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

Computer Science
UNIVERSITY OF TORONTO

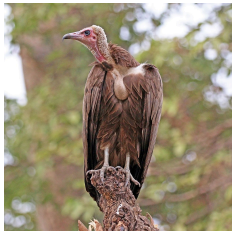
Generalists and specialists

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

Reddit

Games	MakeupAddiction	medicalschool	soccer
math	programming	Cartalk	chromeos
Construction	funny	television	Aquariums

Which is the specialist?

User 1:

$$C = \{\text{China}, \text{nba}, \text{Buddhism}, \text{startrek}\}$$

User 2:

$$C = \{\text{Fitness}, \text{powerlifting}, \text{bodybuilding}, \text{weightroom}\}$$

Which is the specialist?

User 1:

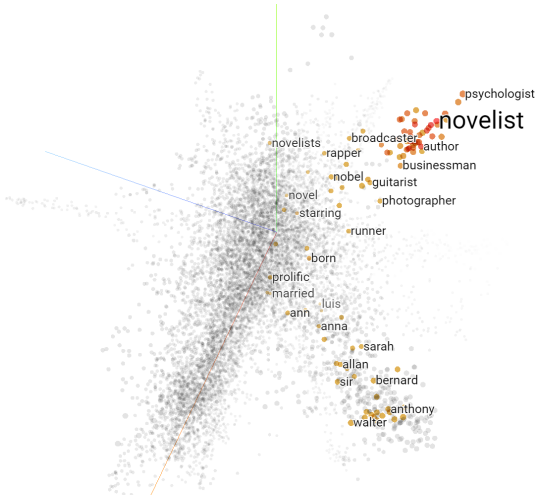
$$C = \{\text{China}, \text{nba}, \text{Buddhism}, \text{startrek}\}$$

User 2:

$$C = \{\text{Fitness}, \text{powerlifting}, \text{bodybuilding}, \text{weightroom}\}$$

$$GS(C) = \ ?$$

Word2vec[1]



[1] Mikolov et al. (2013) Distributed Representations of Words and Phrases and their Compositionality

Word2vec for communities[2,3]

Input: a $(\text{community}, \text{user})$ pair for each comment made in a community

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

[2] Kumar et al. (2018) Community Interaction and Conflict on the Web

[3] Martin (2017) community2vec: Vector representations of online communities encode semantic relationships

Word2vec for communities[2,3]

Input: a $(\text{community}, \text{user})$ pair for each comment made in a community

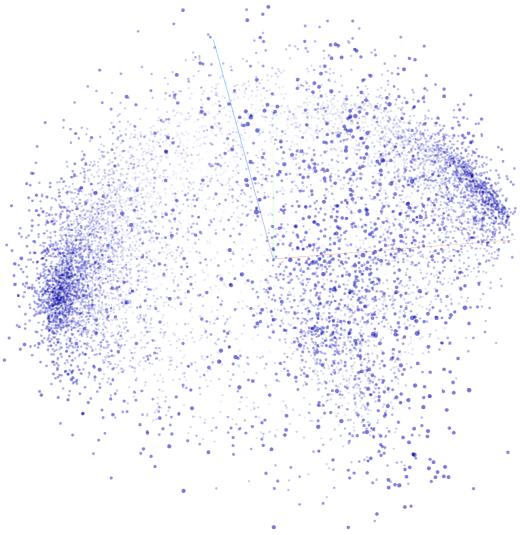
$$(\text{Games}, \text{user1}) \quad (\text{Fitness}, \text{user3}) \quad (\text{medicalschool}, \text{user2})$$
$$(\text{China}, \text{user4}) \quad (\text{Science}, \text{user2}) \quad (\text{weightlifting}, \text{user3})$$

