNLOAD THE SIVA FILES

- ‡ pga get --output path path where to store the siva files - supports HDFS.
- ‡ pga get --jobs n to set the maximum number of concurrent download - default 10.

applications.

ML ON CODE USING PGA

- ‡ Source code analysis
- ‡ Identifier embeddings
- ‡ Topic modeling
- ‡ Code summarization
- ‡ Fuzzy duplicates detection
- ‡ Automatic program repair
- ‡ Code suggestion and completion