Uncovering New Open Source Communities Graphically

DevConf CZ | June 17, 2023

Hema Veeradhi

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

hveeradh@redhat.com hemaveeradhi hemajv California, United States

Oindrilla Chatterjee

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

ochatter@redhat.com oindrillachatterjee oindrillac Boston, United States

slido

What is your role in open source?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Goals

Early Notification

Get early notification of new or evolving projects within a vertical or for an organization.

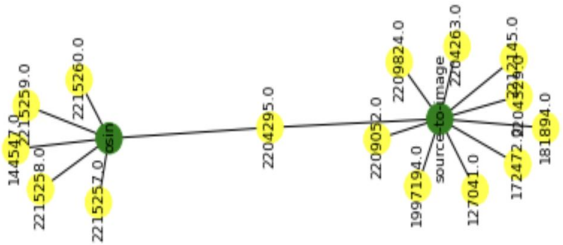
Communities

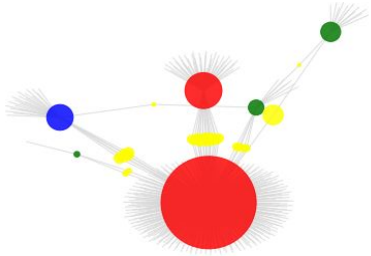
Track important open source communities, usergroups, ecosystems and projects

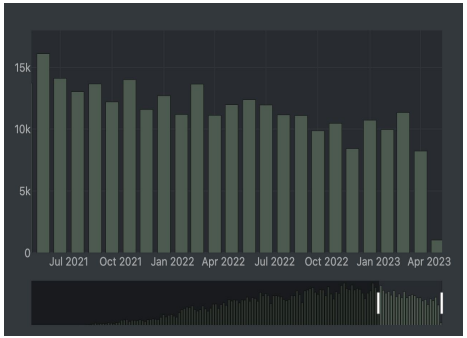
Project Maturity

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

How Network Analysis can help?







Representation

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

Important nodes

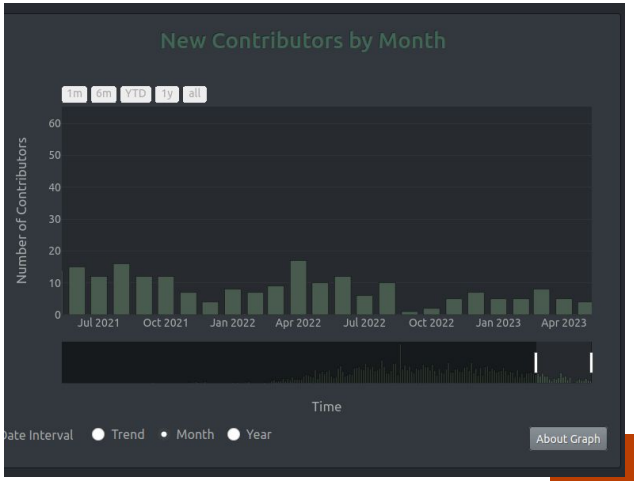
Implement graph centrality algorithms, categorize important nodes in a network and leverage that information to identify important projects and user groups.

Track those over time

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

Project Aspen: Data Science

- **Augur**
  - Open source data collection and organization tool
- **8Knot** – Dash Dashboard application
  - Augur Database Curated and Validated on the Backend
- **Rappel**
  - Open research: developer social network analysis of open source ecosystem

