2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)

World of Code: An Infrastructure for Mining the Universe of Open Source VCS Data

Yuxing Ma[a], Chris Bogart[b], Sadika Amreen[a], Russell Zaretzki[a], Audris Mockus[a]
[a]University of Tennessee, Knoxville [b]Carnegie Mellon University
myxlvxm@gmail.com, cbogart@andrew.cmu.edu,
samreen@vols.utk.edu, russell.zaretzki@gmail.com, audris@utk.edu

***Abstract*—Open source software (OSS) is essential for modern society and, while substantial research has been done on individual (typically central) projects, only a limited understanding of the periphery of the entire OSS ecosystem exists. For example, how are tens of millions of projects in the periphery interconnected through technical dependencies, code sharing, or knowledge flows? To answer such questions we a) create a very large and frequently updated collection of version control data for FLOSS projects named World of Code (WoC) and b) provide basic tools for conducting research that depends on measuring interdependencies among all FLOSS projects. Our current WoC implementation is capable of being updated on a monthly basis and contains over 12B git objects. To evaluate its research potential and to create vignettes for its usage, we employ WoC in conducting several research tasks. In particular, we find that it is capable of supporting trend evaluation, ecosystem measurement, and the determination of package usage. We expect WoC to spur investigation into global properties of OSS development leading to increased resiliency of the entire OSS ecosystem. Our infrastructure facilitates the discovery of key technical dependencies, code flow, and social networks that provide the basis to determine the structure and evolution of the relationships that drive FLOSS activities and innovation.**

***Index Terms*—software mining, software supply chain, software ecosystem**

I. Introduction

Tens of millions of software projects hosted on GitHub and other forges attest to the rapid growth and popularity of Free/Libre Open Source Software (FLOSS). These online repositories include a variety of software projects ranging from classroom assignments to components, libraries, and frameworks used by millions of other projects. Such large collections of projects are currently archived in public version control systems, and, if made available and convenient for analysis, represent a unique opportunity to study FLOSS at large and answer both theoretical and practical questions that rely on the availability of the entirety of FLOSS data. In particular, this infrastructure, referred to as World of Code (WoC) and described below, allows researchers to conduct a census of open source software that would provide types and prevalence across projects, technologies, and practices and serve as a guide to setting policies or creating innovative services. Our infrastructure facilitates the discovery of key technical dependencies, code flow, and social networks that provide the basis to determine the structure and evolution of the relationships that drive FLOSS activities and innovation. Such a large database of software development activities can serve as a basis for "natural experiments" that evaluate the effectiveness of different software development approaches. If preserved, it will also facilitate future anthropological studies of software development [1].

Our objective in the current study is to describe a prototype of an infrastructure that can store the huge and growing amount of data in the entire FLOSS ecosystem and provide basic capabilities to efficiently extract and analyze that data at that scale. Our primary focus is on types of analyses that require global reach across FLOSS projects. A good example is a software supply chain where software developers correspond to the nodes or producers, relationships among software projects or packages represent the "chain", and changes to the source code represent products or information (that flow along the chain) with corporate backers representing "financing."

Several formidable obstacles obstruct progress towards this vision. The traditional approaches for obtaining the repository of a project or a small ecosystem does not scale well and may require too many resources and too much effort for individual researchers or smaller research groups. Thus, the community needs a way to scale and share the data and analytic capabilities. The underlying data are also lacking context necessary for meaningful analysis and are often incorrect or missing critical attributes [2]. Keeping such large datasets up-to-date poses another formidable challenge.

In a nutshell, our approach is a software analysis pipeline starting from discovery and retrieval of data, storage and updates, and transformations and data augmentation necessary for analytic tasks downstream. Our engineering principles are focused on using the simplest possible techniques and components for each specific task ranging from project discovery to fitting large-scale models. The result is a conceptual implementation loosely following the microservices architecture [3] where the design and performance of the loosely coupled components can be independently evaluated, each service can utilize a database that is optimal for its needs, and the most computationally-intensive components are extremely portable to ensure they run on any high-performance platform. More specifically, our prototype appears to capture a large portion of publicly available source code in version control systems and it will update quickly enough that the latency of updates on the existing hardware platform does not exceed one calendar month. Finally, a number of research tasks were effectively supported by the existing prototype.

We begin with an overview of related work in Section II,

2574-3864/19/$31.00 ©2019 IEEE
DOI 10.1109/MSR.2019.00031

describe the architecture of the prototype implementation in Section III, provide details of the components of the pipeline in Sections III-A to III-F. We conclude with a description of the experiences describing the attempts to enhance the prototype and to conduct several software analytics tasks in Section IV.

II. Related Work

While we are not aware of a complete census of FLOSS with an analysis engine, several large-scale software mining efforts exist and may be roughly subdivided into attempts at preservation, data sharing for research purposes, and construction of decision support tools.

Software development is a novel cultural activity that warrants preservation as a cultural heritage. The software source code, the only representation of software that contains human readable knowledge, needs to be preserved to avoid permanent loss of knowledge [1]. Software Heritage [1] is a distributed system involved in collecting and storing large amount of open source development data from various open source platforms and package hosts. It currently has software from GitHub, GitLab, Debian, PyPI, etc., and contains 88M projects, 1.2B commits, and 5.5B source files. The main drawback of this particular effort is the lack of focus on enabling applications to software analytics. The API provided allows for quick query of every historical particle in a software project and meets the preservation need, however, it does not grant the access to the full relationships (e.g., the set of projects containing a given commit) among these particles across entire collection of software. Quick access to these relationships is crucial in conducting software analytics such as identification of dependencies among artifacts and authors as well as code spread in open source community.

One potential value of archiving software lies in the reuse of software artifacts. For example, Nexus [4] repository manager, allows developers to share software artifacts in a standard way and provides support for building and provisioning tools (e.g. Maven) to access necessary components such as libraries, frameworks and containers.

Commercial efforts, such as BlackDuck or FOSSID[1] have proprietary collections they use to determine if their clients have included open source software within their proprietary software code. It is generally not clear how complete these collections are nor if the companies involved might consider opening them for research purposes.

In addition to source code and binaries, large scale collection of other software development resources could be integrated with the source code data. For example, GHTorrent [5]–[9] attempts to record every event for each repository hosted on GitHub and provides multiple approaches (SQL request and MongoDB data dump) for data access. The primary limitation is that the collected metadata is specific to GitHub and it does not include the underlying source code as well. Therefore, obtaining dependencies encoded within the source code cannot be accomplished. FLOSSmole [10] collects open source meta data from various forges as a base for academic research but only focuses on software project metadata.

Another platform is Candoia [11]–[14] which provides software development data collections abstraction for building and sharing Mining Software Repository (MSR) applications. In particular, Candoia contains many tools for artifact extraction from different VCSs and bug databases and it also support projects written in different languages. On top of these artifacts, Candoia created its general data abstraction for researchers to implement ideas and build tools upon. This design increased portability and applicability for MSR tools by enabling application on software repositories across hosting platforms, VCSs and bug recording tools. The approach is focused on the design and benefits of creating a specialized software repository mining language. While it abstracts a number of repository acquisition tasks, it also makes it more difficult to handle operational data problems that tend to occur at much lower levels of abstraction and tend to be too idiosyncratic for generalized abstraction. The main drawbacks of Candoia are that it only supports limited programming language (JS and Java) based projects, and ecosystem-wide research might be difficult to implement since Candoia relies on users to provide software related data (e.g., targeted software repository URL) and eco-system wide compliance is generally low.

