GitHub Network Analytics

Abhinav Singh
Catherine Yu Miao
Eddie Eustachon
Elaine Wang
Qinpei Zou
Thomas Bruce

Business Problems

Two-fold:

1) Who should one follow in the Machine Learning (ML) & Web Development (WD) fields as a beginner/newbie in either community?

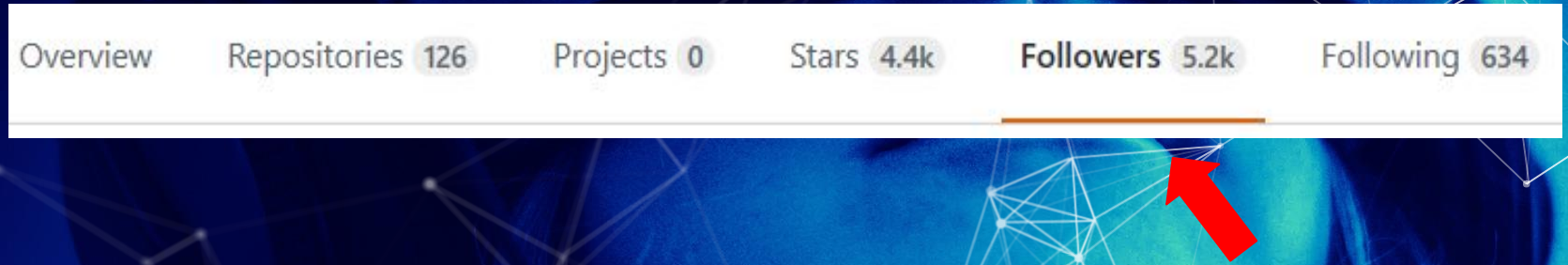
2) How often do the two communities collaborate? What can GitHub do to facilitate greater collaboration?

Data Overview

Description: A large social network of GitHub developers which was collected from the public API in June 2019.

Nodes: are developers who have starred at least 10 repositories

Edges: are mutual follower relationships between them.

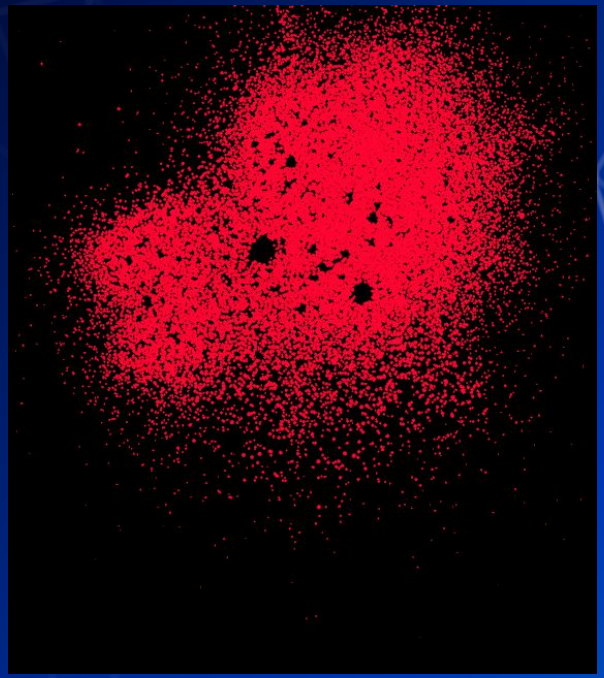


Centrality scores for the top WD & ML members

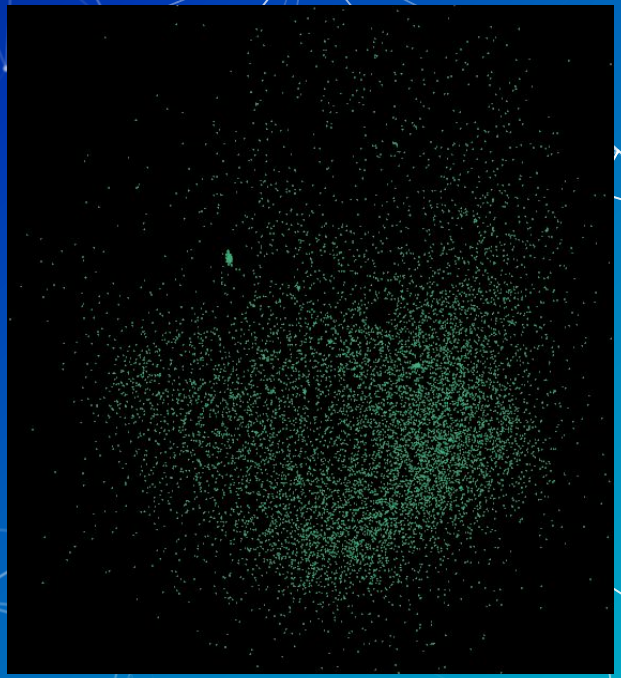
		Degree	Betweenness	Closeness	Centrality
1	Web Developer	0.296	0.293	0.546	1.136
1	Machine Learning Engineer	0.065	0.116	0.331	0.512