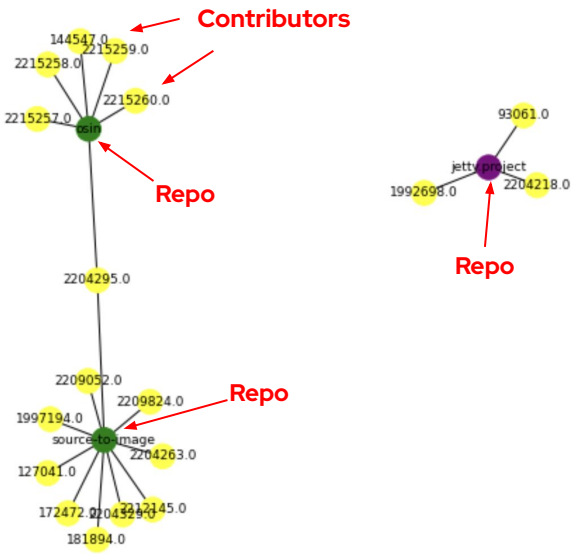


ASPEN

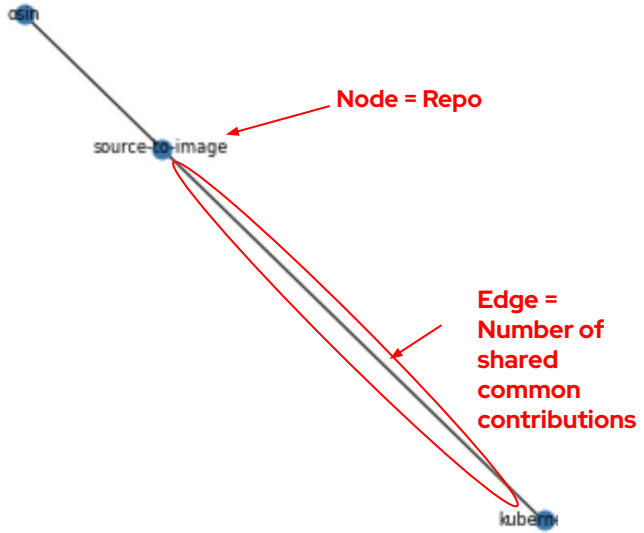
https://metrix.chaoss.io : A New, Hosted CHAOSS Software Solution

Graphical Representation Open Source Ecosystems

Projects and Contributors as Nodes

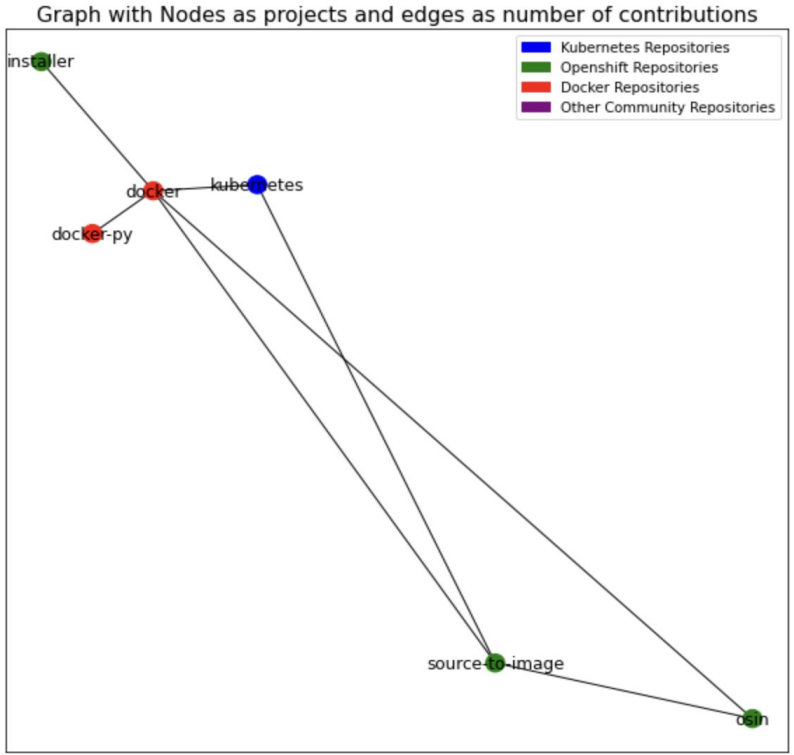


Projects as nodes & contributors as edges



Representing projects as nodes and shared contributions as edges

- In this representation style, we connect project repositories by **aggregating the shared activity** 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.



