Caring on Reddit: The Linguistic Lives of Loyalists and Vagrants

Justine Zhang
Tianze Shi
Skyler Seto

Motivation

How invested are users in online communities? It depends! (Gilbert&Karahalios)

Some users are more invested than others **on reddit**.

Subreddit	% of Comments by Top 10% of users
gameofthrones	54
starcraft	66
PoliticalDiscussion	70

In a platform with multiple subcommunities, users take the opportunity to **explore**. (Tan&Lee)

Intuitions

For a community, define two types of users...

Loyalists: users who spend a sizeable portion of their online life in a particular community.

Vagrants: users who spend most of their time in other communities, but occasionally visit.

Why might a user belong to one particular class?

interest, personal investment, expertise, ...

Intuitions

Hence, our question:

How do loyalists and vagrants interact within an online community?

→ are there **linguistic differences**?

Dataset

Reddit, duh.

Some nice things about Reddit:

- We have a lot of **data**.
- There are **multiple communities**, and users can move freely between them.
- Lots of interest-based communities which possibly **inspire different tiers of interest/investment/expertise**.

Some other things about Reddit:

- Communities are about **diverse range of things** (spoiler alert: tricky)
- We have **a lot of** data (and only 1 TB of RAM).

Dataset (What we've looked at)

10 subreddits

starcraft, Android, Music, gaming, PoliticalDiscussion, gameofthrones, nfl, CFB, books, MakeupAddiction

5 (ish) years

2009-2013 and/or subsets thereof, **separated by months**, contingent on data size

(... but it would be nice to look at everything)

Definitions

For **each month**, take all users who contribute > 15 comments. (Never look at the rest again.)

For **each month and subreddit**, compute a user's activity fraction in that subreddit:

*# **unique links** commented on in subreddit / # unique links commented on in all of reddit*

Loyalists: Top 10th Percentile

Vagrants: Bottom 90th Percentile

footnote: tweaking the definitions probably also produces a legitimate, possibly better, dataset.

Definitions - Depth is Confounding

Preliminary analysis:

1. Loyalists and vagrants comment at different depths.
2. Shallow comments (**level <= 1**) are linguistically different from deep ones.

Point 2 is super interesting but not our problem. Hence we split our analysis for shallow and deep comments and mostly focused on the shallow ones.

> *footnote: looking at deep comments gets into the conversational behavior of users, which is interesting and hard.*

Preliminary Statistics

Distribution of Loyalist vs Vagrant comments, numbers of Loyalist/Vagrant users

Average # of comments/unique links per loyalist, or per vagrant user.

The story: there are numerical differences, so are there content ones?

NLP is just counting words

Hypothesized symptom of loyalty: **putting effort into posts** .

loyalists should write longer posts than vagrants (despite writing more!)

	Android	books	CFB	GOT	Makeup Addiction	Music	NFL	Political Discussion	Star craft
WC	**80, 0.635**	**89, 0.618**	**32, 0.667**	**38, 0.633**	**29, 0.604**	**59, 0.447**	**56, 0.712**	**31, 0.646**	**18, 0.75**
SC	**39, 0.619**	**42, 0.568**	**4, 0.167**	**0.0, 17, 0.567**	**2, 0.083**	**47, 0.712**	**19, 0.487**	**12, 0.5**	**4, 0.33**

format: uhhhhhh

NLP is just throwing LIWC at things

Hypothesized symptom of loyalty: **personal investment** ...maybe.

Us:

loyalists should use more personal/concrete pronouns than vagrants.

Deja Review paper:

vagrants are more likely to contribute personal gut reactions than loyalists (“experts”)

NLP is just throwing LIWC at things

Hypothesized symptom of loyalty: **personal investment** ...maybe.

Results:

	Android	books	CFB	GOT	Makeup Addiction	Music	NFL	Political Discussion	Star craft
ipro	**10^-23 95 0.754**	**10^-9 92 0.639**	**0.01 42 0.875**	**10^-9 51 0.85**	0.07 21 0.4375	**10^-31 93 0.7045454545 45**	**10^-8 53 0.67948717948 7**	**10^-5 41 0.854166666667**	**10^-7 22 0.916666666667**

LIWC is pretty fun though

Other stuff:

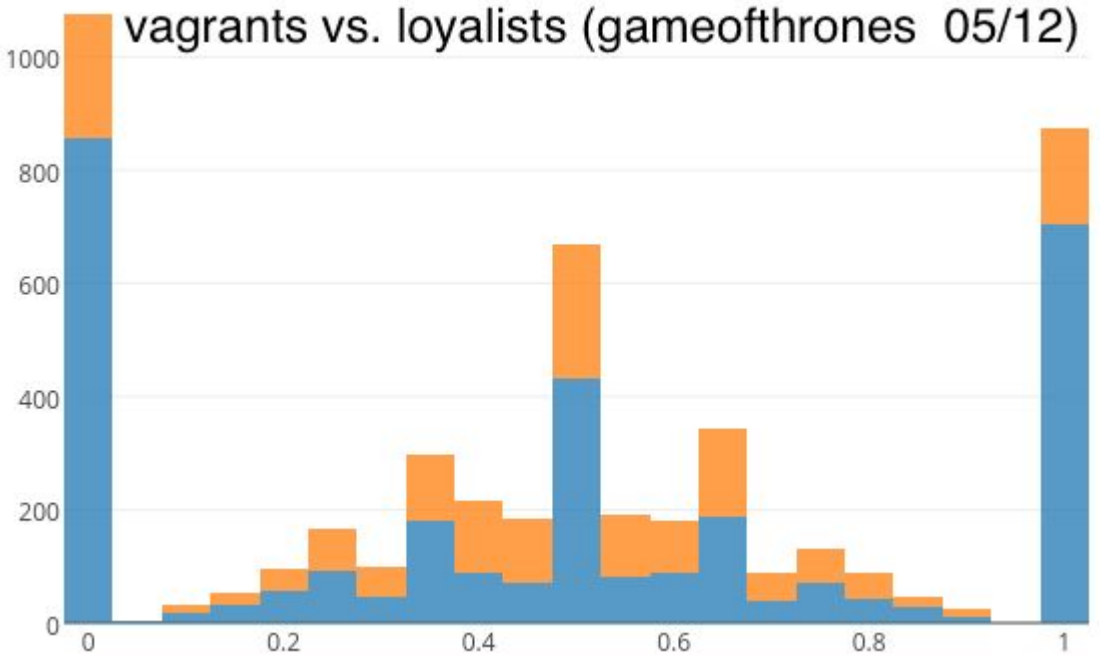
	Android	books	CFB	GOT	Makeup Addiction	Music	NFL	Political Discussion	Star craft
SW	**10^-293 16 0.32**	**10^-26 46 0.32**	**10^-21 14 0.29**	**10^-32 13 0.22**	**10^-16 4 0.08**	**10^-18 9 73 0.56**	**10^-8 32 0.41**	**10^-123 21 0.44**	**10^-11 0 0**
Qs	**0.0 40 0.32**	**0.0 0 0**	**0.0 2 0.04**	**0.0 2 0.03**	**0.0 2 0.04**	**0.0 27 0.21**	**0.0 0 0**	**0.0 21 .438**	**10^-238 0 0**

Is loyalty rewarded?

Hypothesis:

communities prefer loyalists.

Results:



The consistency of loyalty

Jaccard similarity of loyalist sets between months.

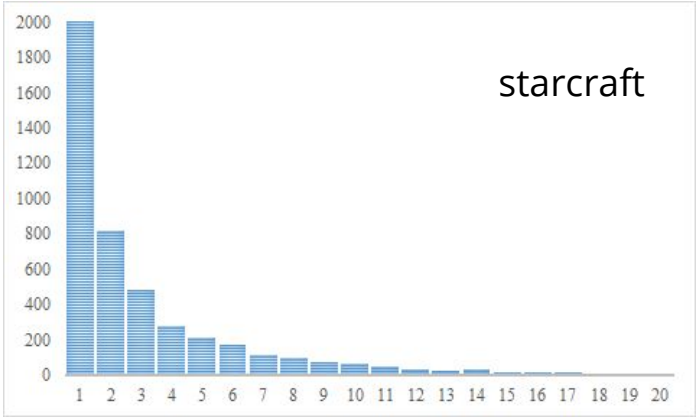
	1 month	2 months	3 months
Android	0.3260	0.2575	0.2252
books	0.1546	0.1105	0.0904
CFB	0.3861	0.2896	0.2315
gameofthrones	0.2045	0.1243	0.0883
gaming	0.2242	0.1684	0.1405

