OSS Capstone Project Team Updates

# 09-07-2021

This meeting focused on starting questions regarding the project.

* What are our project objectives?
  + Research and document [source code hosting sites](https://en.wikipedia.org/wiki/Comparison_of_source-code-hosting_facilities)
  + Select one or more sites to focus our efforts
  + Collect available data from this source for network analysis
    - We may also explore diffusion of software packages and dependencies for R and python
  + Create documentation of our process and data
  + Model the network and conduct analysis
    - Specifics are to be determined based on group interest and data availability
* Understand ecosystem:
  + Who are the players, what are the projects, eco system
  + Good understanding about what the data sources are?
  + Source code hosting platforms (include Gitee not on wiki link)
  + As a data source what that platform can allow of us to have
  + Text mining methods to understand the purpose of the project
  + Gizem will send categories for data sources - how many projects; how many developers; what is the way to extract information; are we going to use api, webscrapping (Exclude github)
  + First deliverable - internal eg: spreadsheet/slides
* Data collection:
  + Data collection through webscraping or api
  + Select a source which are impactful
  + Brandon will tell us a format on how to organize it
* What networks can we generate?
  + Granularity
  + Developer n/w, institution, country, enterprise, where the come from, institutions, unit of analysis
  + Tools to visualize
  + Calculate the statistics - look at the structure
* What will be our success metrics?
  + Outputs
    - Good documentation
    - Collected data
    - Network analysis
  + To be further discussed in a few weeks
* Give us an idea of how it was in the past so that we can understand the time commitment and deliverables.
* Does weekly time commitment translate to any deliverables? If yes, how can we work towards creating them?
  + Weekly meetings
* Recommended literature to review
* Develop understanding of the data
* Can we get Rivanna Time
  + yes, they have allocation of Rivanna
* Learning for graph theory (graph fourier transformation)
* Getting started with network analysis

Notes from 9-7-21 Meeting

* OSS developed by researchers is counted as impact of investment in research (compared to publications) but there’s no way to track and quantify the impact of this.
* What are the interactions between developers (contributing to the same projects)
* Impact through dependencies (other software is built on top of yours) sort of like a citation
* Networks between collaborators, networks between software packages and dependencies
* What do these interactions look like and who are the key players, which countries are contributing the most? Which institutions? Which projects are more impactful than others
* Businesses which build on open source projects
* Programming languages which are open source - package manager (CRAN for R, PyPI for python)
* Action items
  + Discuss when we want to meet
  + Ask Abbas if we could flip our time slot
  + Discuss which projects were interested in
  + Divide up the data source platforms to research and document

# 09-16-2021

* -create this doc in one drive to keep track of notes
* -May be look at R packages
* - Add all the hosting facilities - add description
* - why or why not we will pursue it as a data source
* -For BML - see table 2 in the proposal
* -training network and predict and classification can be to predict if there is a link
* -It can also be community related

Our mentor Abbas has requested we outline the following information in our weekly listing of what progress we’ve made:

* What did you do
* What was the result
* What challenges did you run into
* What did you do about it / what was the end result

# 09-23-2021

We met as a group on 09-19-2021 to discuss our progress up to this point and decide on what directions we are interested in exploring for this project.

We’re interested in pursuing python packages and maybe R packages.

Docker containers – is the package usage captured somewhere else? Would it be captured in package downloads? How does docker container usage affect the quantification of OSS usage?

Questions:

* Are we only interested in OSI licensed software?
* Can we get a copy of all the papers referenced in the TGIF PowerPoint
  + Camille wasn’t able to access a copy through UVA’s library of Greenstein, S., and Nagel, F. Digital Dark Matter and the Economic Contribution of Apache,” NBER Working Paper 2013. What is an efficient method to measure citations of OSS?
* Has work already been done to characterize and model package dependencies in R and python?
  + Look at package manager directly
  + Look at the existing github page

Shilpa

* Gathered some information about Phabricator, Gitea, OW2 Consortium, Rosetta code, ProgramOnChain and SEUL.
* Almost all sites except Gitea and Rosetta Code everything needs an account and though the service is free the repositories might not be public.
* Rosetta Code might be a good site to document but will be hard to know the usage and other statistics. May be # of citations can be gathered.
* Reviewed the proposal and pna document and some slides.
* Would like to explore some python and R packages this week.