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

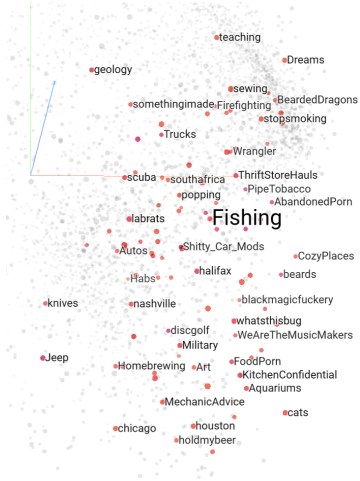
[2] Kumar et al. (2018) Community Interaction and Conflict on the Web

[3] Martin (2017) community2vec: Vector representations of online communities encode semantic relationships

A first embedding



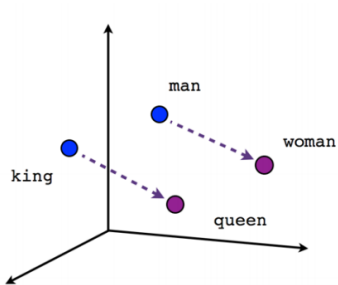
A first embedding



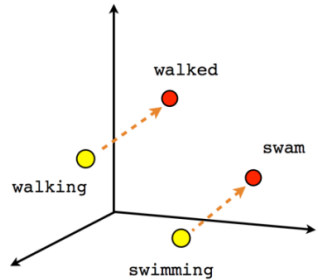
Nearest points in the original space:

discgolf	0.168
Shitty_Car_Mods	0.191
PipeTobacco	0.197
Trucks	0.205
surfing	0.206
flyfishing	0.217
DippingTobacco	0.221
Jeep	0.222
itookapicture	0.236
AbandonedPorn	0.237
beards	0.254
FoodPorn	0.259
halifax	0.261
ThriftStoreHauls	0.265
KitchenConfidential	0.265
labrats	0.270