Other platforms are aimed at improving reproducibility by providing a repository of datasets for researchers to share their data. These include PROMISE Repository [15], Black Duck OpenHub [16], and SourcererDB [17]. PROMISE Repository is a collection of donated software engineering data. It was created to facilitate generations of repeatable and verifiable results as well as to provide an opportunity for researchers to extend their ideas to a variety of software systems. Black Duck OpenHub is a platform that discovers open source projects, tracks the development and provides the functionality of comparison between softwares. Currently, it is tracking 1.1M repositories, connecting 4.2M developers and indexing 0.4M projects. SourcererDB is an aggregated repository of 3K open source Java projects that are statically analyzed and cross-linked through code sharing and dependency. On top of providing datasets, it also provides a framework for users to create custom datasets using their projects.

Apart from providing datasets (repository) for potential users, platforms such as Moose [18], RepoGrams [19], Kenyon [20], Sourcerer [21], and Alitheia Core [22] are more focused on facilitating building and sharing MSR tools. Moose is a platform that eases reusing and combining data mining tools. RepoGrams is a tool for comparing and contrasting of source code repositories over a set of software metrics and assists researchers in filtering candidate software projects. Kenyon is a data platform for software evolution tools. It is restricted to supporting only software evolution analysis. Sourcerer is an infrastructure for large scale collection of open source code where both meta data and source code are stored in a relational database. It provides data through SQL query to researchers and tool builders but is only focused on Java

[1] blackducksoftware.com,fossid.com

projects. Alitheia Core is a platform with a highly extensible framework and various plug-ins for analyzing software on a large database of open source projects' source code, bug records, and mailing lists.

Furthermore, there were efforts to standardize software mining data description for enhanced reproducibility [23]. None of the listed platforms focus on both collection and analysis of the dependencies of the entirety of FLOSS source code version control data. Further, they contain either limited collections (e.g. only GitHub, no source code, have only donated data, or do not contain an analysis engine). For example, it is not possible to answer simple questions such as "In which projects has a file been used?", "What projects/codes depend on a specific module?", "What changes has a specific author made?" etc.

Some large companies have devoted substantial effort to develop software analysis platforms for the entire enterprise, aiming to improve the quality of software they build and to help the enterprise achieve its business goals by providing recommendations to software development organizations/teams, monitoring software development trends, and prioritizing research areas. For example, Avaya, a telecommunications company, built a platform [24], which collects software development related data from most of its software development teams and third parties and enabled systematic measurements and assessments of the state of software. CodeMine [25], is a software platform developed by Microsoft that collects a variety of source code related artifacts for each software repository inside Microsoft. It is designed to support developer decisions and provide data for empirical research. We hope that similar benefits can be realized with the WoC platform targeted to the entire FLOSS community.

Large scale software mining efforts also include domain specific languages. Robert Dyer et al. developed Boa [26]–[31], both as a domain specific language and as an infrastructure, to ease open source-related research over large scale software repositories. The approach is focused on the design and benefits of an infrastructure and language combination. However, the lack of explicit tools to deal with operational data problems make it of limited use to achieve our aims. Their collection procedures -discovery, retrieval, storage, update, and completeness issues (for example, only certain languages are supported)- are not the primary focus of this effort. The tools to deal with operational data problems common in version control data are also lacking in Boa.

The system described in this paper is loosely modeled after a system described a decade ago [32], [33]. In comparison, at that time, git was just beginning to emerge as a popular version control system, but now it dominates the FLOSS project landscape. The number of software forges and individually hosted projects was much larger then in contrast to the consolidation of forges and the overwhelming dominance of GitHub. Furthermore, the scale of the FLOSS ecosystem is more than an order of magnitude larger now and it continues to experience very rapid growth. WoC could not, therefore, reproduce that design closely and, instead, is focused on preserving the original git objects and on creating a design that enables both efficient updating of this huge database and ways to cross-reference it so that the complete network of relationships among code and people is readily available.

III. Architectural Considerations

The process of mining individual git repositories is complex to begin with [34], but becomes even more difficult on a large scale [35]. More specifically, using operational data from software repositories requires resolution to three major problems [2]: the lack of context, missing attributes or observations, and incorrect data. This makes critical tasks such as debugging and testing complex and time consuming. To cope with these big data challenges we employed both vertical and horizontal prototyping [36]–[39]. Most big data systems use the layered data approach where initial layers approximate raw data and later layers include cleaned/augmented data.

In this section we present a prototype WoC implementation. It has four stages: project discovery, data retrieval, correction, and reorganization as shown in Figure 1.

A. Project Discovery

Millions of projects are developed publicly on popular collaborative platforms/forges such as GitHub, Bitbucket, GitLab, and SourceForge. Some of the FLOSS projects can be identified from the registries maintained by various package managers (e.g., CRAN, NPM) and Linux distributions (e.g., Debian, Fedora). Other project repositories, however, are hosted in personal or project-specific sites. A complete list of FLOSS repositories is, therefore, difficult to compile and maintain since new projects and forges are created and older forges disappear. There is a tendency for the FLOSS repositories to migrate to (or be mirrored on) several very large forges [40]. A number of older forges provide convenient approaches to migrate repositories to other viable forges before being shut down. This consolidation has alleviated some of the challenge of discovering all FLOSS projects [32], though the task remains nontrivial. We discuss several approaches to project discovery below. To package our project discovery procedure we have created a docker container[2] that has the necessary scripts.

Using Search API: Some APIs may also be used to discover the complete collection of public code repositories within a forge. The APIs are specific to each forge and come with different caveats. Most APIs tend to be rate limited (for user or IP address) and the retrieval can be sped up by pooling the IDs of multiple users.

Using Search Engine: Search engines (e.g., Google or Bing) can supplement the discovery of FLOSS project repositories on collaborative forges when the forge does not provide an API, or when the API is broken. The primary drawback is the incompleteness of the repositories discovered.

Keyword Search: Some forges provide keyword based search of public repositories, which is a complementary approach when a forge does not provide APIs for the enu-

[2]https://github.com/ssc-oscar/gather

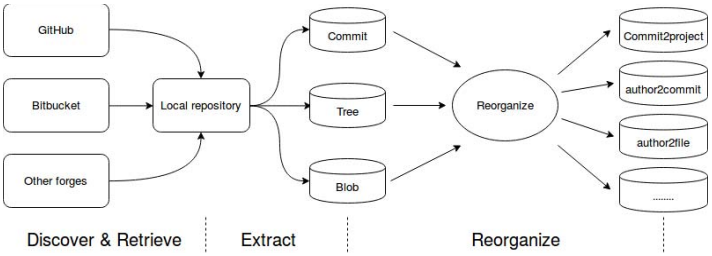

Fig. 1. Overarching data flow

meration of repositories and the results returned from search engines are lacking.