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?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

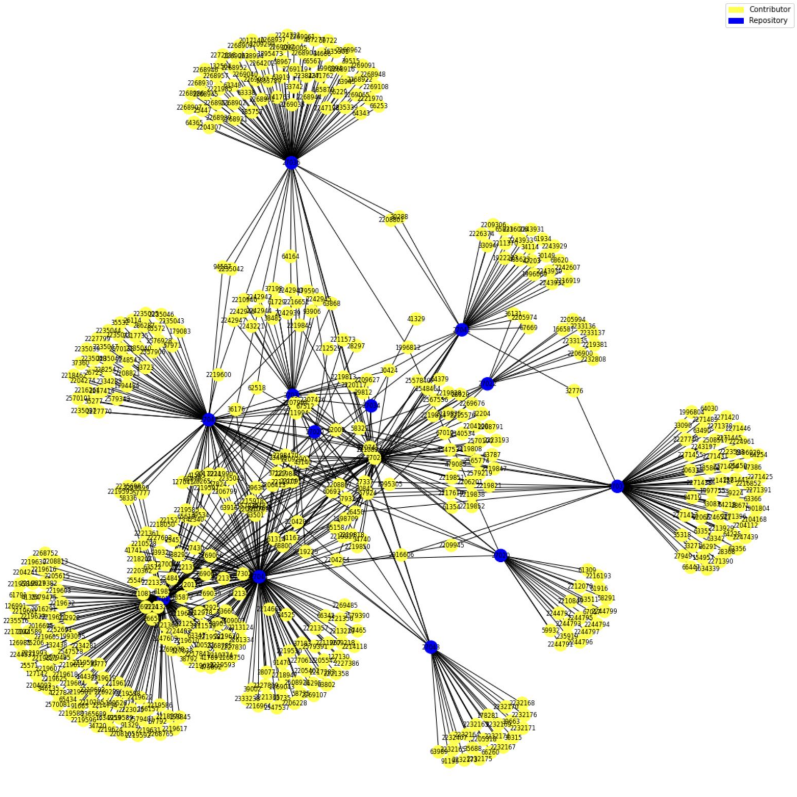
slido

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

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Centrality Algorithms

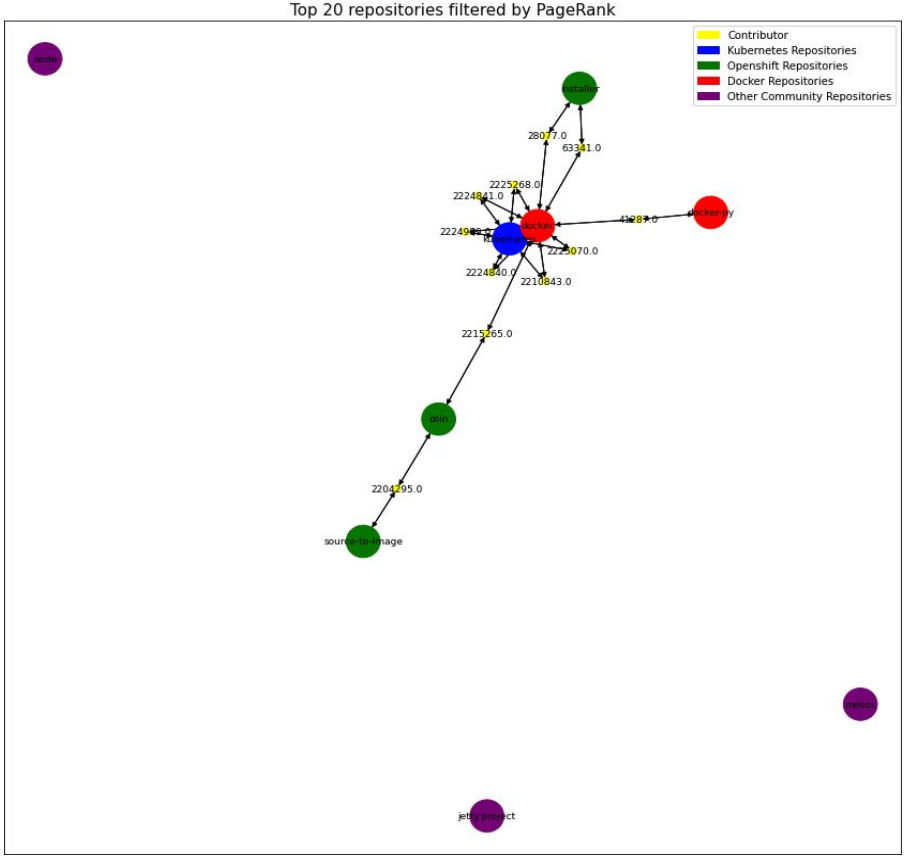
- When applied to graphs with lots of interconnections can provide **rankings** to identify important nodes.
- Commonly used in **identifying most influential users** in social networks, key infrastructure nodes in urban networks, etc.