Word analogies

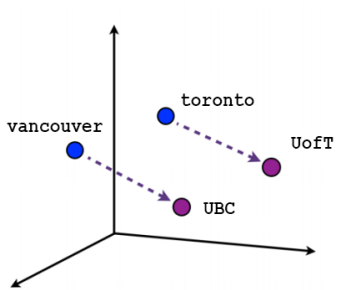


Male to female

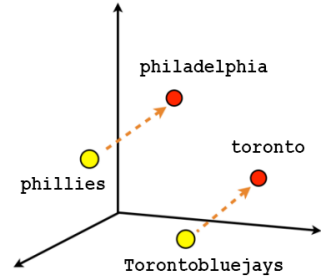


Verb tense

Community analogies



University to city

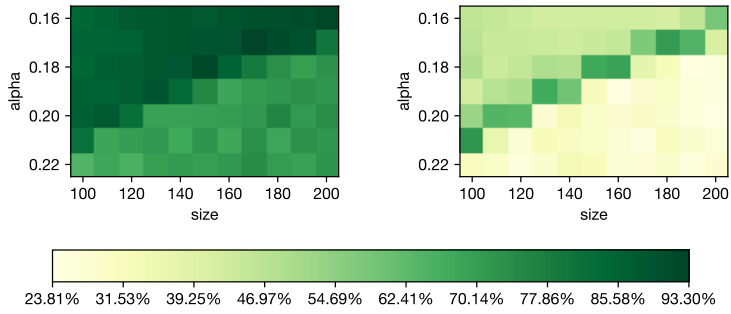


Sports team to sport / city

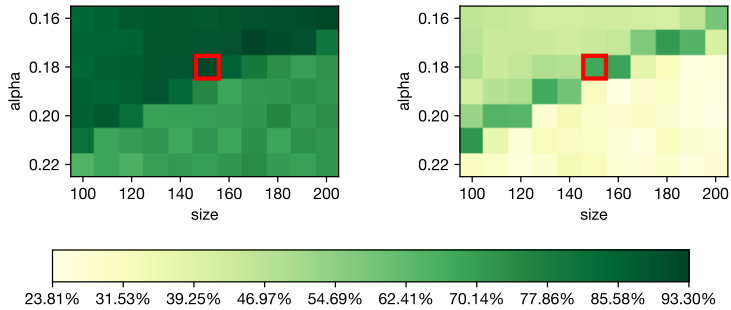
4,392 analogies total

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

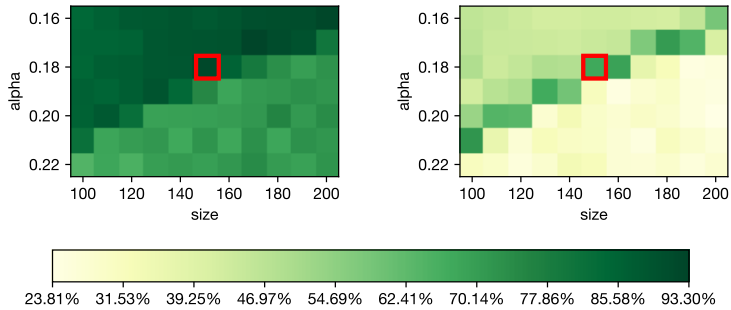


Hyperparameter search



72% perfect, 93% top 5

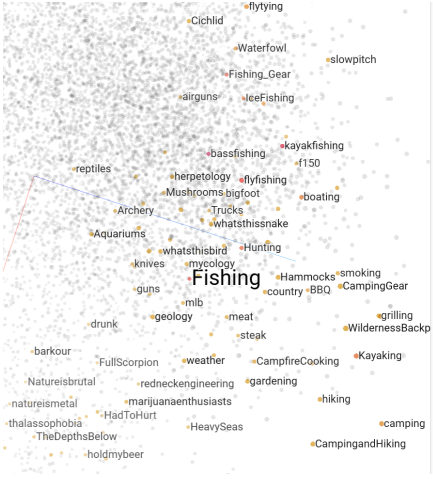
Hyperparameter search



72% perfect, 93% top 5

cycling + swimming + running = triathalon

Our better embedding



Nearest points in the original space:

bassfishing	0.226
kayakfishing	0.260
flyfishing	0.325
Fishing_Gear	0.357
Hunting	0.403
IceFishing	0.440
Kayaking	0.457
boating	0.478
Spearfishing	0.533
flytying	0.542
camping	0.543
Waterfowl	0.548
BBQ	0.562
Outdoors	0.579
bowhunting	0.581
hiking	0.584

Back to generalists and specialists

User 1:

$$C = \{\text{China}, \text{nba}, \text{Buddhism}, \text{startrek}\}$$

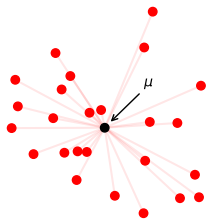
User 2:

$$C = \{\text{Fitness}, \text{powerlifting}, \text{bodybuilding}, \text{weightroom}\}$$

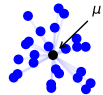
$$GS(C) = \ ?$$

GS-score

generalist

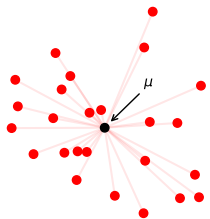


specialist

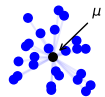


GS-score

generalist



specialist



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

GS-score

User 1:

$$GS(\{\text{China}, \text{nba}, \text{Buddhism}, \text{startrek}\}) = \underset{\textbf{24}^{\textbf{th}}\ \textbf{percentile}}{0.69}$$

User 2:

$$GS(\{\text{Fitness}, \text{powerlifting}, \text{bodybuilding}, \text{weightroom}\}) = \underset{\textbf{72}^{\textbf{nd}}\ \textbf{percentile}}{0.89}$$