Using these and other opportunistic approaches helps ensure that they complement each other in approximating the publicly available set of repositories though it does not guarantee the completeness. We expected that various ways of crowdsourcing the discovery (with incentives to share a project's git URL) would help increase the coverage in the future.

B. Project Retrieval

This data retrieval task can be done in parallel on a very large number of servers but requires a substantial amount of network bandwidth and storage. The simplest approach is to create a local copy of the remote repositories via git clone command. As of December 2018, we estimate over 62M unique repositories (excluding GitHub repositories marked as forks, repositories with no content and private repositories). A single thread shell process on a typical server CPU (we used Intel E5-2670) with no limitations on network bandwidth clones randomly selected 20K to 50K repositories (the time varies dramatically with the size of a repository and the forge) in 24 hours. To clone 60M repositories in one week would, therefore, require from two to four hundred servers. We do not possess dedicated resources of such size and, therefore, optimize the retrieval by running multiple threads per server and retrieving a small subset of the repositories that have changed since the last retrieval. Specifically, we use five Data Transfer Nodes of a cluster computing platform[3].

C. Data Extraction

Code changes are organized into commits that typically change one or more source code files within the project. Once the repository is cloned as described above, we extract Git objects[4] from each repository and store these git objects in a single database.

*1) **Data Model**:* Git [41] is a content-addressable filesystem containing four types of objects. The reference to these objects is a SHA1[5] [42] calculated based on the content of that object. **commit** is a string including the SHA1's of commit parent(s) (if any), the folder (tree object), author ID and timestamp, committer ID and timestamp, and the commit message. **tree**: A tree object is a list that contains SHA1's of files (blobs) and subfolders (other trees) contained in that folder with their associated mode, type, and name. **blob**: A blob is the compressed version of the file content (the source code) of a file. **tag**: A tag is the string (tag) used to associate readable names with specific versions of the repository.

Fig. 2 illustrates relationships among objects described above. The snapshot at any entry point (commit) is constructed by following the arrows from left side to right side.

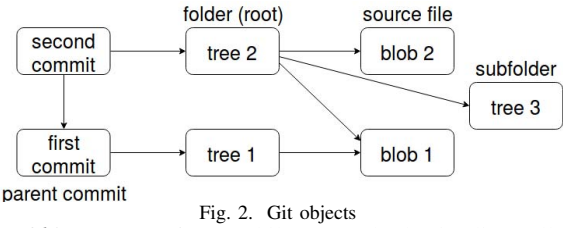

Fig. 2. Git objects

*2) **Object Extraction**:* While a standard Git client allows extraction of raw git objects, it displays them for manual inspection. For the bulk extraction need, first we list all objects within the git database, categorize them, and create bulk extractor based on a portable pure C implementation of *libgit2*[6]. We run listing and extraction using 16 threads on each of the 16-CPU node on a cluster[7]. The process takes approximately two hours for a single node to process 50K repositories. The extraction procedure represents a microservice.

D. Data Storage

The collection of public Git repositories as a whole replicate the same git object hundreds of times [32]. Without removing this redundancy, the required storage for the entire collection exceeds 1.5PB, and it also makes analytics tasks virtually impossible without extremely powerful hardware. Many reasons for this redundancy exist, such as pull-based development, usage of identical tools or libraries, and copying of code.

To avoid redundancy of git object among repositories, we store all git objects into a single database. The database is organized into four parts corresponding to each type of git object. Each part is further separated into a cache and content. The cache is used to rapidly determine if the specific object is already stored in our database and is necessary for data extraction described above. Furthermore, the cache helps determine if a specific repository needs to be cloned. If the heads (the latest commits in each branch in .git/refs/heads) of a repository are already in our database, there is no need to clone the repository altogether.

Cache database is a key-valued database, with the twenty byte Git object SHA1 being the key and the packed integer (indexing the location of the object in the corresponding value

[3]No. node: 300, Bandwidth up to 56 Gb/s
[4]https://git-scm.com/book/en/v2/Git-Internals-Git-Objects
[5]https://en.wikipedia.org/wiki/SHA-1
[6]https://libgit2.org/
[7]CPU: E5-2670, No. node: 36, No. core: 16, Mem size: 256 GB

database) being the value. The value database consists of an offset lookup table that provides the offset and the size of the compressed git object in a binary file (containing concatenated compressed git objects). While this storage allows for a fast sweep over the entire database, it is not optimal for random lookups needed, for example, when calculating diffs associated with each commit. For commits and trees, therefore, we also create a key value database where key is SHA1 of the git object and value is the compressed content of the said object. Cache performance is relatively fast: a single thread on Intel E5-2623 is capable of querying of 1M git objects in under 6 seconds, or over 170K git objects per second per thread. This can be multi-threaded and run on multiple hosts, thus reaching any desired speeds with expanded hardware.

Needless to say, with 12B objects occupying over 80TB we need to use parallel processing to do virtually anything. Thankfully, we can use SHA1 itself to split the database into pieces of similar size. We, therefore, split each of the database into 128 slices based on the first seven bits of Git object SHA1. This results in 128 key-offset cache databases for all four types of objects, 128 content databases as flat files for the four types of objects, and 128 key value databases for commits and trees: 128*(4+4+2) databases with each capable of being placed on a separate server to speed up parallel tasks. The individual databases containing content range from 20MB for tags up to over 0.5TB for blobs. The largest individual cache databases are over 2Gb for tree object SHA1s.

Databases are fragile and may get corrupted due to hardware malfunction, internet attack, pollution/loss by unrecoverable operation, etc. To enhance the robustness and reliability and to avoid permanent data loss, we maintain three copies of the databases: two copies on two separate running servers and one copy on a workstation that is not permanently connected to Internet. In the future, we will consider keeping a copy using a commercial cloud service.

Furthermore, due to the size of the data and complexity of the pipeline, some of the objects may have been missed or have been retrieved but are not identical to originals. Techniques to validate the integrity of the data at every stage of the process are necessary. We therefore, include numerous tests to ensure that only valid data gets propagated to the next stage.

In particular, the errors when listing and extracting objects are captured and the operation is repeated in case a problem occurs. The extracted objects are validated to ensure that they are not corrupt and also to ensure that they are not going to damage the database or the analytics layer. To validate correctness, the object is extracted per git specifications and recreated from scratch. The SHA1 signature is compared to ensure it matches that of the original object. A substantial number of historic objects have issues due to a bug in git that has since been fixed. Furthermore, a much smaller number of objects also had issues that we assume are either caused by problematic implementations of git or problems in operation (zero-size objects that may be occasionally created when git runs out of disk space during a transaction).

Despite the scrubbing and validation efforts, some of the data may still be problematic or missing, therefore a continuous process of checking the database for missing or incorrect data is needed. We plan to add missing object recovery service that identifies missing commits, blobs, and trees, and retrieves and stores them (in case they are still available online).