Web Developers have higher centrality scores because there are more Web Developers in the dataset.

Networks for ML & WD



Web developer network



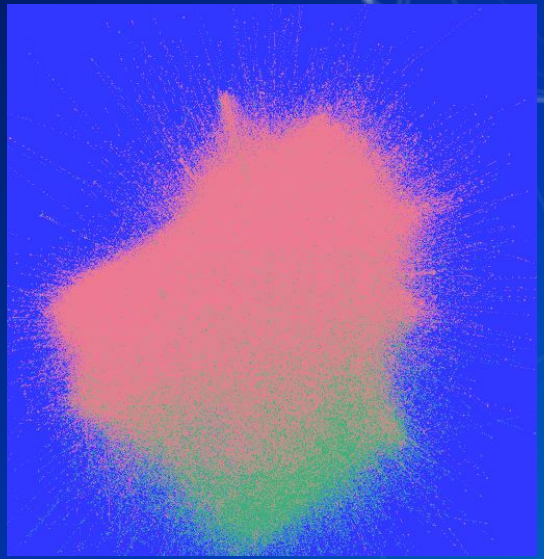
Machine Learning network

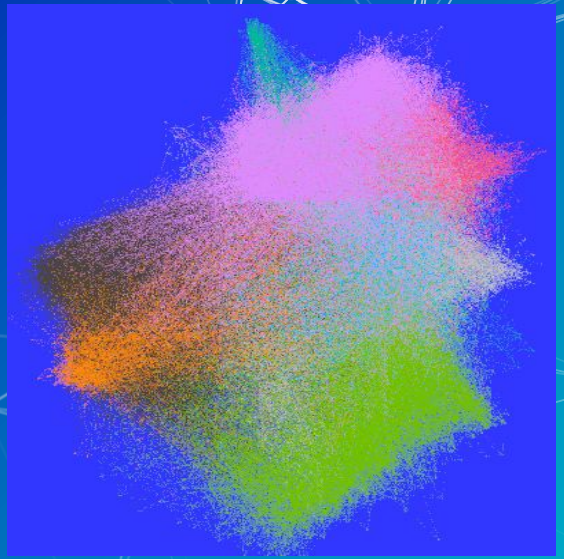
Network for WD + MLE: Nodes

Software Developer

Machine Learning Engineer

Software Developers engage in more diverse activities than Machine Learning Engineer





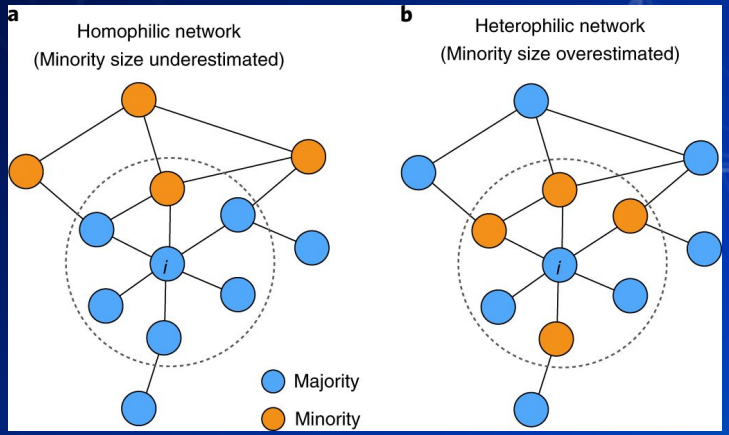
Network for WD + MLE: Edges

Within - Edge

Cross - Edge

Circled part (roughly) indicates the ML community, more green within the community than WD's

Homophily Detection



Expected probability of a cross-edge

0.383

Actual probability of a cross-edge

0.155

Is one group more collaborative?