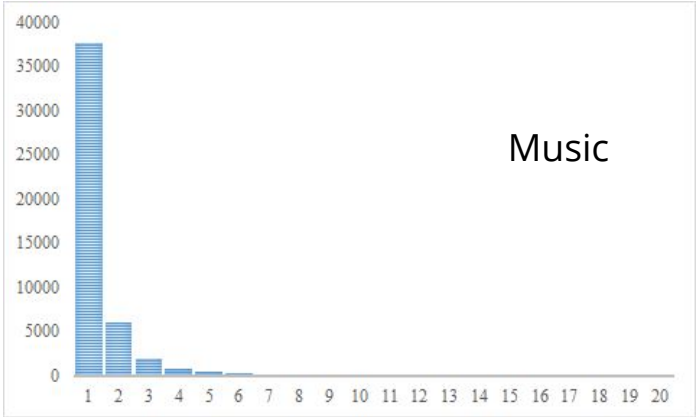
	1 month	2 months	3 months
MakeupAddiction	0.3032	0.2026	0.1508
Music	0.0732	0.0569	0.0480
nfl	0.3764	0.2873	0.2394
PoliticalDiscussion	0.2979	0.2219	0.1792
starcraft	0.3665	0.2775	0.2307

The consistency of loyalty

Average number of “loyal” months

Android	3.007	MakeupAddiction	2.289
books	1.659	Music	1.415
CFB	3.341	nfl	3.326
gameofthrones	1.892	PoliticalDiscussion	2.587
gaming	2.239	starcraft	3.135





Language Models

How vagrants and loyalists differ in language use?

Train language models and see how well they can predict the language!

Language Models

“Training set” (set of comments used to train language models):

- For each (subreddit, month), randomly choose 2,000 comments
- Vocabulary: top 1,000 words for all time, others mapped to <RARE>
- Predict other comments

Bunches of language models:

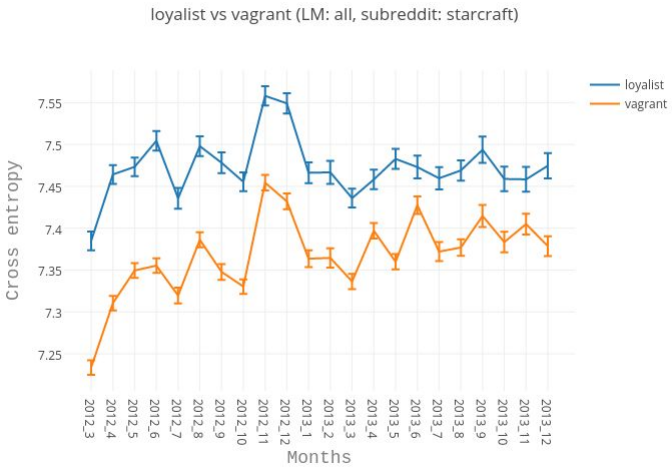
- All comments; vagrant comments; loyalist comments;
- shallow comments; deep comments

Cross Entropy

Use the trained language models to predict other comments

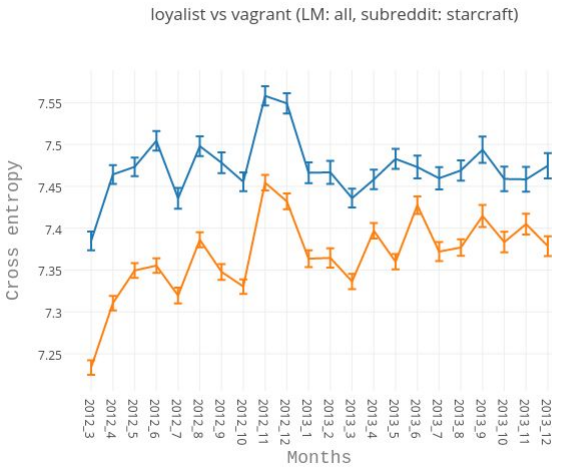
Control for subreddit:	LMs are subreddit-specific
Control for time:	LMs are month-specific
Control for length:	take first 30 words of each comment
Control for type of users	
Control for depth of comments	