E. Update

The process of cloning all GitHub repositories takes an increasing amount of time with the growth in size of existing repositories and the emergence of new ones, given fixed hardware. Currently, to clone all git repositories (over 90M including forks), we estimate the total time to require six hundred single-thread servers running for a week and the result would occupy over 1.5PB of disk space. Fortunately, git objects are immutable and we can leverage that to simplify and speed up the updates. More generally, to get acceptable update times, we use a combination of two approaches:

- Identify new repositories, clone and extract Git objects
- Identify updated repository and retrieve only newly added Git objects

The work flow is illustrated in Fig. 3.

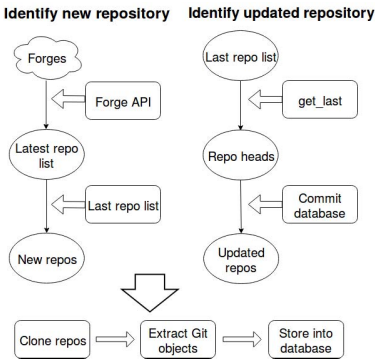

Fig. 3. Update workflow

In fact, only approximately three million new projects were created and an additional two million updated during Dec, 2018.

*1) **Procedures for new repositories**:* Forge-specific APIs are utilized to obtain the complete list of public repositories as described above. A comparison with prior extract yields new repositories. The list may include renamed repositories and forks. We can exclude forks for GitHub, since it is an attribute returned by GitHub API. Other forges contain fewer repositories, so the forks are not large enough to be a concern.

*2) **Procedures for updated repositories**:* First we need to identify updated repositories from the complete list of repositories. Since we are not sure how GitHub determines the latest update time for a repository, we use a forge-agnostic way of identifying updated repositories. We modified the *libgit2* library so that we can directly obtain the latest commit of each branch in a Git repository for an arbitrary Git repository URL, without the need to clone the repository. If any of the heads contain a commit that is not already in our database, the repository must have had updates and needs to be obtained.

We are working on a strategy to reduce the amount of

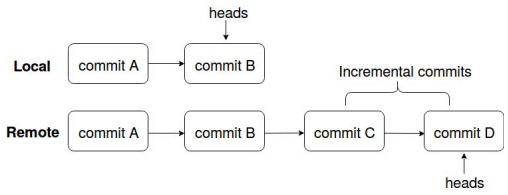

Fig. 4. Incremental commits

bandwidth needed to do the updates. Instead of cloning an updated repository, we'd like to retrieve only incremental Git objects (see Fig. 4) that are generated during the time gap between two consecutive updates. This can be easily done via git fetch for a git repository, but since we do not keep the original git repository and it is time consuming to prepopulate it with git objects, we plan to customize git fetch protocol by inserting additional logic in order to use our database backend that comprises git objects from all repositories. The procedure consists of two steps:

1) Customize git fetch protocol[8] to work without git's native database.
2) Keep track of the heads for each project that we have in our database so that we can identify latest commits to the modified git fetch.

For the second step, the database backend will use the project name as input and provide the list of heads for the project. These heads are then sent to the remote so that the set of latest commits (and related trees/blobs) will be calculated out and transferred back as illustrated in Figure 5. By following this strategy, we could drastically speed-up mining incremental Git objects from repositories in each update.

F. Data Reorganization for Analytics

Objects in Git are organized in a way for fast reconstruction of a repository at each commit/revision. In fact even the seemingly simple operation of identifying what files changed in a commit is computationally intensive. Furthermore, there is no consideration for the projects, files, or authors as first-class objects. This limits the usability of the git object store for research and suggests the need for an alternative data design. Since our objective is to obtain relationships among projects, developers, and files, we have created an alternative database that allows both a rapid lookup of these associations and sweeps through the entire database that make calculations based on such relationships.

*1) **Analytic Database**:* The scale of the desired database limits our choices. For example, a graph database [9] like neo4j would be extremely useful for storing and querying relationships, including transitive relationships. However, it is not capable (at least on the hardware that we have access to) of handling hundred's of billions of relationships that exist within the entire FLOSS. In addition to neo4j, we have experimented with more traditional database choices. We evaluated common relational databases MySQL and PostgreSQL and key value databases or NoSQL [43] databases MongoDB, Redis, and Cassandra. SQL like all centralized databases [44] has limitations handling petabyte datasets [45], [46]. We, threfore, focus on NoSQL databases [47] that are designed for large scale data storage and for massively parallel data processing across a large number of commodity servers [47].

For the specific needs of the cache database and for key value stores for the analytics maps we use a C database library called TokyoCabinet (similar to berkeley_db) using a hash-indexed as described above, to provide approximately ten times faster read query performance than a variety of common key value databases such as MongoDB or Cassandra. Much faster speed and extreme portability lead us to use it instead of more full-featured NoSQL databases.

*2) **Maps**:* Apart for the general requirement to be able to represent global relationships among code, people, and projects, we also consider the basic patterns of data access for several specific research tasks as use cases in order to design a database suitable for accomplishing research tasks within a reasonable time frame. The specific use cases are:

1) Software ecosystem research would need the entire set of repositories belonging to a specific FLOSS sub-ecosystem, e.g., the set of all repositories that use Python language.
2) Developer behavior research would need to identify all projects that a specific developer worked on, the files they authored, and software technologies they used.
3) Code reuse research would need to identify all projects where a specific piece of code occurs and determine how it got there.

To support the first task, a mapping from file names to project names would be necessary. The second task would require author to project, file, and to content of the versions of the file authored by that developer (in order to access the source code and identify what components or libraries were employed). The last task would require a map between blobs (that contain snippets of code) and projects. It would also require a map between blobs and commits in order to identify the time when the specific piece of code was introduced.

We have identified a number of objects and attributes of interest here: projects, commits, blobs, authors, files, and time. The complete set of possible direct maps for an arbitrary pair is 30. Since author and time are properties of the commit and are not properties of projects, blobs, or files, it makes sense to place commit at the center of this network database. The author-to-file map can then be constructed as a composition of author-to-commit and commit-to-file maps; and author-to-project map can be constructed via author-to-commit and commit-to-project maps. We also need to associate file names with the corresponding blobs since a single commit may create multiple files. Out of the 12 maps[10], only 10 need to be instantiated because commit-to-author and commit-to-time maps are embedded as the properties of the commit object.

[8] git fetch downloads only new objects from the remote repository

[9] a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data

[10] bidirectional maps between the commit and five objects/attributes and between file and blob

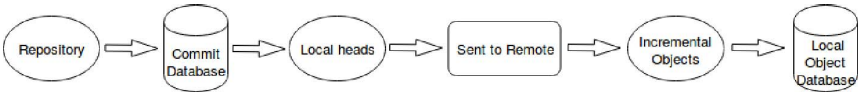

Fig. 5. Future workflow

In addition to having the commit at the center, for certain tasks we also needed to have a blob-to-file map as well. For example, we want to identify module use in Python language files. First, we need to identify relevant files via suitable extension (e.g., .py), then we can determine all the associated commits via file to commit map. These commits, however, may involve other files and if we use commit to blob map to identify associated blobs, we would get blobs not just for python, but also for all files that were modified in commits that touched at least one python file. The file-to-blob map allows us to reduce the number of blobs that need to be analyzed dramatically.

