THE LINUX FOUNDATION
OPEN SOURCE SUMMIT

Discovering Emerging Open Source Communities Through Graphical Analysis

OSS NA | May 12, 2023

Hema Veeradhi

Senior Data Scientist at **Emerging Technology | Red Hat**

hveeradh@redhat.com in hemaveeradhi hemajv California, United States

Oindrilla Chatterjee

Senior Data Scientist at **Emerging Technology | Red Hat**

ochatter@redhat.com in oindrillachatterjee oindrillac Boston, United States

Goals

Early Notification

Using social network analysis techniques, get early notification of new or evolving projects within a vertical or for an organization.

Communities

Track new important emerging open source communities, usergroups, ecosystems and projects

Project Maturity

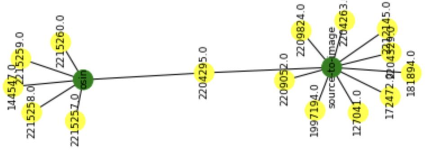
Graphically visualize where a project is situated in relationship to its peers

slido

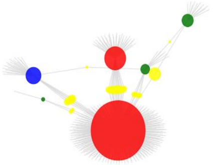
What is your role in open source?

ⓘ Start presenting to display the poll results on this slide.

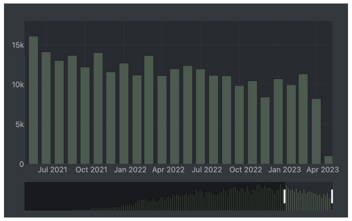
How Network Analysis can help?



Use graphical network representation techniques to depict open source community data.



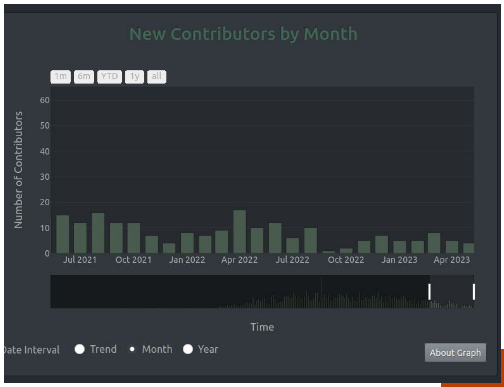
Implement graph centrality algorithms, categorize important nodes in a network and leverage that information to identify important projects and user groups.



Apply above graph centrality techniques on historical GitHub data to track the emergence of important projects

Project Aspen: Data Science

- 8Knot - Dash Dashboard application
  - Cloud-native container deployment strategy
  - Augur Database Curated and Validated on the Backend
  - Python-native data science toolchain
- Rappel
  - Open research
  - Current focus: developer social network analysis of open source ecosystem



https://metrix.chaoss.io : A New, Hosted CHAOSS Software Solution

Project Aspen

- 8Knot - Dash dashboard application
  - Cloud-native container deployment strategy
  - Uses Augur as datasource
  - Python-native data science toolchain
- Rappel
  - Open research
  - Current focus: Developer social network analysis of open source ecosystem

Project Aspen

Community impact

Achieving sustainability

Risk factors can now be measured at many levels.

- Internal project health can still be measured
- Impact on community within the broader ecosystem can now be quantified
- Early detection of risk factors can inform business decisions