Proportion of cross-edges initiated by ML	0.485
Proportion of cross-edges initiated by WD	0.515

Most influential ML GitHub accounts - bradfitz

Brad Fitzpatrick
bradfitz

Follow

Xoogler. Ex **@golang** team (2010-2020). My side project is **@perkeep**.

Seattle
brad@danga.com
https://bradfitz.com/

PRO

Overview | Repositories 75 | Projects 0 | Stars 244 | Followers 7.4k | Following 37

Pinned

golang/go
The Go programming language
Go ★ 69.3k 9.9k

perkeep/perkeep
Perkeep (née Camlistore) is your personal storage system for life: a way of storing, syncing, sharing, modelling and backing up content.
Go ★ 4.7k 353

http2 Archived
old repo for HTTP/2 support for Go (see README for new home)
Go ★ 1.7k 138

goimports
(old repo) Tool to fix (add, remove) your Go imports automatically.
Go ★ 964 73

memcached/memcached
memcached development tree
C ★ 9.8k 2.7k

latlong
The latlong package maps from a latitude and longitude to a timezone.
Go ★ 321 32

997 contributions in the last year

bradfitz - iOS, Android software for Smart Home Technology

Most influential ML GitHub accounts - antirez

Salvatore Sanfilippo

antirez

Sponsor

Follow

Computer programmer based in Sicily, Italy. I mostly write OSS software. Born 1977. Not a puritan.

Redis Labs

Campobello di Licata, Sicily, Italy

antirez@gmail.com

Overview | Repositories 62 | Projects 0 | Stars 73 | Followers 13.3k | Following 3

Pinned

redis

Redis is an in-memory database that persists on disk. The data model is key-value, but many different kind of values are supported: Strings, Lists, Sets, Sorted Sets, Hashes, Streams, HyperLogLogs,...

C ★ 41.3k 16.1k

rax

A radix tree implementation in ANSI C

C ★ 591 78

dump1090

Dump1090 is a simple Mode S decoder for RTLSDR devices

C ★ 1.5k 928

linenoise

A small self-contained alternative to readline and libedit

C ★ 2.4k 480

kilo

A text editor in less than 1000 LOC with syntax highlight and search.

C ★ 4.4k 551

load81

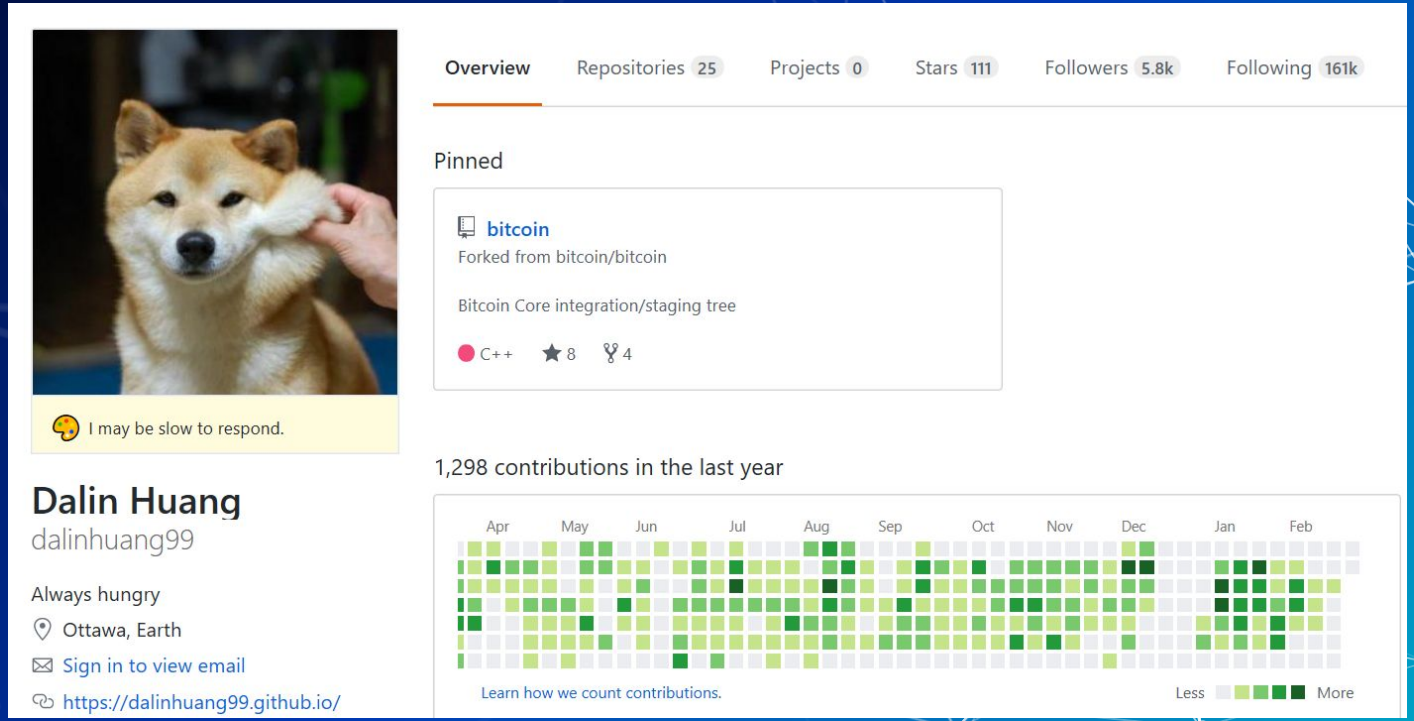
SDL based Lua programming environment for kids similar to Codea