In addition to these basic maps we create additional maps, such as the author ID to author ID map for IDs that have been established to belong to the same person (see Section IV-B), and project to project maps to adjust for the influence of forking. Project-to-project maps are based on the transitive closure of the links induced between two projects by a shared commit. Explicit forks that can be obtained as a GitHub project property do not generalize to other forges and, even on GitHub, represent only a fraction of all repositories that have been cloned from each other and then developed independently. Project-to-project map also handles instances where repositories exist on multiple forges or when they are renamed.

As with the original data we utilize multiple databases and use compressed files for sweep operations and TokyoCabinet for random lookup. We separate maps into 32 instead of 128 databases we use for the raw objects since maps tend to be much smaller in size than, for example, blobs. For commits and blobs we use the first character of SHA1 for database identification. For authors, files, and projects, we use the first byte of FNV-1a Hash [11]. Both approaches yield quite uniform distribution over bins.

As noted above, the maps from commit to meta data are not difficult to achieve because meta data are part of the content of a commit object. However, git blobs introduced or removed by a commit are not directly related to the commit and need to be calculated by recursively traversing trees of the commit and its parent(s). A Git commit represents the repository at the-state-of-world and contains all the trees (folders) and blobs (files). To calculate the difference between a commit and its parent commit, i.e., the new blobs, we start individually from the root tree that is in the commit object, traverse over each subtree and extract each blob. By comparing two sets of blobs of each commit, we obtain the new blobs for the child commit. This step requires substantial computational resources, but the map from the commit to the blobs authored in a commit is used in numerous research scenarios and, therefore, is necessary. On average, it takes approximately one minute to obtain changed files and blobs for 10K commits in a single thread. With 1.5B commits, the overall time for a single thread would take 104 days, but it needs to be done only on approximately 20-40M new commits generated each month.

[11] http://www.isthe.com/chongo/tech/comp/fnv/index.html#FNV-1a

IV. Applications

To evaluate if the experimental platform is capable of supporting research tasks conducted as a part of actual investigations and to provide a set of vignettes for other researchers, we conducted two types of studies. First, we implemented several basic and involved research tasks that require the entirety of FLOSS data as a part of the investigation. Furthermore, we also recruited three researchers external to our group to either conduct investigations of their own utilizing WoC or to provide us with their research problems that can only be solved by using WoC. Below we report both the experiences and results from these experiments.

A. Use of programming languages

Language popularity may influence developers decisions as it may affect the market for their software as well as their job prospects. For example: What language-specific API should developer provide for their component? What language should the developer use to implement their product?

To plot, for example, Java language use trend we use WoC to identify all files with .java extension. Then, via file-to-commit map, obtain the complete set of commits authoring these files. Commit dates are used to plot the time trends of language-specific commits, authors (property of a commit), projects (via commit to project map) and, if desired, lines of code changed. The entire process is highly parallelisable since each map is separated into 32 instances and can be processed independently. The entire calculation, while not interactive on our hardware, can be performed in tens of minutes. For illustration, we show the ratio of the number of commits over the number of developers (a measure of productivity) each month in Fig. 6. The ratio decreases for most languages, perhaps because as a language becomes more popular, the less experienced contributors join and lower the average productivity.

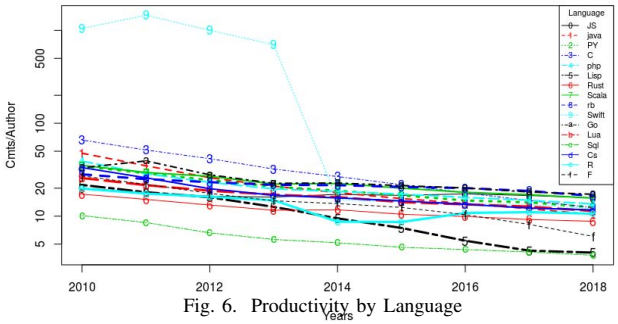

Fig. 6. Productivity by Language

B. Correcting Developer Identity Errors

One of the particularly troubling data quality issues with version control systems is developer name disambiguation. Often, names and emails of developers are missing, incomplete, misspelled or duplicate [48], [49]. Performance of any disambiguation algorithm depends on the distribution of the actual misspellings in the underlying data. In order to design and evaluate corrective algorithms, it is important to study a large collection of actual data and unearth patterns of irregularities that compromise data quality. WoC contains a nearly complete collection of git author ids (name and email combinations) and is, thus, more representative of such irregularities than any specific project.

To obtain author IDs we use author-to-commit map containing roughly 30 million distinct author IDs. Common error patterns include organizational ids and emails (Mozilla, Linux, Google etc), names of tools and projects (OpenStack, Jenkins, Travis CI), roles such as (admin, guest, root etc.) and words that preserve anonymity (student, nobody, anonymous etc) as a part of their credentials. We also found a large number developer IDs to be misspelled.

Traditional identity correction approaches rely on the misspelling patterns of author ID (the full name and email) [49]–[51]. With WoC data, we can enhance the traditional string matching with behavioural comparison, by creating similarity measures between author IDs using files modified by developers, time patterns of commits, and writing styles in commit messages. For illustration — two author IDs that modify a similar set of files may suggest that these IDs belong to the same developer. To implement file-based similarity, we used author to commit and commit to file maps to obtain the set of files modified by a single author ID. Then file-to-commit and commit-to-author maps were used to calculate similarity using weighted Jaccard measure. Commit message text was used to fit a Doc2Vec [52] model to associate each author ID with their writing style. Traditional and behavioural similarities were used to train highly accurate machine-learning model [53].

This experiment demonstrates the utility of WoC data for designing tools to solve common and vexing data quality problems when constructing developer networks. It is also an example of how WoC can be enhanced by incorporating such techniques and providing corrected data to researchers.

C. Cross-ecosystem comparison studies

A second research group used the database to gather comparative statistics about different software ecosystems. The purpose was to supplement other comparative data about those ecosystems in support of a study of how ecosystem tools and practices influence development behavior. The ecosystem study involved a survey, interviews, and data mining over 18 ecosystems whose repositories listed more than 1.2M packages. Some questions about ecosystem practices could be mined from metadata available elsewhere; for example detailed information about dependencies, release frequency, and version numbering practices can be easily extracted from libraries.io[12]. However deeper questions about project content would have been out of reach without WoC; independently building the mechanism to collect all of these projects, building a database of blobs, files, projects, and authors, and comparing them using various metrics would have been too much work for too little gain without the availability of this research platform.

*1) **File cloning across ecosystems**:* One such statistic is rate of file cloning. It was theorized that in ecosystems with more flexible support for dependencies and a tolerance for the risk of breaking changes, developers would be more likely to use dependency management tools to make use of functionality from other projects, rather than copying those files in directly; hence in such ecosystems we should find relatively few commits adding a blob that already exists in any other project available through the ecosystem's dependency management system.