Business impact

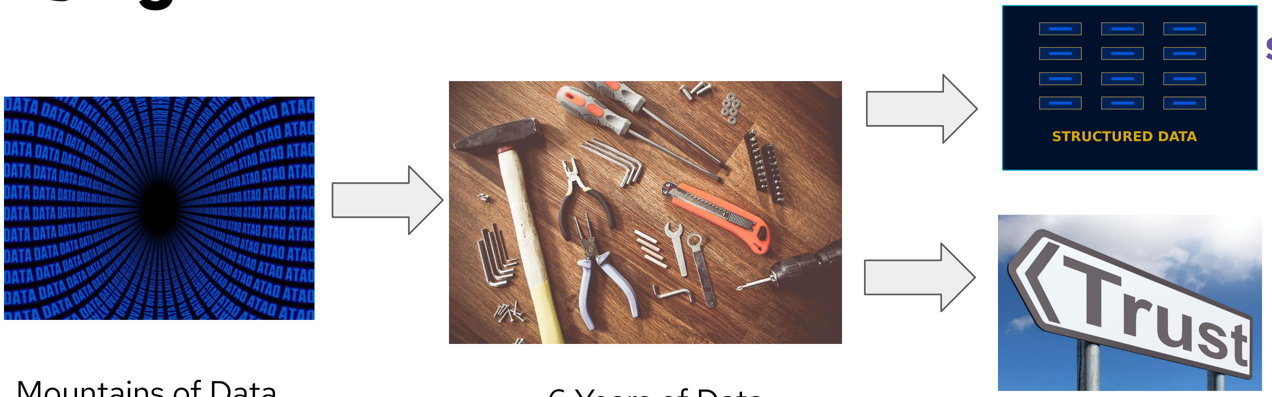
Finding the ever-elusive ROI

All of these tools are well placed to help measure impact of business and community decisions.

- Targeted marketing initiatives can be tracked in a new source
- Resources can be calibrated to community health

Augur: A Path to Data Science

augur



Mountains of Data

6 Years of Data Carpentry

Structured Data

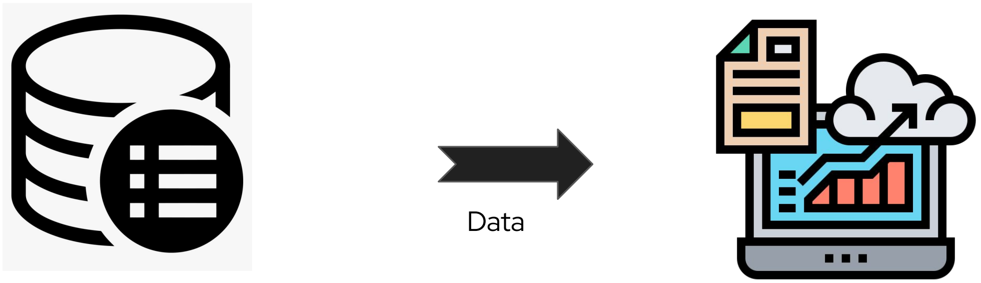
Validated Data

https://metrix.chaoss.io : A New, Hosted CHAOSS Software Solution

Augur

- Open source data collection and organization tool
- Creates database of events in OSS projects (e.g. Github repos)
    - Commits, PRs Issues, Reviews, Contributors, etc.
- Community Health Analytics Open Source Software (CHAOSS) Foundation
- Prof. Sean Goggins, University of Missouri

Augur Database **8Knot Dashboard**

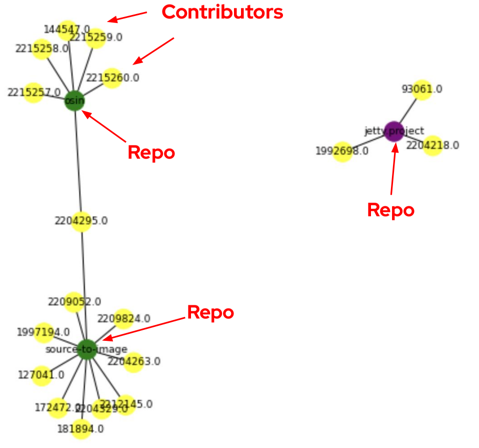


Open source relational database with organized Github data with enforced relationship structure

Dash-Ploty dashboard with the structure to visualize any analysis of the Augur data