PageRank Algorithm

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

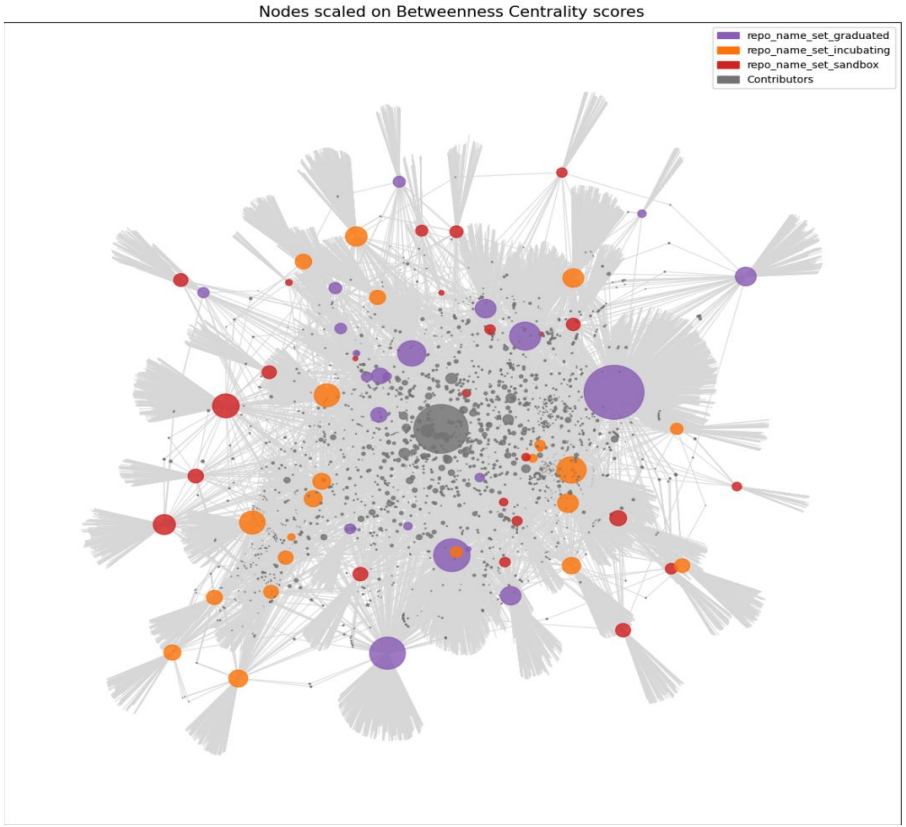
Lower rank => "More important"



Betweenness Centrality

- Betweenness centrality measures the **extent to which a node lies on paths between other nodes in the graph.**

Higher betweenness => "More influential"



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.

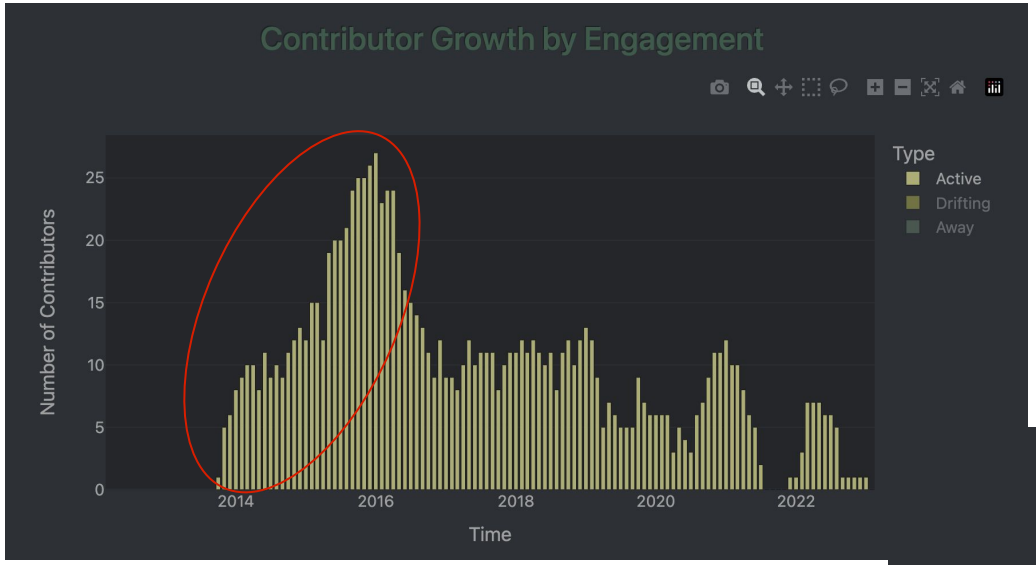
High centrality scores => "Influence the entire network most quickly"