Using WoC, this analysis was straightforwardly accomplished by joining blob-to-commit and commit-to-project mappings, filtering for blobs that appeared in multiple projects, and identifying pairs with one commit in the time frame, and at least one older commit. Such blobs were discarded when the files were very small (since these often turned out to be empty or trivial files duplicated by chance or by tools) resulting in a set of duplicates that, on visual inspection of a sample, did appear to represent genuine examples of reuse-by-cloning.

Contrary to our expectations, the ecosystem with the most propensity for cloning was the one with the modern and flexible dependency system: npm. Despite the strengths of npm's dependency management system, there is a strong tradition of copying dependencies like jQuery into projects rather than letting npm retrieve them. Figure 7 summarizes the findings for a selection of ecosystems.

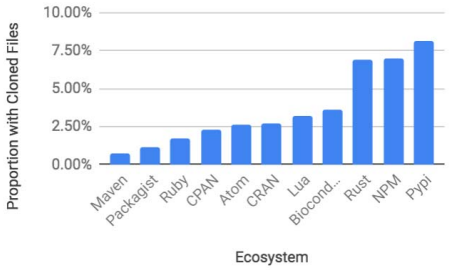

Fig. 7. Proportion of repository packages that added at least one cloned code file over 1kb in 2016.

*2) **Developer migration across ecosystems**:* Another metric of interest was developer overlap between ecosystems. Our ecosystem comparison had included a survey of values and practices in the 18 ecosystems of interest, and we hypothesized that ecosystems might be similar if many developers were actually working in both ecosystems, or had migrated from one to the other.

This question was answered by joining author-to-commit and commit-to-project data for the 1.2M projects in our study,

[12]https://libraries.io/

and relying on the identity matching technique described in Sec IV-B.

Over all pairs of ecosystems, we found a sizable correlation between similarity of average responses on ecosystem *practice* questions (things like frequency of updating, collaboration with other projects, means of finding out about breaking changes), and overlap in committers to those ecosystems (Spearman $\rho = 0.341, p < .00001, n = 16$ ecosystems). Interestingly, perceived *values* of the ecosystem (such as a preference for stability, innovation, or replicability) do *not* seem to align with developer overlap ($\rho = -0.05, p = 0.44$). While more research is needed, we hypothesize that developers may carry practices over from other languages and platforms they have used in the past, in a sometimes cargo-cult-like way, despite recognizing that a new ecosystem is designed to accomplish different ends.

In our very large-scale, wide-ranging study, these questions of developer migration and cloning were of great interest, but would likely have been too expensive to pursue alongside other lower-hanging fruit, absent WoC's prepared set of precomputed maps between files, blobs, authors, projects, and timestamps. The dataset with its analytical maps was not designed with these particular ecosystem comparison in mind, but its design happens to make such ecosystem questions relatively easy to answer.

D. Python ecosystem analysis

An external researcher wanted to use WoC to investigate open source sustainability by identifying source code repositories for packages in PyPI ecosystem and to measure package usage directly. While over 90% of npm packages provide repository URLs, less than 65% of Python Package Index (PyPI) packages do.

The researcher obtained all packages from PyPi and calculated blob SHA1s for *setup.py* file of the first PyPI releases of each package. We filter out resulting 101584 blobs to exclude empty or uninformative blobs (blobs that appear in more than one commit using blob-to-commit map). The 54218 informative blobs are then mapped to 54062 unique commits and commits to 51924 unique projects (adjusted for forking as described in Section III-F). Repositories were recovered for 96% of the 54218 original packages in approximately 20 minutes of computation. To ensure that these repositories are, in fact, used to version control corresponding packages, they can be matched via additional blobs for *setup.py* and other files obtained from PyPi for that package.

Another problem being solved by this researcher was identifying which of the seemingly abandoned projects may be "feature complete," i.e. already have the intended scope and do not require further maintenance [54]. Feature complete projects should be widely used in contrast to abandoned projects. Proxies of project usage, e.g., GitHub stars or forks can be used to identify such projects [54]. WoC, however, lets us measure the extent of use directly. As described in Section IV-A, all commits modifying Python files are identified (file-to-commit map) and the resulting commits are mapped to projects (commit-to-project map). Blobs associated with these commits (commit-to-blob map) are then used to extract imports from these files. The entire procedure could be completed in approximately four hours using the parallelism of the analytic maps (32 databases) and blob content maps (128 databases).

The reported usage was compared to project development activity, i.e the total number of adoptions versus the total number of commits. In some cases, usage was not accurately reflected in the number of commits. Common examples are packages providing console scripts and CMS-like projects. In the former case, packages are not reused in programmatic code and thus don't get into statistics. In the latter case, website builders often do not publish their code and thus such usage remains unobserved. Therefore, while the number of public reuses provides some extra information about package use, it should be adjusted for package type.

E. Repository filtering tool

Millions of repositories on GitHub and other forges also include projects that are completely unrelated to software development. GitHub is widely used for education and other tasks such as backing up text files, images, or other data. Researchers investigating education may need to focus on tutorials, while other researchers may need a sample of actual software development projects. Furthermore, a way to select specific subsets of software development projects in order to conduct, for example, "natural experiments" would also be highly beneficial. WoC can support such project segmentation tasks in a variety of ways. An external education researcher wanted to understand the impact of self-administered programming tutorials. To do that, WoC was used to identify developers who participated in tutorials by searching the set of projects in WoC via keywords related to education: "assignment", "course", "homework", "class", "lesson", "tutorial", "syllabus", "mooc", "udacity". The search yields over 1M projects. While it is only a small fraction of all projects in WoC but it represents a large sample in absolute terms. Further filtering was needed to find developers who also worked on actual software projects to measure the impact of self-administered tutorials. The project-to-commit map identified 605K users of tutorials and, when these users were mapped to all projects they participated in, we determine that only half of them contribute to non-tutorial projects. These 300K individuals are potential subjects of tutorial-impact study. Further information (such as their commit activity and project participation) can be obtained from WoC and combined other data, be used in this research. WoC can be extended with other approaches to segment projects[13]. For example, identification of projects with sound software engineering practices [55] relies on a combination of factors easily obtainable in WoC, such as history, license, and unit tests.

V. Future work

To have an impact on research practice, the WoC prototype needs to be exposed via reliable services that help with

[13]Section IV-B shows how WoC can also be used to improve them

research and do not overwhelm the platform. WoC should also accommodate additional data and computational procedures needed for discovering, correcting, cleaning, augmenting, and modeling the underlying data. Processing hundreds of terabytes of data on powerful clusters may be out of reach for most research groups. Therefore, to accommodate massive queries WoC would require more powerful hardware. Such hardware can be obtained from cloud vendors, but the costs of hosting and analyzing data on these platforms might be high. An alternative might be a few high-throughput services that work on the hardware we currently employ.

The differentiating features of WoC are the completeness of the collection and access to global relationships. Specifically, two basic services would be difficult to replicate outside WoC, yet be capable of high throughput on the limited hardware. First, a reporting service that considers prevalence of certain features, such as languages, tools, and other technologies as well as the information about contributors might provide services akin to those provided by a population census. The second basic service would focus on identifying all entities linked to a specific entity, such as files modified by a developer, all repositories containing a specific code, or all files that use a specific module or technology. These two capabilities, in conjunction with MSR technology already in use, would provide both, population-level data and complete links within entire FLOSS ecosystem. It would then be up to researchers to retrieve additional data on individual projects based on the stratified samples from the first service or derived from the relationships obtained from the second service.