Graphical Representation Open Source Ecosystems

Projects and Contributors as Nodes



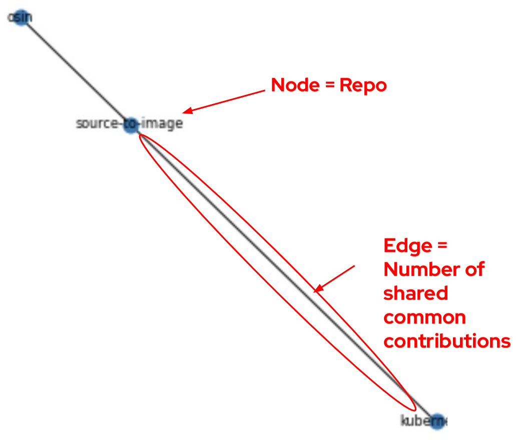
Contributors

Repo

Repo

Repo

Projects as nodes & contributors as edges



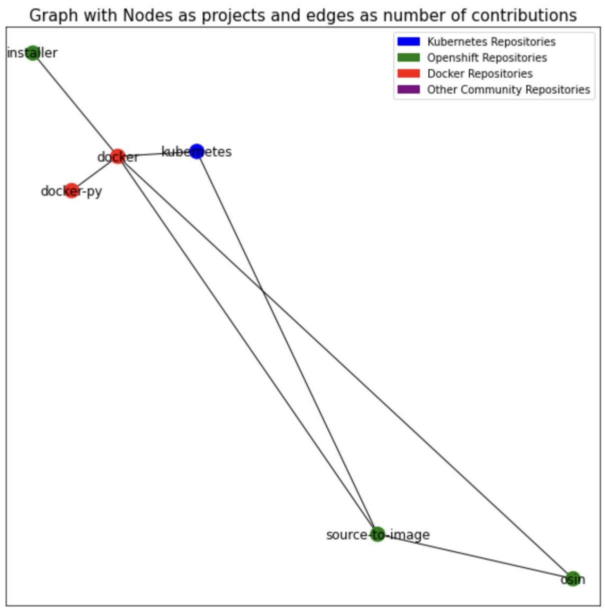
Node = Repo

Edge = Number of shared common contributions

Representing projects as nodes and shared contributions as edges

In this representation style, we connect project repositories by **aggregating the shared activity** between them. We then plot the graph in which the nodes (projects) which have shared contributions are connected to each other and the distance between them represent the frequency of shared contributions between them.

This technique turned out to be an effective way to filter out new repositories which are linked to well-known repositories and identify how closely connected they are.


Graph with Nodes as projects and edges as number of contributions
🟦 Kubernetes Repositories
🟩 Openshift Repositories
🟥 Docker Repositories
🟪 Other Community Repositories
installer
docker
kubernetes
docker-py
source-to-image
osin

What counts as shared activity?

What makes 2 projects connected

- Issues
- PRs
- Commits
- PR Reviews

by the same contributor

Strength of connection (Weights of Edges) is impacted by

Degree of participation

- Maintainer
- Core Contributor
- Developer

To find emerging projects, filter these repos by

- Projects started in the last year
- Number of forks
- Number of Stars
- Growing activity trend

slido

What makes a project rapidly emerging to you?

ⓘ Start presenting to display the poll results on this slide.

slido

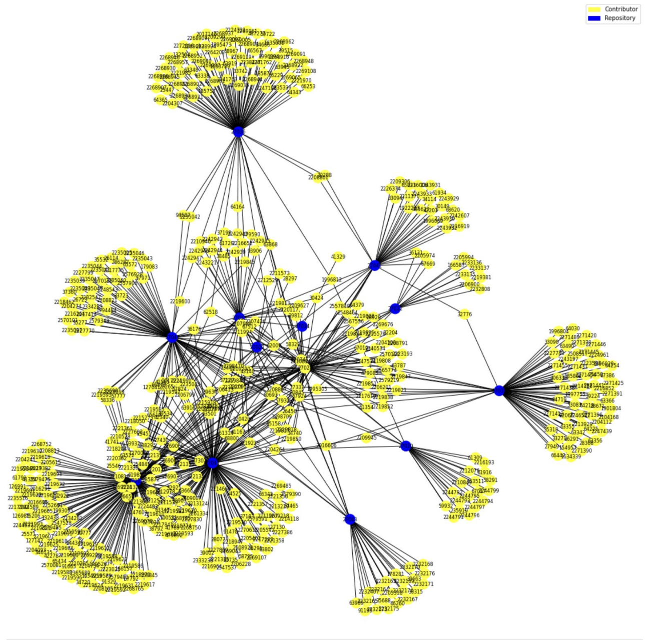
What’s the most important insight that you’re looking for from your community?