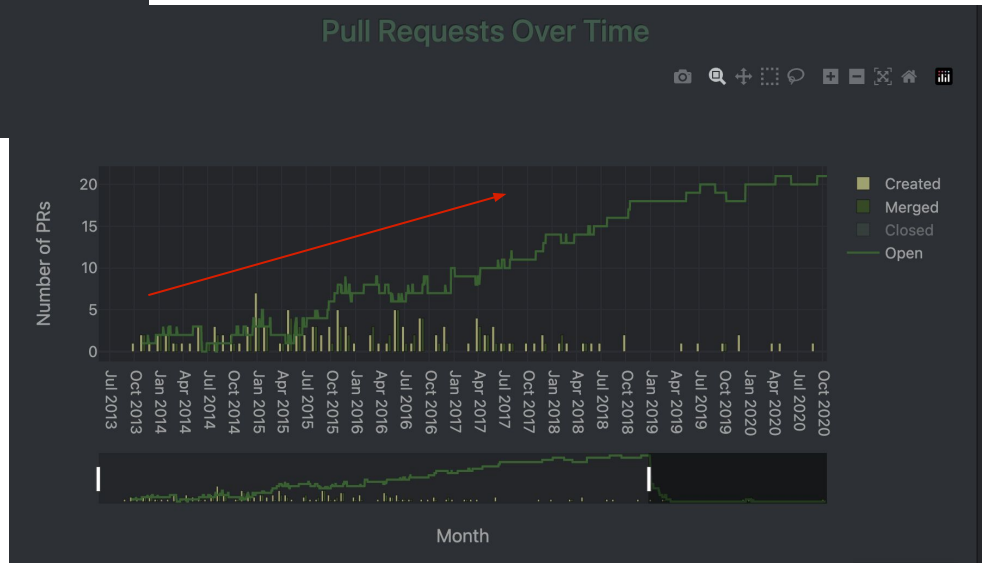
	repo	page_rank	betweenness_centrality	closeness_centrality
0	etcd-io/etcd	0.009532	0.037473	0.707071
1	helm/helm	0.025805	0.115359	0.813953
2	helm/chartmuseum	0.001640	0.006454	0.560000
3	helm/chart-testing	0.001721	0.006197	0.551181
4	helm/chart-releaser	0.000804	0.002418	0.560000
...	...	...	...	...
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

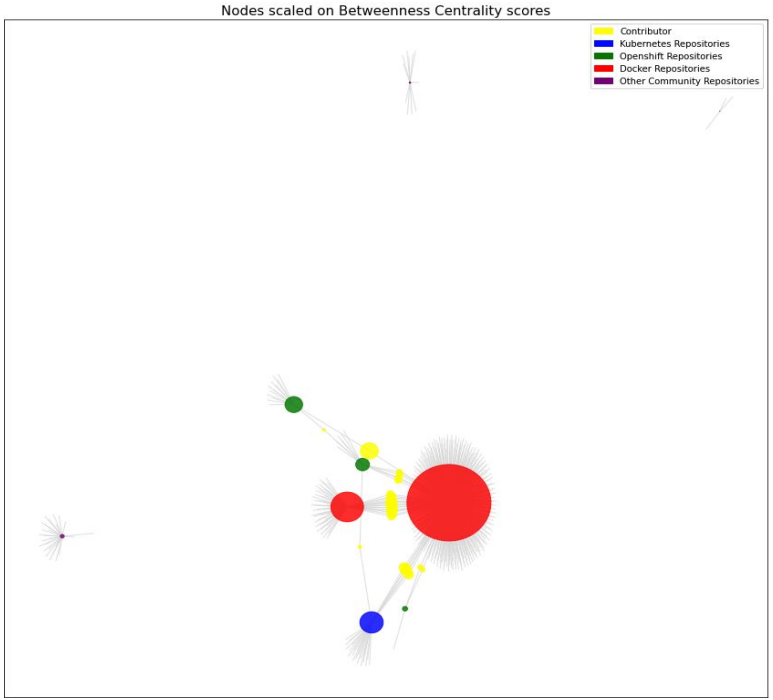
- **Well-known projects:** Kubernetes, Docker
- **Emerging projects:** OpenShift
- **Other communities:** Apache Hadoop, Apache Mesos, Node, Eclipse jetty.project



