

Community Stats and Models

Methods:

Say we would like to better understand not only who is participating in our community, but the trends and volume of participation as well. To do this, we can use a number of source datasets and indicators to tell us when something big is brewing in our community. While far from perfect, there are a number of quantitative measures and methodologies one can use to provide insight.

Navigating the Messy World of Open Source Contributor Data

<https://thenewstack.io/navigating-the-messy-world-of-open-source-contributor-data/>

Check out the paper “[Which contributions count? Analysis of attribution in open source](#)” (arxiv: 2013.11007) for research on what is considered an attributable community contribution.

Open Source Guides: Open Source Metrics

<https://opensource.guide/metrics/>

Potential Data Sources for Analysis

Git Issues, Surveys, Forums, Site Analytics, Legal/Finance, Document Reads.

PII (personal identifiable data):

Open source projects and the GDPR (European Privacy Standards)

<https://www.privacypolicies.com/blog/gdpr-open-source-projects/>

Using Open Source resources (Education and Evaluation audience)

<https://merlcenter.org/guides/Dispelling-myths-qualifying-assumptions/>

Review of open source tools (Energy audience)

<https://www.sciencedirect.com/science/article/pii/S1364032121005773>

Github Analytics:

20+ tools to help you mine and analyze GitHub and Git data

<https://livablesoftware.com/tools-mine-analyze-github-git-software-data/>

Export the Github Statistics of your Organization Contributors

<https://beaudry-maxime.medium.com/export-the-github-statistics-of-your-organization-contributors-19a40bbe2784>

Release Downloads (PyPi example: <https://pypistats.org/packages/devolearn/>):

[PyPI page](#)

[Home page](#)

Author: Mayukh Deb, Ujjwal Singh, Bradly Alicea

License:

Summary: Accelerate data driven research in developmental biology with deep learning models

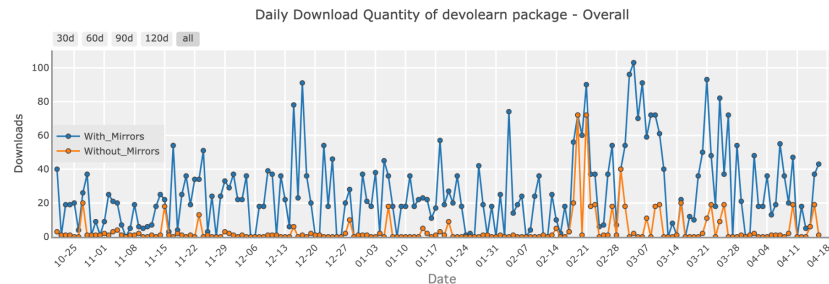
Latest version: 0.3.0

Required dependencies: [cycler](#) | [decorator](#) | [efficientnet-pytorch](#) | [future](#) | [imageio](#) | [imgaug](#) | [imutils](#) | [joblib](#) | [kiwisolver](#) | [matplotlib](#) | [munch](#) | [networkx](#) | [numpy](#) | [opencv-python](#) | [pandas](#) | [pillow](#) | [pretrainedmodels](#) | [pyparsing](#) | [python-dateutil](#) | [pytz](#) | [pywavelets](#) | [scikit-image](#) | [scikit-learn](#) | [scipy](#) | [segmentation-models-pytorch](#) | [six](#) | [sklearn](#) | [threadpoolctl](#) | [tifffile](#) | [timm](#) | [torch](#) | [torchvision](#) | [tqdm](#) | [typing-extensions](#) | [wget](#)

Downloads last day: 0

Downloads last week: 26

Downloads last month: 114



Slack Analytics:

Slack Frontiers 2020: Use analytics to increase Slack adoption and maturity

<https://www.youtube.com/watch?v=zxFKK2mNtXI>

Understand the Data in Your Slack Analytics Dashboard

<https://slack.com/help/articles/360057638533-Understand-the-data-in-your-Slack-analytics-dashboard>

Time-series Analysis:

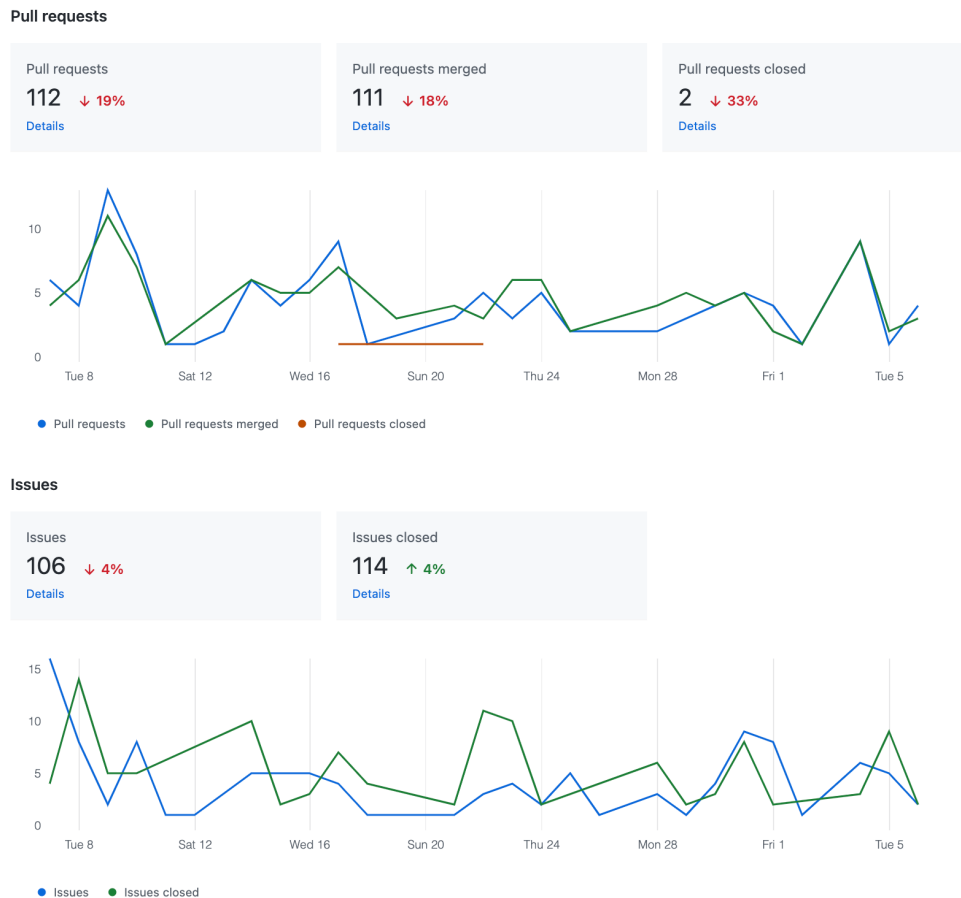
One thing Github (and many blog analytics engines) allows you to do is examine data over specific time intervals. This provides a graph with aggregate data over a temporal series, which can be good for understanding obvious trends, but it looks quite noisy otherwise. Extracting these data from their source can be done either manually or using various command line tools.

Export Github Analytics to csv:

<https://github.com/keplarllp/google-analytics-export-to-csv>

Once the data is extracted, you must first smooth the data to remove noise. Your choice of smoothing technique depends on the source of noise. Is there an automated update that runs on a regular basis? Do you have a lot of periodic fluctuations (e.g. most contributors commit on weekends)? If so, you can adjust the data accordingly. The data can be segmented into overlapping windows to compare between meaningful intervals (days, weeks). You might also use tools such as frequency or autocorrelation analysis to discover trends in the smoothed and segmented dataset. More traditional hypothesis testing is also an option, but remember that

time-series data is dependently distributed (as opposed to the independently distributed assumption of randomly sampled data).



Complex Networks (Rokwire Community example data):

<https://github.com/rokwire/rokwire-community/tree/master/RokComm%20Graph/Version%201>

Complex networks allow us to examine the interactions between community members using select criteria. Each contributor is treated as a node that is connected to other nodes by edges, which are weighted connections that represent specific community attributes.

EXAMPLE: A community of 10 contributors are involved in three activities: documentation, repository maintenance, and discussion leadership. Only a subset of contributors is involved in each activity, but all 10 are involved in at least one of the three. The produces three subgraphs, which connect the contributors in different ways.