ⓘ Start presenting to display the poll results on this slide.

Centrality Algorithms

- Centrality algorithms when applied to graphs with lots of nodes and interconnections can help provide rankings which identifies important nodes.

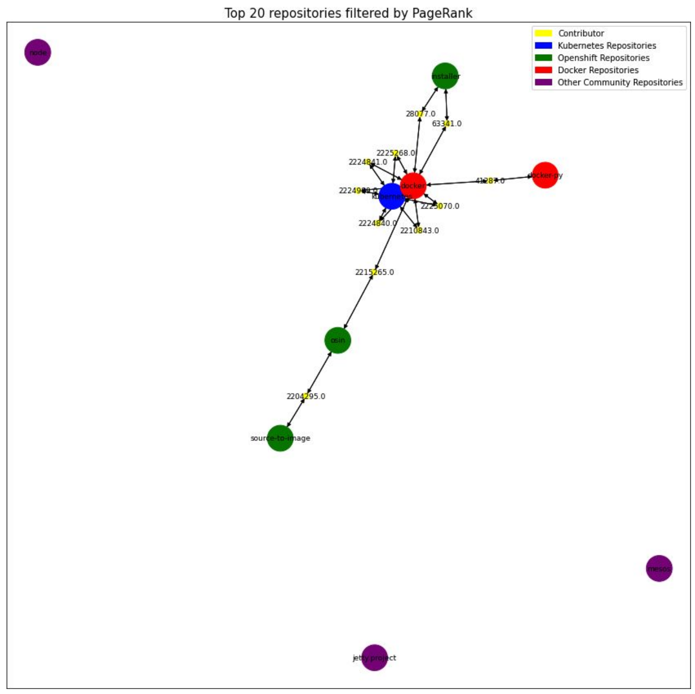
- These are commonly used in identifying most influential user in social networks, key infrastructure nodes in urban networks, etc.



PageRank Algorithm

- PageRank ranks important nodes by analyzing the **quantity and quality of the links** that point to it. In our case, important projects are ranked based on the number of contributors and the quality of the contributor nodes (how well-connected they are)

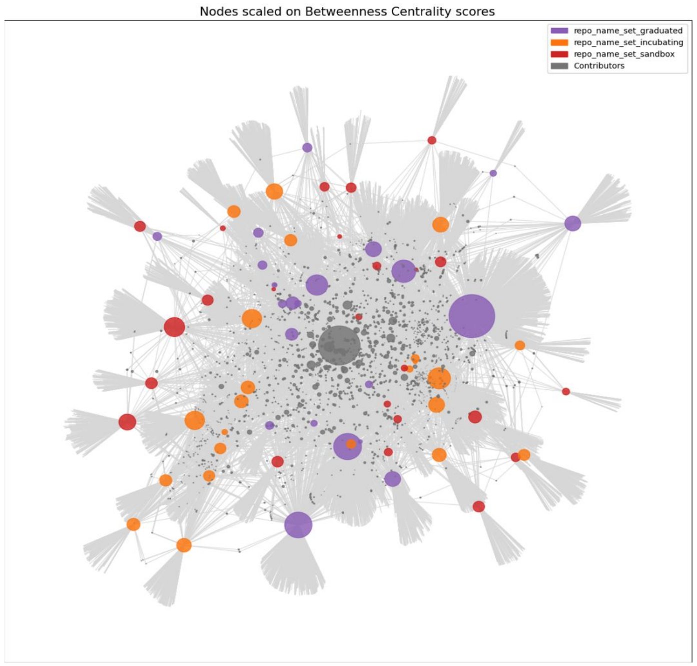
Since PageRank gives importance to the quantity of links pointing to it, if a repository has a lot of contributors and especially if these contributors count as important nodes, they are ranked high. This ends up showing us prominent and well connected project nodes, but is failing to narrow down on important nodes “in relation to” well-known nodes.