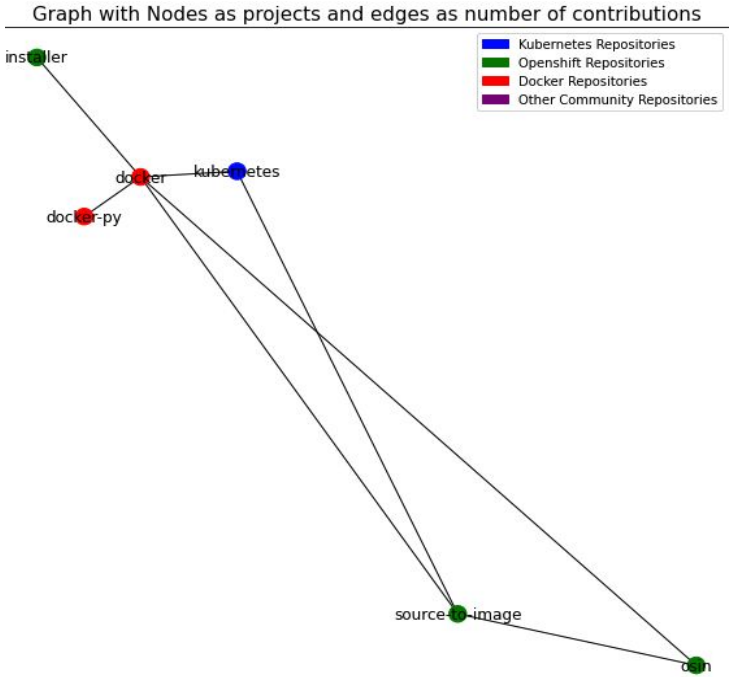


8Knot Dashboard:
https://eightknot.osci.io/overview

Results of Use Case 1



OpenShift repos show higher betweenness scores compared to other projects



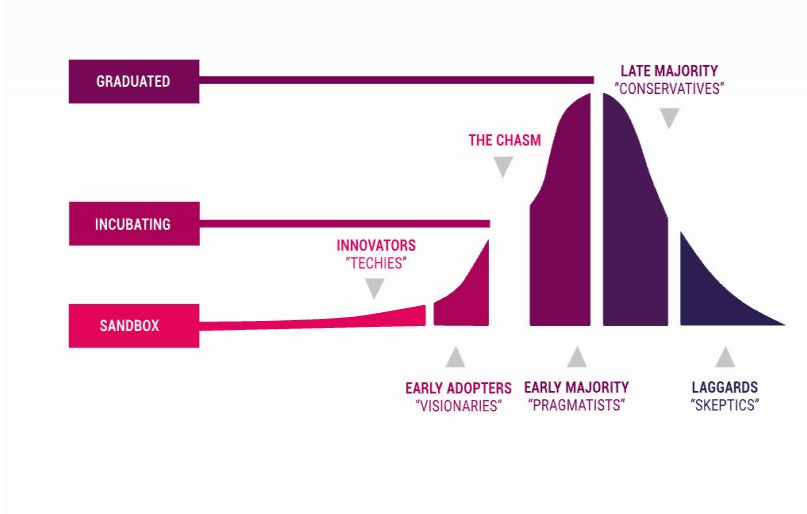
Repos not tightly connected were filtered out and we see closest projects