Network analysis can be an interesting way to discover subgroups in your community, or to find underutilized interpersonal relationships and affinities to bring contributors together.

Learn more about this topic:

Social Network Analysis and Open-source Communities:

Tamburri DA, Lago P, van Vliet H (2013). Organizational social structures for software engineering. ACM Comput Surv 46(1):3,1–3,35. <https://doi.org/10.1145/2522968.2522971>.

Tamburri DA, Lago P, van Vliet H (2013). Uncovering latent social communities in software development. IEEE Soft 30(1):29–36. <https://doi.org/10.1109/MS.2012.170>.

Community Structure:

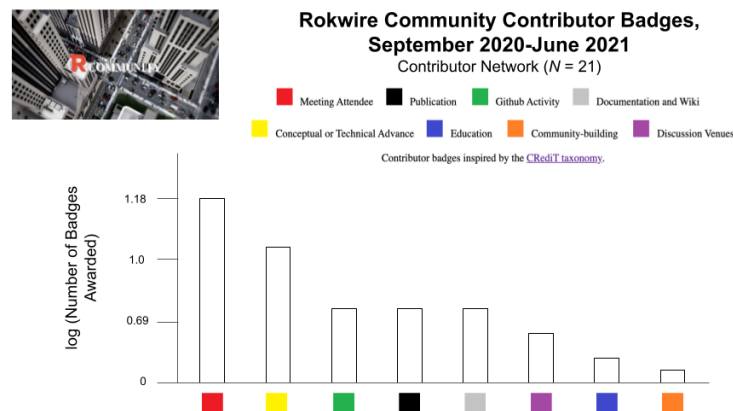
Tamburri, D.A., Palomba, F., Serebrenik, A., and Zaidman, A. (2019). Discovering community patterns in open-source: a systematic approach and its evaluation. Empirical Software Engineering, 24, 1369–1417 <https://doi.org/10.1007/s10664-018-9659-9>.

Social Network Analysis and Applications in Python:

<https://towardsdatascience.com/social-network-analysis-from-theory-to-applications-with-python-d12e9a34c2c7>

NetOpen (workshop at Networks 2021):

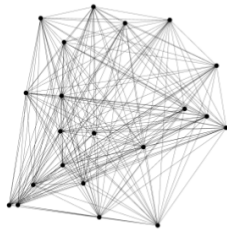
<https://interactiondatalab.com/netopen21/>



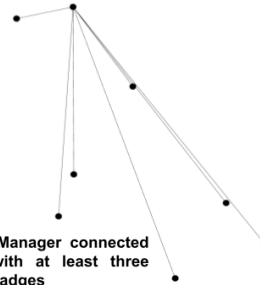


Rokwire Community Contributor Badges Network

Nodes: Unique Community Members,
Connectivity: number of badges in common



Full community network (at least one contributor badge in common).

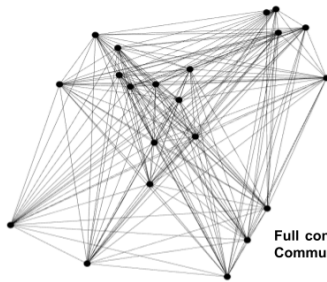


Community Manager connected to people with at least three contributor badges



Rokwire Community Contributor Badges Network

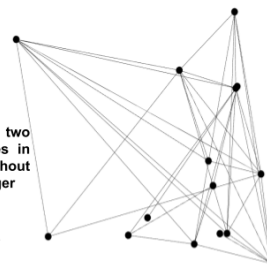
Nodes: Unique Community Members,
Connectivity: number of badges in common