Betweenness Centrality

- Betweenness centrality measures the **extent to which a node lies on paths between other nodes in the graph.** Nodes with “higher betweenness” have more influence within a network. Repositories with higher centrality scores can thought to be influential in connection to other repositories in the network.

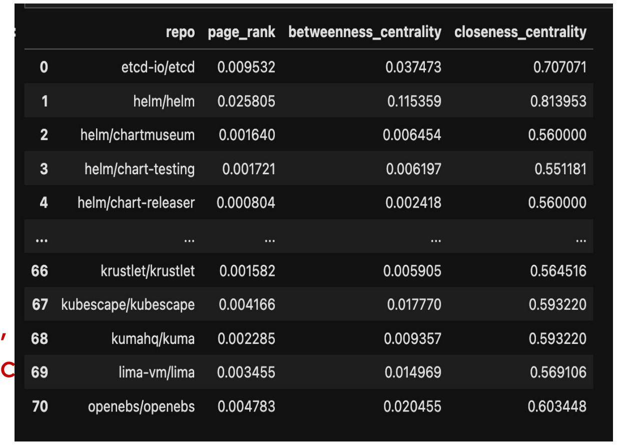
Betweenness centrality highly ranks CNCF “graduated” repos in comparison to repositories in the category of “sandbox”. This is a good metric for us, as using this we are able to better capture relative importance of repositories. In our case since we start with examples of well-known repos, we can use this algorithm to find other repos which are important in connection to these well-known repos.



Closeness Centrality

- Closeness centrality indicates **how close a node is to all other nodes in the network**. It is calculated as the average of the shortest path length from the node to every other node in the network. Nodes with a high centrality scores are best places to influence the entire network most quickly

By calculating the closeness centrality scores for each project, we get a better understanding of the "influence" a given projec can have on the larger Open Source community.



	repo	page_rank	betweenness_centrality	closeness_centrality
0	etcd-io/etcd	0.009532	0.037473	0.707071
1	helm/helm	0.025805	0.115359	0.813953
2	helm/chartmuseum	0.001640	0.006454	0.560000
3	helm/chart-testing	0.001721	0.006197	0.551181
4	helm/chart-releaser	0.000804	0.002418	0.560000
...	...	...	...	...
66	krustlet/krustlet	0.001582	0.005905	0.564516
67	kubescape/kubescape	0.004166	0.017770	0.593220
68	kumahq/kuma	0.002285	0.009357	0.593220
69	lima-vm/lima	0.003455	0.014969	0.569106
70	openebs/openebs	0.004783	0.020455	0.603448

Use Case 1 : Identifying OpenShift as downstream of Kubernetes

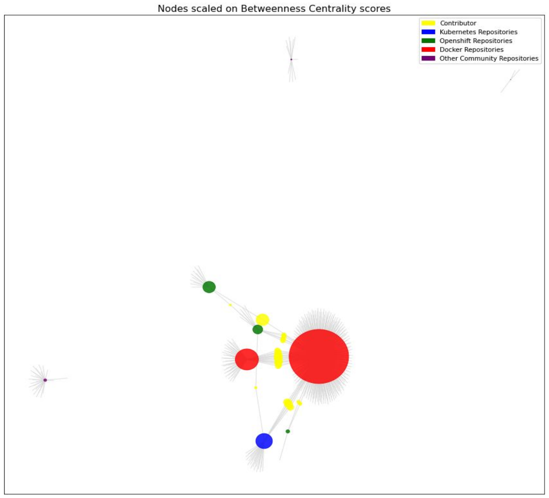
Collected 3 groups of repositories in the **2011-2014** time range - period when OpenShift was becoming popular