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

Data

All comments in 2017

All commits, pull requests, forks, watches, and stars in 2017

900M comments, 11.4M distinct users

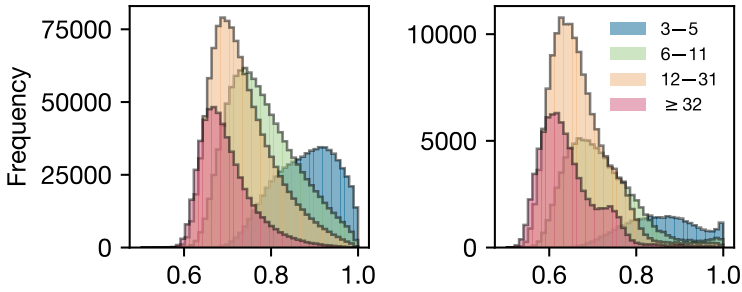
413M actions, 8.3M distinct users

Top 10,000 subreddits by activity

Top 40,000 repos by number of stars

Sources: pushshift.io, gharchive.org

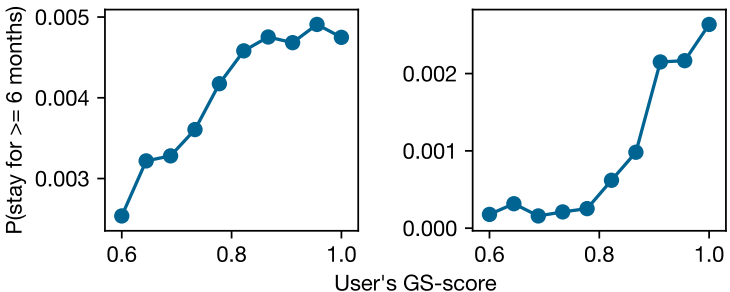
Results



Reddit (left) and GitHub (right)

Results

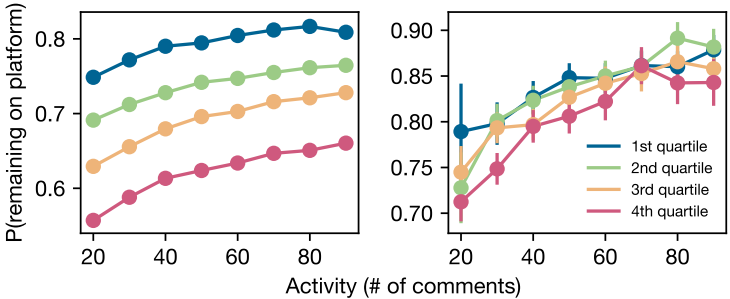
Specialists stay engaged with **communities** longer