VI. Limitations

We tried to make the assumptions and rationale for specific decisions clear within each section but it is important to reiterate at least some of the limitations. Despite a large size (the collection contains over 1.45B commits), there is no guarantee it closely approximates the entirety of public version control systems as the project discovery procedure is only an approximation. Our focus on git (due to the simplified global representation) excludes older version control systems that have not been converted to git yet. We regularly identify issues with data being incomplete due to collection, cleaning, or processing and we are working on an approach to continuously validate and correct it. The particular design decisions were focused on the particular computing capabilities that were available to us at the time and could/should be revisited as the prototype evolves. The entirety of research tasks that WoC provides is not exhausted by the few examples we have investigated and certain tasks may require different solutions. We do, however, think that the micro-services approach allows for simpler addition/extension/replacement of components as needs or opportunities arise than would be possible with a more monolithic architecture.

How to reliably clean, correct, integrate, and augment the collected data so that the resulting analyses accurately reflect the modeled phenomena is a concern. To ensure the performance of the analytics layer certain objects are filtered from it. For example, some of the public repositories are created to test the performance/capabilities of git and contain many millions of files/blobs in a single commit. Such commits are excluded from the analytics layer to speed-up the commit-to-file and commit-to-blob maps. The nature of the data may also create performance problems. For example, the most common blob is an empty file. Mapping such blobs to all commits that create them or to all files does not make sense, since there are millions of commits that have created empty files. These performance-related modifications may affect some arguably superficial analyses, e.g., what are the commits with the largest number of files? We explicitly highlight these modification in the WoC code to minimize potential confusion.

Reproducibility may pose an issue in a constantly updated database. Since git objects are added incrementally and order in which they are stored is preserved, we can reconstruct any past version of the object store. For the analytic layer, which depends on the set of git objects available at the time, we create versions, where each of the maps described above is tagged with a version identifying the state of git object store. Preserving these past versions ensures reproducibility of the results obtained from them.

The research use cases presented do not constitute an empirical evaluation of WoC usability but, instead, focus on presenting vignettes that are effective for these scenarios. Some of these vignettes went through several iterations until the simplest and fastest implementations were obtained.

VII. Conclusions

We introduce WoC: a prototype of an updatable and expandable infrastructure to support research and tools that rely on version control data from the entirety of open source projects and discuss some of the research problems that require such global reach. We discuss how we address some of the data scale and quality challenges related to data discovery, retrieval, and storage. Furthermore, we implement ways to make this large dataset usable for a number of research tasks by doing targeted data correction and augmentation and by creating data structures derived from the raw data that permit accomplishing these research tasks quickly, despite the vastness of the underlying data. Finally, we evaluated WoC by conducting actual research tasks and by inviting researchers to undertake investigations of their own. In summary, WoC can provide support for diverse research tasks that would be otherwise out of reach for most researchers. Its focus on global properties of all public source code will enable research that could not be previously done and help to address highly relevant challenges of open source ecosystem sustainability and of risks posed by this global software supply chain. Transforming the WoC prototype into a widely accessible platform is, therefore, our immediate priority.

All source codes can be found in a public repository.[14]

Acknowledgment

This work was supported by the National Science Foundation NSF Award 1633437.

[14]https://github.com/ssc-oscar/Analytics

REFERENCES

[1] R. Di Cosmo and S. Zacchiroli, "Software heritage: Whyandhowtopreservesoftwaresourcecode. ipres 2017," 2017.

[2] A. Mockus, "Engineering big data solutions," in *ICSE'14 FOSE*, 2014. [Online]. Available: papers/BigData.pdf

[3] S. Newman, *Building microservices: designing fine-grained systems*. " O'Reilly Media, Inc.", 2015.

[4] "Nexus repository," https://www.sonatype.com/nexus-repository-oss, accessed: 2019-01-02.

[5] G. Gousios and D. Spinellis, "Ghtorrent: Github's data from a firehose," in *Mining software repositories (msr), 2012 9th ieee working conference on*. IEEE, 2012, pp. 12–21.

[6] G. Gousios, "The ghtorrent dataset and tool suite," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 233–236. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487132

[7] G. Gousios, M. Pinzger, and A. v. Deursen, "An exploratory study of the pull-based software development model," in *Proceedings of the 36th International Conference on Software Engineering*. ACM, 2014, pp. 345–355.

[8] G. Gousios and A. Zaidman, "A dataset for pull-based development research," in *Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM, 2014, pp. 368–371.

[9] G. Gousios, B. Vasilescu, A. Serebrenik, and A. Zaidman, "Lean ghtorrent: Github data on demand," in *Proceedings of the 11th working conference on mining software repositories*. ACM, 2014, pp. 384–387.

[10] J. Howison, M. Conklin, and K. Crowston, "Flossmole: A collaborative repository for floss research data and analyses," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 1, no. 3, pp. 17–26, 2006.

[11] N. M. Tiwari, G. Upadhyaya, and H. Rajan, "Candoia: A platform and ecosystem for mining software repositories tools," in *Proceedings of the 38th International Conference on Software Engineering Companion*. ACM, 2016, pp. 759–764.

[12] N. M. Tiwari, G. Upadhyaya, H. A. Nguyen, and H. Rajan, "Candoia: A platform for building and sharing mining software repositories tools as apps," in *MSR'17: 14th International Conference on Mining Software Repositories*, May 2017.

[13] G. Upadhyaya and H. Rajan, "On accelerating ultra-large-scale mining," in *Proceedings of the 39th International Conference on Software Engineering: New Ideas and Emerging Results Track*. IEEE Press, 2017, pp. 39–42.

[14] ——, "On accelerating source code analysis at massive scale," *IEEE Transactions on Software Engineering*, 2018.

[15] J. Sayyad Shirabad and T. Menzies, "The PROMISE Repository of Software Engineering Databases." School of Information Technology and Engineering, University of Ottawa, Canada, 2005. [Online]. Available: http://promise.site.uottawa.ca/SERepository

[16] B. D. Software, "Black duck open hub," https://www.openhub.net/, accessed: 2018-12-18.

[17] J. Ossher, S. Bajracharya, E. Linstead, P. Baldi, and C. Lopes, "Sourcererdb: An aggregated repository of statically analyzed and cross-linked open source java projects," in *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*. IEEE, 2009, pp. 183–186.

[18] S. Ducasse, T. Gîrba, and O. Nierstrasz, "Moose: an agile reengineering environment," in *ACM SIGSOFT Software engineering notes*, vol. 30, no. 5. ACM, 2005, pp. 99–102.

[19] D. Rozenberg, I. Beschastnikh, F. Kosmale, V. Poser, H. Becker, M. Palyart, and G. C. Murphy, "Comparing repositories visually with repograms," in *Proceedings of the 13th International Conference on Mining Software Repositories*. ACM, 2016, pp. 109–120.