C ★ 462 49

782 contributions in the last year

antirez - Redis DBMS

Most influential WD GitHub accounts - Dalinhuang99



Dalinhuang99 - Tweepy, LeetCode , Blockchain & cryptocurrency Algorithms

Most influential WD GitHub accounts - nfultz

Neal Fultz

nfultz

Sponsor

Computational statistician, programmer and data scientist.

@njnmco

Los Angeles, CA

Sign in to view email

Overview | Repositories 126 | Projects 0 | Stars 4.4k | Followers 5.2k | Following 634

Pinned

nb2mail
Send a notebook as an email
Python ★ 110 10

tron
R package for easy logging
R ★ 36 3

jupyterlab-theme-simple-extension
A simpler theme from a more civilized time
CSS ★ 10 3

AWR.Athena
Short R Wrapper for Athena JDBC connections
R ★ 17 8

grpc
gRPC clients and servers in R
C++ ★ 39 10

stackoverflow
An R package for snippets from stackoverflow
R ★ 13 3

nfultz - computational statistician, R, and Python packages developer

Influential but not cross-collaborative - rasbt

Sebastian Raschka

rasbt

Follow

Machine Learning researcher & open source contributor. Author of "Python Machine Learning." Asst. Prof. of Statistics @ UW-Madison.

UW-Madison

Madison, WI

Overview | Repositories 64 | Projects 0 | Stars 69 | Followers 11.9k | Following 33

Pinned

python-machine-learning-book-3rd-edition

The "Python Machine Learning (3rd edition)" book code repository

Jupyter Notebook ★ 816 ⑂ 266

deeplearning-models

A collection of various deep learning architectures, models, and tips

Jupyter Notebook ★ 11.6k ⑂ 2.7k

biopandas

Working with molecular structures in pandas DataFrames

HTML ★ 238 ⑂ 53

mlxtend

A library of extension and helper modules for Python's data analysis and machine learning libraries.

Python ★ 2.8k ⑂ 605

stat479-machine-learning-fs19

Course material for STAT 479: Machine Learning (FS 2019) taught by Sebastian Raschka at University Wisconsin-Madison

Jupyter Notebook ★ 593 ⑂ 183

stat453-deep-learning-ss20

STAT 453: Intro to Deep Learning @ UW-Madison (Spring 2020)

Jupyter Notebook ★ 264 ⑂ 46

Rasbt - Statistics professor and machine learning researcher

Conclusion & Insights

- Bipartite network confirmed for the two groups of gitHub users - WD and ML
- There is strong evidence of homophily in both communities
- Neither community seems to follow the other more
- Popular members within their own community are not necessarily popular between both communities