Results

Specialists stay engaged with **communities** longer

but generalists stay engaged with the **platform** longer



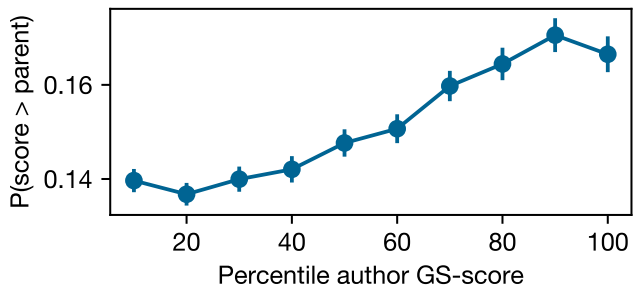
Results

Specialists stay engaged with **communities** longer

but generalists stay engaged with the **platform** longer

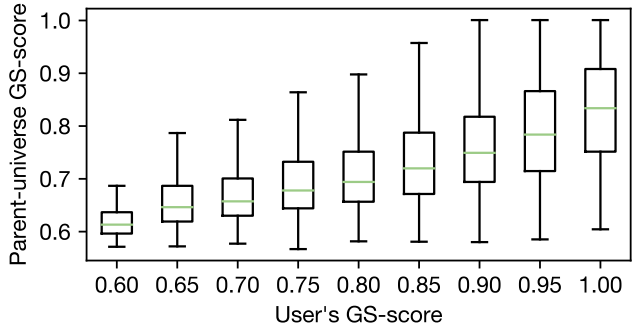
Results

On Reddit, specialists tend to be make more **exceptional comments**



Results

but generalists are exposed to a **more diverse set of users**

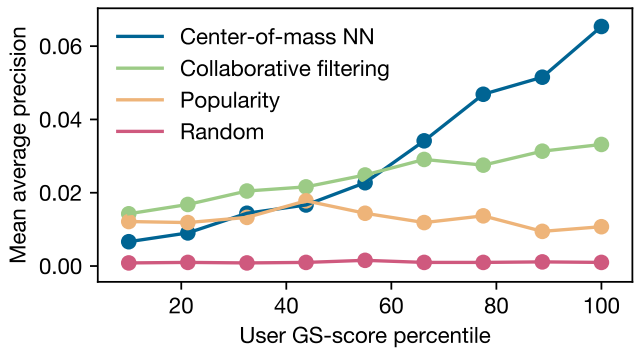


Results

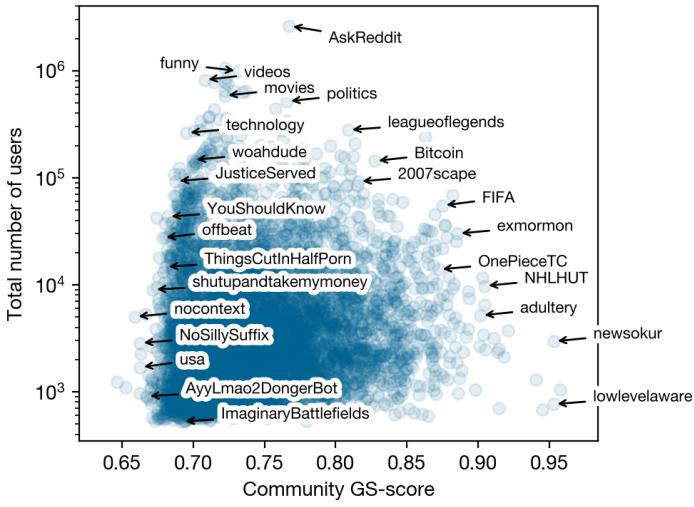
Can GS-score predict new communities a user joins?

Results

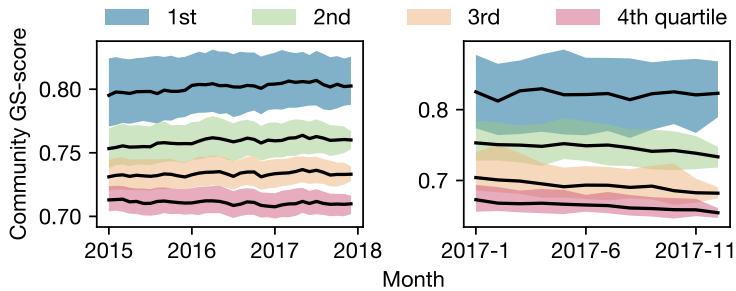
Can GS-score predict new communities a user joins?



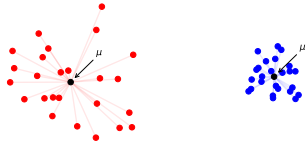
Community GS-scores



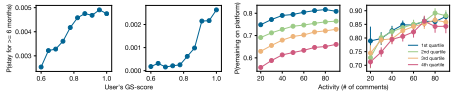
Community GS-scores



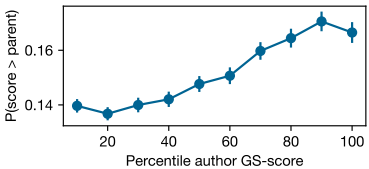
In summary



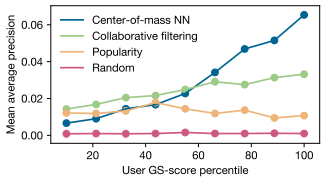
Users on Reddit and GitHub range from generalist to specialist



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



On Reddit, specialists are more likely to make exceptional comments



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...
Computer Science
UNIVERSITY OF TORONTO
A project of the Computational Social Science lab at the University of Toronto (@isaacwaller and @ashton1anderson).
Scroll down to continue
Interactive mode
Thank you!
tiny.cc/gsscore