Camille

* Read the provided presentations and papers of previous and relevant work.
* Conducted preliminary research on Bitbucket, Gforge, Ourproject.org, SourceForge, Assembla, OSDN, Helix TeamHub.
* I’ve found some of the sites have APIs but I’m not always sure what information we can query from them. Some require that an app be made to generate the necessary keys but then I’m not sure if there are restrictions on what projects can be accessed from there.
  + For the Github data collection previous work used the [Ghost.jl repo](https://github.com/uva-bi-sdad/GHOST.jl)

Derek

Nick

* Conducted preliminary research on several of the OSS Sites listed on the excel doc
  + Several of the sites I researched claimed to be able to host public repositories but I was unable to find any of them. Unsure if a direct link to the repository is necessary or if more digging can be done to find them.

Why did we choose the PyPI project?

* Ease of access and interest because we use these products

What is the goal of the project? What analysis / outcomes do we have in mind?

* Explore and catalog open source software projects (must be marked with OSS license)
* Map networks and relationships between repositories, developers, institutions / countries

Measures of impact

* Downloads and contributions by developers
* Citations

For next week:

* Get more information on highlighted repositories and docker
* Start collecting pypi packages
* Meeting with mentor
  + Data summary (downloads, summary statistics, etc.)
  + Review network material with Abbas

Gizem will look for R and python previous work data.

Brandon will prep the materials for the network demo and share with us

# 09-30-2021

Review of networks during meeting

Shilpa

* Pulled gitlab public rep data with licenses and data for the contributors to those projects

Camille

* Created notebook to access Gitea. The notebook pulls the orgs (1,187) and repos (not sure how many total). Couldn’t get a list of users. License info is not listed in API call. If you pull the contents of the repo you can see if it has a license file. However, it appears that the contents are encrypted so you can’t tell which license it is.
* Sample data is in drive
* Gitea has this at the top of their webpage:
* 

Derek

Nick

* Created webscraper for SourceForge and uploaded sample data

To Do

* Nick – pursue contributors, languages, all packages if possible
* Shilpa – to pull all public projects and contributors, try to get more information of activity of contributors
* Split pypi data tasks to collect data
* Can we get contributors from github repos associated with package development?
* Split up assignments between people, get a sample data set, review all columns of data document what we are going to potentially use it for, document what we did along the way and share with Abbas.
  + Document all the steps we take, struggles we run into, how we addressed them, etc. Share with Abbas
* Look for examples of the data products we might want to produce / replicate
* Review network resource Abbas shared with us

# 10-07-2021

Fall break next week, no meeting.

Camille

* Was sick for part of this week and did not accomplish as much as intended
* Got connected to Big Query
* Started a data dictionary of the two Big query PYPI tables
* Reviewed work done by Derek and Shilpa

Derek

Nick

* Added capabilities to scrape contributor data to source forge webscraper

Shilpa

* Gathered name, version, author, license, author\_email info for pypi from big query
* Used that dta to get list of contributors for few popular packages through github api

Challenges:

* Only public contributors or already authorized contributor list is available. To get an entire list we need to get permission from each contributor to get their info
* Same with contributor emails, orgs
* Maybe we should focus on released versions to keep the data clean and easily manageable
* Github api has an api rate limit
* Understanding the api and exploring what is available is taking time

Questions:

* How do downloads, contributors, subscribers translate into node lists? Do we need to collect counts or individual data?

Action items:

* Home\_page - figure out which websites where development is happening
* Continue exploring data and developing data dictionary
* Nick – keep working on SourceForge data
* Look over the presentation about networks see if we have any questions
* We'll review the network visualization software next time

No meeting next week

Meeting with Abbas

* Documenting all of the work we’re doing. Detailing all findings for the semester. The work we did along the way and the decisions we made in our analysis. Share with Abbas before our next meeting in two weeks.
* How can we gather citation data?
  + Google scholar?
  + Start looking into this and see what options we have.
    - Ask Gizem and Brandon
* How do you build a graph with missing data?
* What information will we be incorporating into the graph? Have an answer for next meeting
  + Contributors
  + Packages
  + Dependencies
  + Etc...
* Three weeks have demo data at least to go over with Abbas

# 10-21-21

Camille

* Completed data dictionary for the metadata table
* Researched citations as an impact measure (We won’t be pursuing this further. Instead focusing on number of downloads.)
* Started initial draft of semester documentation requested by our mentor.