Use Case 2: Representing CNCF Projects

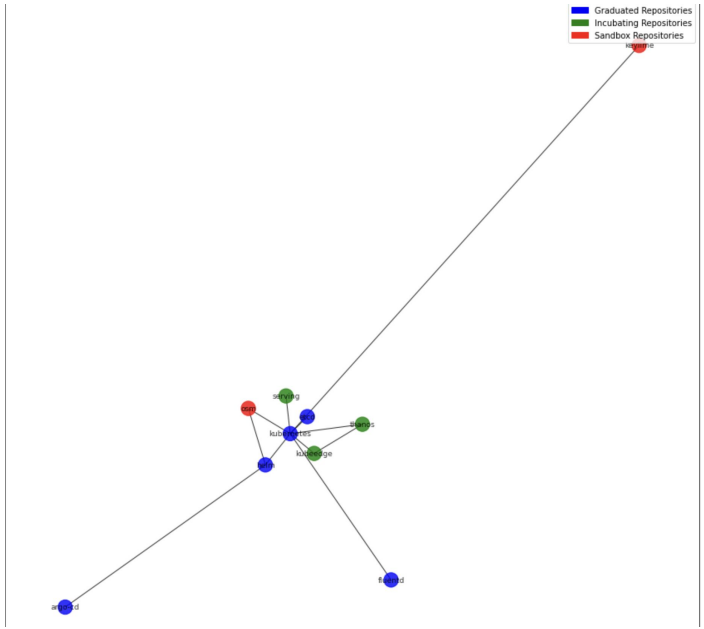
CLOUD NATIVE
COMPUTING FOUNDATION

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

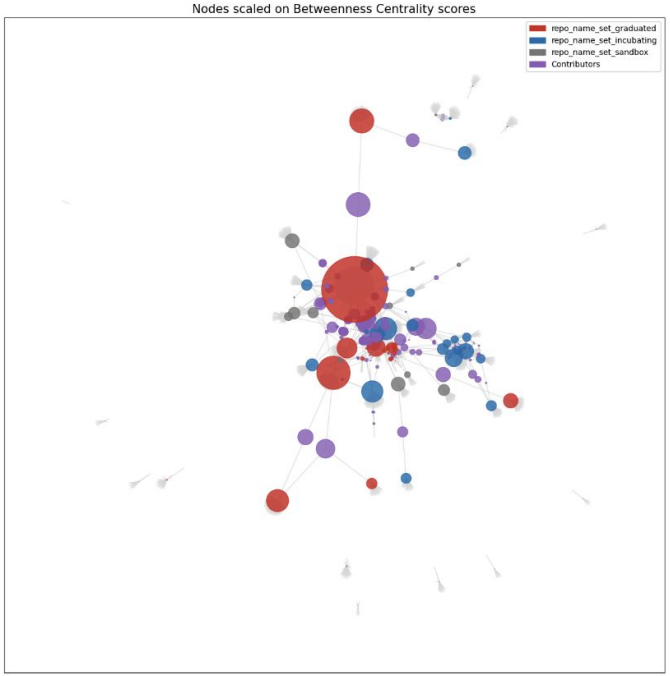
- **Graduated projects:** projects that are considered stable, widely adopted and production ready
- **Incubating projects:** projects used successfully in production by a small number of users
- **Sandbox projects:** experimental projects not yet widely tested in production on the bleeding edge of technology



Results of Use Case 2



Graduated projects more centrally located in comparison to Incubating and Sandbox



In a larger subset of projects, Graduated projects higher ranked in comparison to Incubating and Sandbox

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

*The test set in this experiment does not consist of all CNCF projects

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 network 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?

slido

Audience Q&A Session

ⓘ Click **Present with Slido** or install our Chrome extension to show live Q&A while presenting.

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.

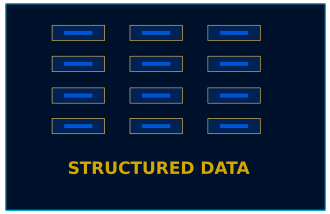
- Targeted marketing initiatives can be tracked in a new source
- Resources can be calibrated to community health

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 Database

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

Data

8Knot Dashboard

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