[20] J. Bevan, E. J. Whitehead Jr, S. Kim, and M. Godfrey, "Facilitating software evolution research with kenyon," *ACM SIGSOFT software engineering notes*, vol. 30, no. 5, pp. 177–186, 2005.

[21] S. Bajracharya, J. Ossher, and C. Lopes, "Sourcerer: An infrastructure for large-scale collection and analysis of open-source code," *Science of Computer Programming*, vol. 79, pp. 241–259, 2014.

[22] G. Gousios and D. Spinellis, "Alitheia core: An extensible software quality monitoring platform," in *Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on*. IEEE, 2009, pp. 579–582.

[23] S. Kim, T. Zimmermann, M. Kim, A. E. Hassan, A. Mockus, T. Gîrba, M. Pinzger, E. J. W. Jr., and A. Zeller, "Ta-re: an exchange language for mining software repositories." in *ICSE'06 Workshop on Mining Software Repositories*, Shanghai, China, May 22-23 2006, pp. 22–25. [Online]. Available: http://dl.acm.org/authorize?804411

[24] R. Hackbarth, A. Mockus, J. Palframan, and D. Weiss, "Assessing the state of software in a large enterprise," *Journal of Empirical Software Engineering*, vol. 10, no. 3, pp. 219–249, 2010.

[25] J. Czerwonka, N. Nagappan, W. Schulte, and B. Murphy, "Codemine: Building a software development data analytics platform at microsoft," *IEEE software*, vol. 30, no. 4, pp. 64–71, 2013.

[26] R. Dyer, H. A. Nguyen, H. Rajan, and T. N. Nguyen, "Boa: A language and infrastructure for analyzing ultra-large-scale software repositories," in *Proceedings of the 35th International Conference on Software Engineering*, ser. ICSE'13, 2013, pp. 422–431.

[27] R. Dyer, H. Rajan, and T. N. Nguyen, "Declarative visitors to ease fine-grained source code mining with full history on billions of AST nodes," in *Proceedings of the 12th International Conference on Generative Programming: Concepts & Experiences*, ser. GPCE, 2013, pp. 23–32.

[28] R. Dyer, "Task fusion: Improving utilization of multi-user clusters," in *Proceedings of the 2013 companion publication for conference on Systems, programming, & applications: software for humanity*, ser. SPLASH SRC, 2013, pp. 117–118.

[29] H. Rajan, T. N. Nguyen, R. Dyer, and H. A. Nguyen, "Boa website," http://boa.cs.iastate.edu/, 2015.

[30] R. Dyer, H. A. Nguyen, H. Rajan, and T. N. Nguyen, "Boa: Ultra-large-scale software repository and source-code mining," *ACM Trans. Softw. Eng. Methodol.*, vol. 25, no. 1, pp. 7:1–7:34, 2015.

[31] ——, "Boa: an enabling language and infrastructure for ultra-large scale msr studies," *The Art and Science of Analyzing Software Data*, pp. 593–621, 2015.

[32] A. Mockus, "Amassing and indexing a large sample of version control systems: towards the census of public source code history," in *6th IEEE Working Conference on Mining Software Repositories*, May 16–17 2009. [Online]. Available: papers/amassing.pdf

[33] ——, "Large-scale code reuse in open source software," in *ICSE'07 Intl. Workshop on Emerging Trends in FLOSS Research and Development*, Minneapolis, Minnesota, May 21 2007. [Online]. Available: papers/ossreuse.pdf

[34] C. Bird, P. C. Rigby, E. T. Barr, D. J. Hamilton, D. M. German, and P. Devanbu, "The promises and perils of mining git," in *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*. IEEE, 2009, pp. 1–10.

[35] I. Gorton, A. B. Bener, and A. Mockus, "Software engineering for big data systems," *IEEE Software*, vol. 33, no. 2, pp. 32–35, 2016.

[36] E. Rosch, "Principles of categorization," *Concepts: core readings*, vol. 189, 1999.

[37] S. Agrawal, V. Narasayya, and B. Yang, "Integrating vertical and horizontal partitioning into automated physical database design," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. ACM, 2004, pp. 359–370.

[38] H. Lichter, M. Schneider-Hufschmidt, and H. Zullighoven, "Prototyping in industrial software projects-bridging the gap between theory and practice," *IEEE transactions on software engineering*, vol. 20, no. 11, pp. 825–832, 1994.

[39] R. Budde, K. Kautz, K. Kuhlenkamp, and H. Züllighoven, "Prototyping," in *Prototyping*. Springer, 1992, pp. 33–46.

[40] Y. Ma, T. Dey, J. M. Smith, N. Wilder, and A. Mockus, "Crowdsourcing the discovery of software repositories in an educational environment," *PeerJ Preprints*, vol. 4, p. e2551v1.

[41] S. Chacon and B. Straub, *Pro git*. Apress, 2014.

[42] D. Eastlake 3rd and P. Jones, "Us secure hash algorithm 1 (sha1)," Tech. Rep., 2001.

[43] N. Leavitt, "Will nosql databases live up to their promise?" *Computer*, vol. 43, no. 2, 2010.

[44] D. J. Abadi, "Data management in the cloud: Limitations and opportunities." *IEEE Data Eng. Bull.*, vol. 32, no. 1, pp. 3–12, 2009.

[45] P. Russom *et al.*, "Big data analytics," *TDWI best practices report, fourth quarter*, vol. 19, no. 4, pp. 1–34, 2011.

[46] H. P. Luhn, "A business intelligence system," *IBM Journal of research and development*, vol. 2, no. 4, pp. 314–319, 1958.

[47] A. Moniruzzaman and S. A. Hossain, "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison," *arXiv preprint arXiv:1307.0191*, 2013.

[48] D. German and A. Mockus, "Automating the measurement of open source projects," in *Proceedings of the 3rd workshop on open source software engineering*. University College Cork Cork Ireland, 2003, pp. 63–67.

[49] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 International Workshop on Mining Software Repositories*, ser. MSR '06. New York, NY, USA: ACM, 2006, pp. 137–143. [Online]. Available: http://doi.acm.org/10.1145/1137983.1138016

[50] W. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage," 01 1990.

[51] W. E. Winkler, "Overview of record linkage and current research directions," BUREAU OF THE CENSUS, Tech. Rep., 2006.

[52] Q. Le and T. Mikolov, "Distributed representation of sentences and documents," in *Proceedings of the 31 st International Conference on Machine Learning*, vol. 32. Beijing,China: JMLR, 2014. [Online]. Available: https://cs.stanford.edu/~quocle/paragraph_vector.pdf

[53] S. Amreen, A. Mockus, C. Bogart, Y. Zhang, and R. Zaretzki, "Alfaa: Active learning fingerprint based anti-aliasing for correcting developer identity errors in version control data," *arXiv preprint arXiv:1901.03363*, 2019.

[54] J. Coelho, M. T. Valente, L. L. Silva, and E. Shihab, "Identifying unmaintained projects in github," in *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 2018.

[55] N. Munaiah, S. Kroh, C. Cabrey, and M. Nagappan, "Curating github for engineered software projects," *Empirical Software Engineering*, vol. 22, no. 6, pp. 3219–3253, 2017.