Generalists and Specialists
Using Community Embeddings to Quantify
Activity Diversity in Online Platforms

Isaac Waller
walleris@cs.toronto.edu

Ashton Anderson
ashton@cs.toronto.edu

University of Toronto

The Web Conference 2019
Generalists and specialists

full-stack developer vs. React developer
family doctor vs. neurosurgeon
Generalists and specialists

- full-stack developer vs. React developer
- family doctor vs. neurosurgeon
- generalist vs. specialist
Generalists and specialists

vulture

generalist

koala

specialist

Koala photo by DAVID ILIFF. License: CC-BY-SA 3.0. Vulture photo by Charles Sharp. License: CC-BY-SA 4.0
Reddit

Games  MakeupAddiction  medicalschool  soccer

math  programming  Cartalk  chromeos

Construction  funny  television  Aquariums
Which is the specialist?

**User 1:**

\[C = \{\text{China, nba, Buddhism, startrek}\}\]

**User 2:**

\[C = \{\text{Fitness, powerlifting, bodybuilding, weightroom}\}\]
Which is the specialist?

**User 1:**

\[ C = \{ \text{China, nba, Buddhism, startrek} \} \]

**User 2:**

\[ C = \{ \text{Fitness, powerlifting, bodybuilding, weightroom} \} \]

\[ GS(C) = ? \]
Word2vec

Word2vec for communities\textsuperscript{2,3}

**Input:** a (community, user) pair for each comment made in a community

\[(\text{Games}, \text{user1}) \quad (\text{Fitness}, \text{user3}) \quad (\text{medicalschool}, \text{user2}) \]
\[(\text{China}, \text{user4}) \quad (\text{Science}, \text{user2}) \quad (\text{weightlifting}, \text{user3}) \]


Word2vec for communities\textsuperscript{2,3}

**Input:** a (community, user) pair for each comment made in a community

- (Games, user1)  (Fitness, user3)  (medicalschool, user2)
- (China, user4)  (Science, user2)  (weightlifting, user3)

**Output:** a vector for each community in the input, where communities with high user overlap are closer to each other


A first embedding
A first embedding
Word analogies

Male to female

Verb tense
Community analogies

University to city

Sports team to sport / city
4,392 analogies total

brocku → stcatharinesON as uakron → akron
angelsbaseball → baseball as LAClippers → nba
nus → singapore as UMT → missoula
Colts → indianapolis as oaklandraiders → oakland
PolkStateCollege → WinterHaven as csun → LosAngeles
Coyotes → phoenix as AnaheimDucks → LosAngeles
FLC → folsom as OxfordBrookes → oxford
phillies → philadelphia as TorontoBluejays → toronto
Hyperparameter search

cycling + swimming + running = triathlon
72% perfect, 93% top 5
Hyperparameter search

72% perfect, 93% top 5

cycling + swimming + running = triathlon
Our better embedding
Back to generalists and specialists

**User 1:**

\[ C = \{ \text{China, nba, Buddhism, startrek} \} \]

**User 2:**

\[ C = \{ \text{Fitness, powerlifting, bodybuilding, weightroom} \} \]

\[ GS(C) = ? \]
GS-score

\[ \text{GS-score} = 1 | C | \sum_{c \in C} w_c \cos(c, \mu) \]

**generalist**

**specialist**
GS-score

\[ GS(C) = \frac{1}{|C|} \sum_{c \in C} w_c \cos(c, \mu) \]
GS-score

User 1:

$$GS(\{\text{China, nba, Buddhism, startrek}\}) = 0.69$$

24th percentile

User 2:

$$GS(\{\text{Fitness, powerlifting, bodybuilding, weightroom}\}) = 0.89$$

72nd percentile

$$GS(C) = \frac{1}{|C|} \sum_{c \in C} w_c \cos(c, \mu)$$
Data

All comments in 2017
900M comments, 11.4M distinct users
Top 10,000 subreddits by activity

All commits, pull requests, forks, watches, and stars in 2017
413M actions, 8.3M distinct users
Top 40,000 repos by number of stars

Sources: pushshift.io, gharchive.org
Results

Reddit (left) and GitHub (right)
Results

Specialists stay engaged with **communities** longer
Results

Specialists stay engaged with **communities** longer but generalists stay engaged with the **platform** longer.
Results

Specialists stay engaged with **communities** longer but generalists stay engaged with the **platform** longer.
Results

On Reddit, specialists tend to make more exceptional comments.
but generalists are exposed to a more diverse set of users
Can GS-score predict new communities a user joins?
Can GS-score predict new communities a user joins?

![Graph showing the mean average precision for different methods across user GS-score percentiles.]

- **Center-of-mass NN**
- **Collaborative filtering**
- **Popularity**
- **Random**
Community GS-scores
Community GS-scores

![Diagram showing Community GS-scores from 2015 to 2018 for different quartiles. The x-axis represents months (2017-1, 2017-6, 2017-11), and the y-axis represents Community GS-scores ranging from 0.70 to 0.80. Each quartile is color-coded: 1st quartile in blue, 2nd quartile in light green, 3rd quartile in orange, and 4th quartile in pink. The line charts show a slight variation in scores across the years, with the 1st quartile consistently maintaining the highest scores.]
In summary

Users on Reddit and GitHub range from generalist to specialist

On Reddit, specialists are more likely to make exceptional comments

Specialists stay engaged with individual communities longer, but generalists stay engaged with the platform longer

Specialists are significantly more predictable than generalists
What does the universe of subreddits look like?

There's a community for almost anything on Reddit. What do these constellations look like? And where are you located within them?

We mapped Reddit to answer this question and find out how users exist in this space. Here's what we found...

A project of the Computational Social Science lab at the University of Toronto (@isaacwalter and @ashton1anderson).

Scroll down to continue

Thank you! tiny.cc/gsscore