Back to the Problem

- 1) Who do we recommend to beginners in a community?
  - Consider deeper dive in community detection efforts
  - Current tag system is comprehensive & well-used
- 2) How do we encourage greater collaboration between communities?
  - Tough for GitHub; both communities look inward typically
  - Super collaborative users are often entrepreneurs; academia is less collaborative

Moving Forward

- Create a suitable recommendation engine
  - Recommend similar topics of ML and WD to encourage collaboration
  - Recommend who to follow based on user data (topic modeling based on collaborative user bios is a possibility)

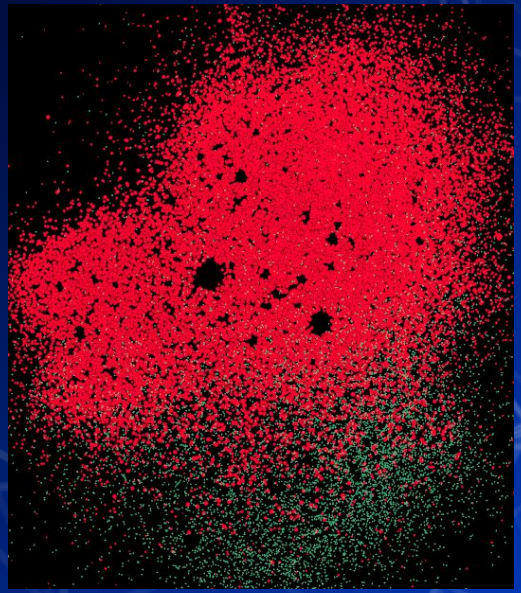
Questions?

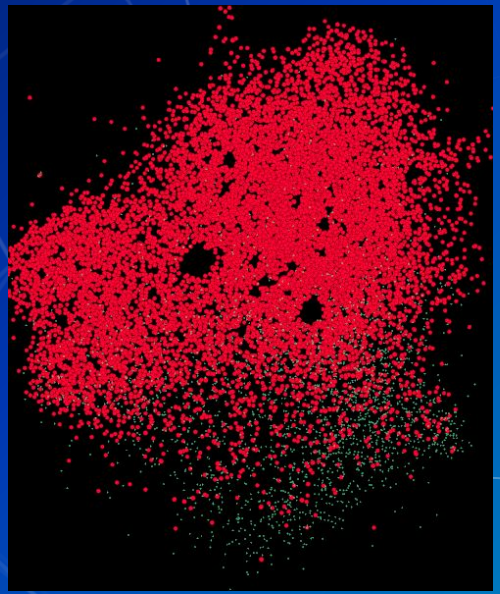
Appendix

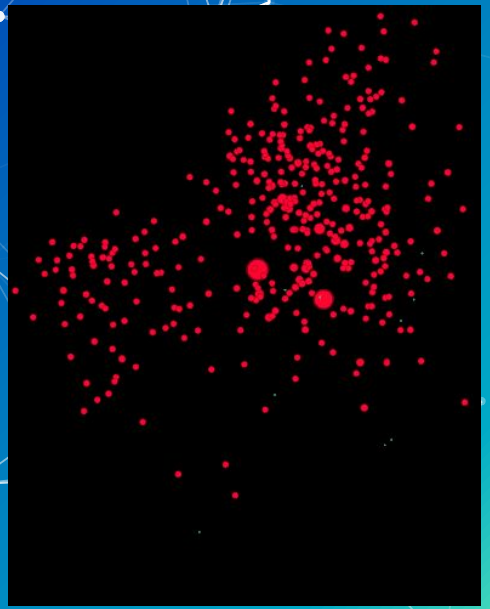
Contents

- Objective: Identify follower networks amongst GitHub developers
- Data Overview
- Network for Web Developer and Machine Learning Engineer
- Cross-Community Network and Homophily Detection
- Conclusion & Insights

WBs are more social connected than MLs







Project Outline

We wanted to see how influential machine learning & web developers are with respect to their own communities & over the entire Github community

- We made two dataframes, one containing just machine learning (ML) & the other with web developers (WD)
- We calculated degreeness, betweenness, & closeness for both these dataframes
- Then we compared these metrics to the degreeness, betweenness & closeness for both dataframes combined to see how influential ML & WD are within their own communities and to the entire community (ML + WD)