- **Well-known projects:** Kubernetes, Docker
- **Emerging projects:** OpenShift
- **Other communities:** Apache Hadoop, Apache Mesos, Node, Eclipse jetty.project

kubernetes
OPENSHIFT

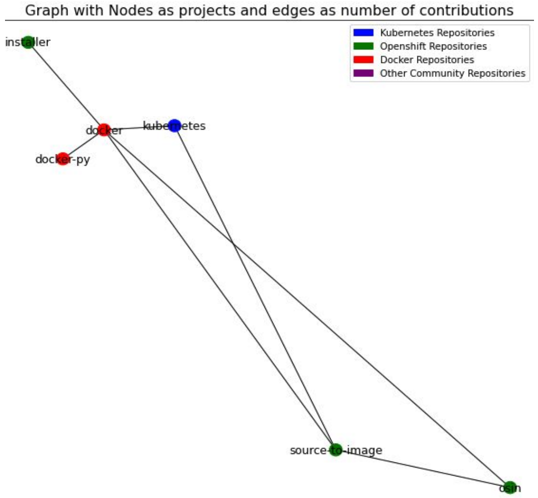
Notebook:
https://github.com/oss-aspen/Rappel/blob/main/notebooks/graph_analysis/approaches/openshift.ipynb

Results of Use Case 1



The size of the nodes in the plot above indicate higher centrality scores. We see that the centrality scores highly rank the docker, kubernetes and openshift repos.

Betweenness Centrality gives us good results and is highly ranking OpenShift repos in comparison to other community repos as this algorithm is able to better capture relative importance of repositories.



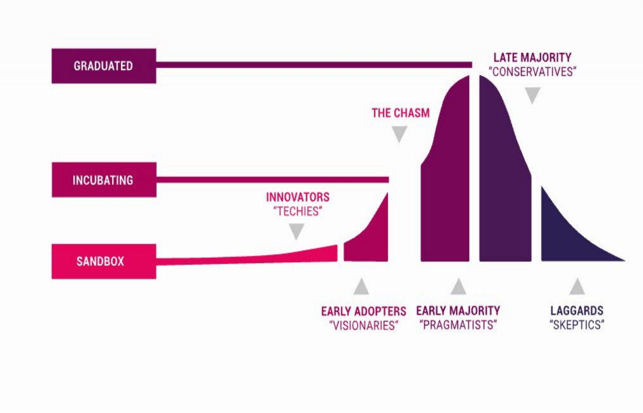
The above graph represents project repositories and how close or far they are to each other based on their degree of connection (number of shared contributions amongst them).

We see that this graph representation effectively filters out the repositories we are most interested in seeing. The repository "closest" to Kubernetes and Docker repositories are 2 OpenShift repositories "installer" followed by "source-to-image" and "osin" and the other unrelated community repositories do not appear on the plot as they are not "connected" to kubernetes

Use Case 2: Representing CNCF Projects

Collected data for the 3 widely categorized CNCF projects in the 2020-2023 time range - https://www.cncf.io/projects/

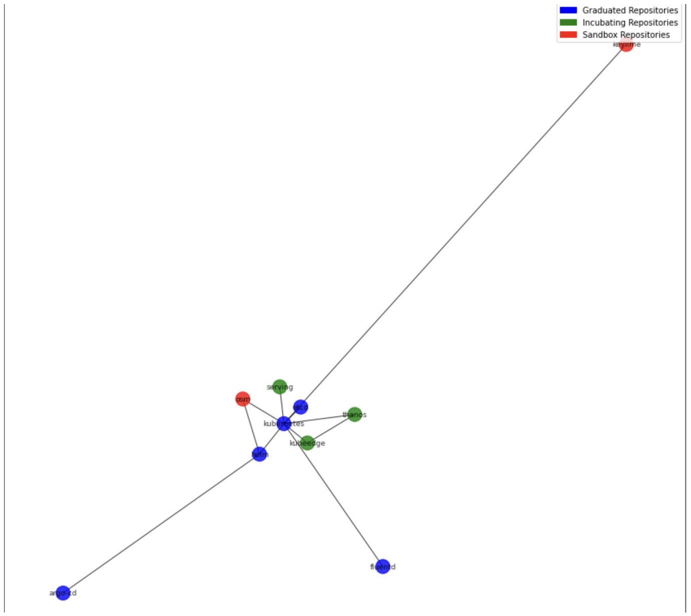
- **Graduated projects**: projects that are considered stable, widely adopted and production ready, attracting thousands of contributors Eg: ***Kubernetes, Prometheus, Helm***
- **Incubating projects**: projects used successfully in production by a small number of users Eg: ***Telemetry, Thanos, Istio***
- **Sandbox projects**: experimental projects not yet widely tested in production on the bleeding edge of technology Eg: ***Keylime, Karmada, Dex***