Shilpa

* Optimizing code to get the datasets

Derek

Nick

* Discussed limitations of current contributor data and researched options for gathering additional data.
* Additional data appeared difficult to gather and not very substantial after a couple of manual run throughs.
* Pivoted to Pypi side of the project and began catching up with where everyone else was

Notes

* Aggregate downloads by package, country
* Can we get the GitHub data – need to work with Brandon for that
* Ask Brandon about dealing with missing data
* Per Gizem, complete pipeline, build node and edges, then after completing this look to see how we can improve our analysis with dealing with missing data

For next week

* Have data set
* Have summary stats and exploration of data
  + Graphs / visualizations
  + Package downloads
  + Which packages are most popular
  + Analyze the home\_page column to figure out how many packages are hosted on github vs other places.
* Next we start building contributor network, build supernodes as we add layers we group by larger categories (person, organization, country, etc.)
* Also dependency network
* Nick to wrap up the SourceForge exploration, update data and notebooks in OneDrive, also add descriptions to the data fields that have been connected.

# 10-28-21

Camille

* Looked for reference articles for our BML project
* Added our documentation file to the OneDrive

Shilpa

* Finished coding to get both datasets
* Working on optimizing code to pull the data and we are stuck with some roadblocks in converting array string format from google api to python data frame

Derek

Nick

Reach out to Neil to ask about batch jobs. If he can’t help with that then Shilpa has a contact in the Rivanna office to get this set up. Neil is also going to get us proper access to a project folder so we can work together.

What deliverables are for this semester and next semester per SDS, ask Abbas.

# 11-04-2021

Camille

* Worked on the EDA for project data
* Got access to a number of articles through interlibrary loan

Shilpa

* Optimized code with team’s help and worked with Neil to set up slurm scripts as well if needed in the future
* Made Datasets available for EDA
* Worked with Brandon to get contributors data from github commits DB. Have a lot of missing data for commits and need to assess that with Brandon and Gizem.

Derek

Nick

Questions for Gizem and team:

* Is the data we have good to start analysis?
* Seems like there is not complete contributors’ data from gh? Ask Brandon’s input and do we need to pursue it?

What deliverables are for this semester and next semester per SDS, ask Abbas.

* Pgadmin – get access and use this to access the github contributor data
* Going to meet with Brandon this weekend to get setup with pgadmin
* There is code written to create edge lists from our data

Scope

* Contributor network
  + Country aggregation level
  + Edge weight - number of commits per contributor

Data cleaning

* Remove version from the dependencies column
  + This is something we can explore later if we want. However, it would involve a few additional steps to infer versions from the commits data
* Package name and dependencies need to match

EDA

* Scatter chart for number of dependencies, number of downloads
* Map chart of some kind

# 11-18-2021

Camille

* Worked on joining the metadata table with the github data tables. Ran into memory errors. Were able to join the tables directly in the database through the rivanna shell

Shilpa

* Also worked on joining the metadata table with github tables

Derek

Nick

* Worked with dependency edge list

Notes

* We don’t actually need the cost data
* Join tables, group by to create a summary table, Brandon to write us a query
* 2009 – 2019 data range available for contributor data
* Do network/graph analysis in python. Python will be used to do the numerical analysis.
  + Network analysis (R, academics)
  + Graph analysis (python networkx, d3)
  + Exploratory analysis
    - Transitivity
    - Distance
    - Modularity
    - Centrality measures
      * Degree centrality
      * Betweenness centrality (network could be too large for this)
      * Page rank
      * In and out degree centrality (for dependency network)
* First need to know how many nodes and how many edges. It will determine whether we can visualize the network