Full contributor network without Community Manager



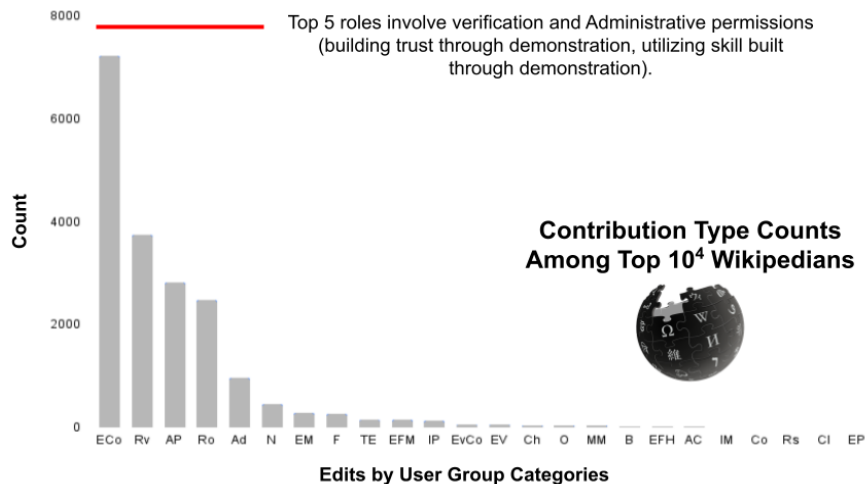
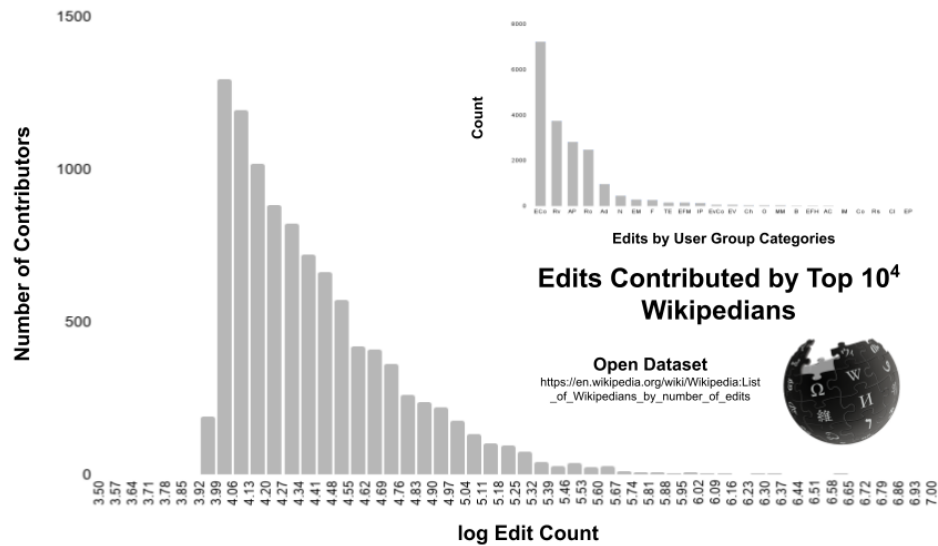
At least two contributor badges in common without Community Manager



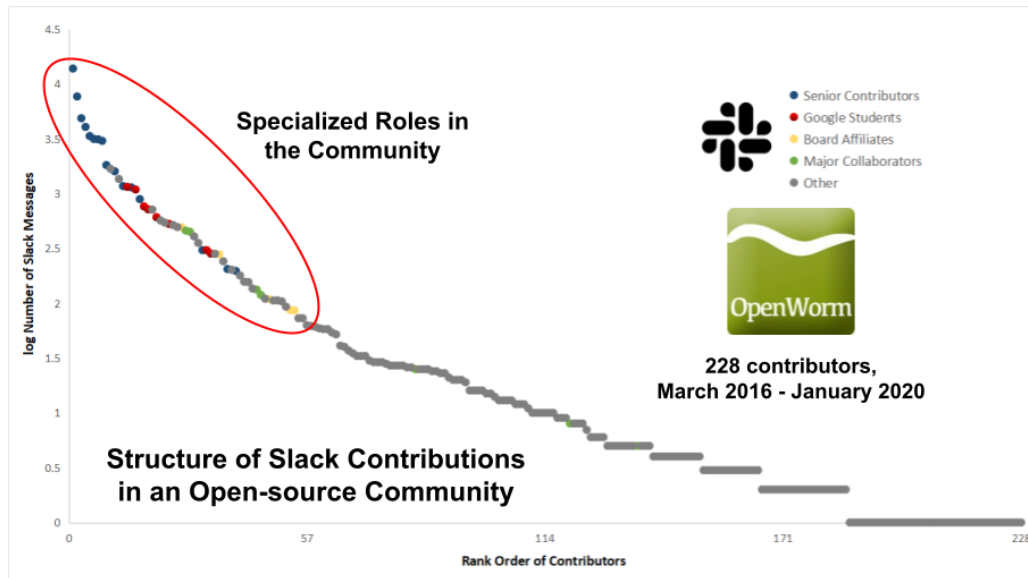
Advanced Interpretation:

Long tail:

Wikipedia (paraphrase): a distribution where the central tendency (region of support) is distinct from the tail, which has many instances occurring over a wide range of values, and are not considered outliers. Such a distribution can be characterized in terms of ranked items (counts) or frequencies of occurrence.



Long tails are properties of an exponential distribution. A nice heuristic is the 80/20 rule, which states that 80% of density in y is represented by 20% of x , while 20% of density in y is represented by 80% of x . This is often used to characterize income distributions (difference between upper and lower income). We can use this same principle to characterize the potential for hierarchical relationships and contribution disparity in our communities. The long tail property can tell us how much of our community is there to participate in a passive manner, or is a potential source of contribution not yet realized.



Mathematical Models and Division of Labor:

One theme that arises from both the Rokwire Community and OpenWorm data is the classification of contributors into skill and experience grades. How do we characterize the potential stable forms of collaboration such a community might take? One way is through the application of agent-based models. [NetLogo](#) has a number of good models for modeling competition and cooperation using computational agents. NetLogo simulations allow for possible scenarios to be generated and understood, and data can be extracted from the simulations.



























Information Theory, Reconfigurability:

One of the great benefits in open source collaboration is the ability to recombine expertise. This is something I have called the “reconfigurability of expertise”. Taking the Rokwire Community badge definitions as an example, we can calculate the information content of the community, as well as for subgroups in the community.

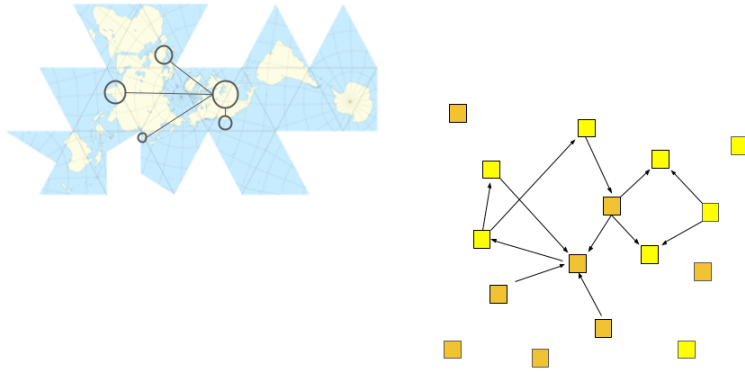
We can measure the information content using Shannon Information (H). Shannon Information is a calculation of the diversity of category types in a given group of contributors. For the Rokwire Community example, if every member of a subgroup is of the same category, the value of H is small, while a broader representation of categories results in a large value for H .

$$H = - \sum p_i * \log p_i$$

Source of badge data: <https://rokwirecommunity.web.illinois.edu/onboarding.htm>

Contributor or Participant (Name or Github handle)	Colored Badges Earned
Kat Asejo	
Peter Ondish	
Robert Belson	
Angela Risius	
Yun Huang	
ortizreyes	 
njo2	 
Krishna Katyal	
Jesse Parent	
Bryan Ranchero	 
Jeff Mahaffa	
Birdal Serbest	 
Kartik Jadhav	 
Chris Nnabuihe	
Kathryn Courtney	
Uros Marjanovic	
Todd Nelson	
Jake Fava	 
Linda Larsen	 

We can also measure the geographic distribution of contributors by time zone. Time zone displacement is important for synchronizing working groups and meeting participation. For example, the difference between a contributor located in Chicago (UTC -5) and a contributor located in Istanbul (UTC +3) is 8 hours. We can also measure the variability in time zone membership using information theory, although in this case, a smaller H value is desirable. Time zone information can be collected via survey or through Slack.



Reconfigurability of Disciplinarity

Optimists and Pessimists