GRADUATED
INCUBATING
SANDBOX
INNOVATORS "TECHIES"
THE CHASM
LATE MAJORITY "CONSERVATIVES"
EARLY ADOPTERS "VISIONARIES"
EARLY MAJORITY "PRAGMATISTS"
LAGGARDS "SKEPTICS"

Results of Use Case 2

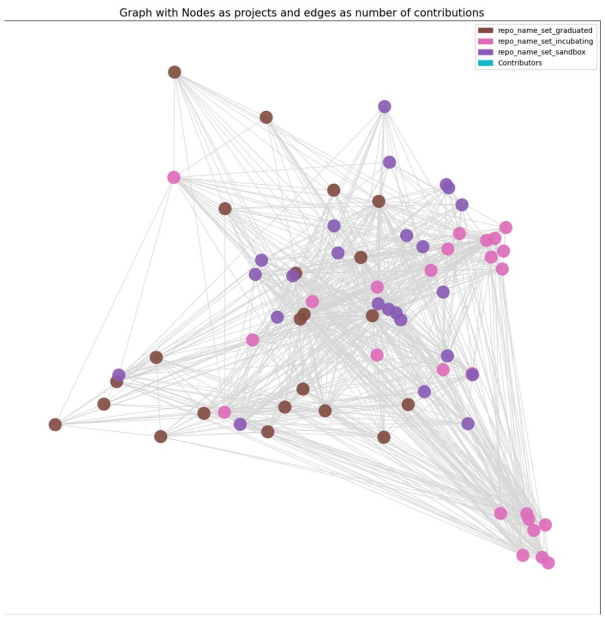
We applied the graph algorithms to the different CNCF projects (**graduated, incubating, sandbox**) to see if they can distinguish and highlight the relevant groups of projects.

We see that the "blue" nodes are more centrally located in the graph compared to the farther "red" nodes. This indicates that the "red" nodes have fewer contributions and is probably a project in its early stages (i.e. sandbox), whereas the "blue" nodes are projects which already have a large number of contributions and is probably more well developed (i.e graduated)



We see that this graph representation effectively filters out the repositories we are most interested in seeing. The repository "closest" to the "graduated projects" are more "incubating projects" and very few "sandbox project". This indicates that the more developed and established projects are central in the graph and the lesser known projects are emerging from it.

The edge lengths and the distance between the nodes can be used to filter out the most connected repos that we are interested in. Thus this graph representation turns out to be an effective way to filter out emerging repositories in relation to already prominent communities.



Notebook:
https://github.com/oss-aspen/Rappel/blob/main/notebooks/graph_analysis/approaches/cncf.ipynb

Top 5 Graduated Projects

Project	Total Score
kubernetes/kubernetes	3
helm/helm	1.49
argoproj/argo-cd	1.36
envoyproxy/envoy	1.09
prometheus/prometheus	1.04

Top 5 Incubating Projects

Project	Total Score
istio/istio	1.08
thanos-io/thanos	0.89
open-telemetry/opentelemetry-collector-contrib	0.70
istio/istio.io	0.66
kubeedge/kubeedge	0.61

Top 5 Sandbox Projects

Project	Total Score
k3s-io/k3s	0.85
strimzi/strimzi-kafka-operator	0.59
openebs/openebs	0.39
karmada-io/karmada	0.37
dexidp/dex	0.35

Ongoing Efforts & Resources

- Look into identifying **influential user groups/contributors** using graphical techniques
- Periodically come up with **lists of potentially important repositories** to feed into Augur and run graph analysis on
- Prototype **reports and visualizations** that can serve as additions to Aspen dashboards.
- **Project Repo:** https://github.com/oss-aspen/Rappel
- **Notebooks:** https://github.com/oss-aspen/Rappel/tree/main/notebooks/graph_analysis

THANK YOU!

Qs?