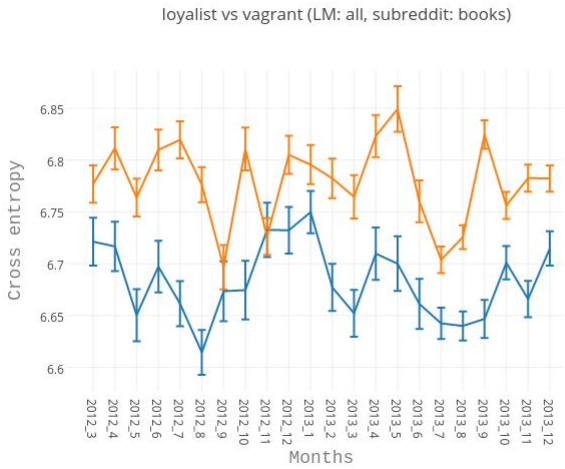
Cross Entropy - Vagrants vs. Loyalists

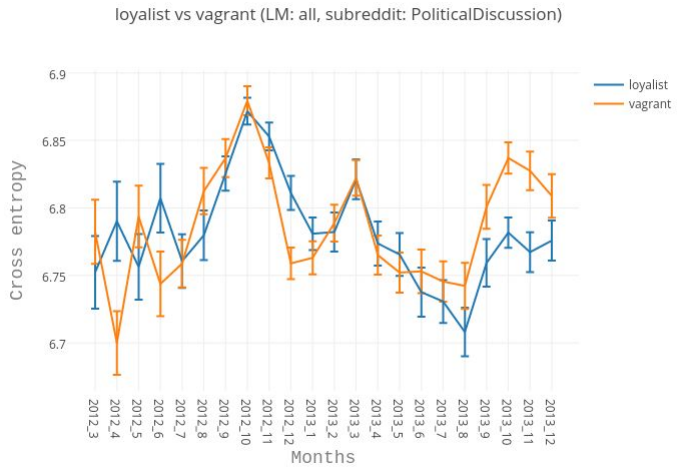


Cross Entropy - Vagrants vs. Loyalists

But, for different subreddits?







Cross Entropy - Vagrants vs. Loyalists

Relation to consistency of royalty (avg. number of loyal months)

Android

3.007

books

1.659

CFB

3.341

gameofthrones

1.892

gaming

2.239

MakeupAddiction

2.289

Music

1.415

nfl

3.326

PoliticalDiscussion

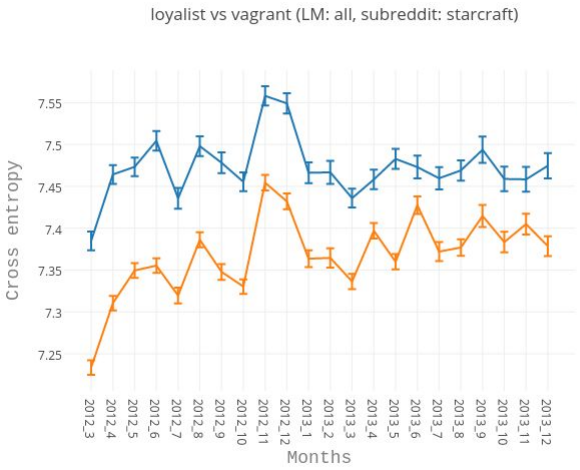
2.587

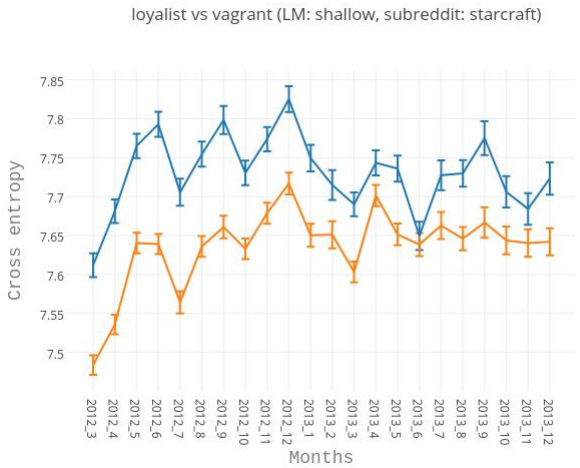
starcraft

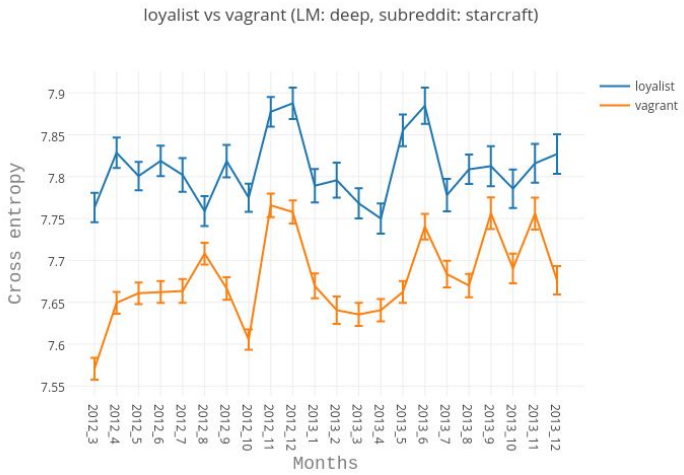
3.135

Cross Entropy - Vagrants vs. Loyalists

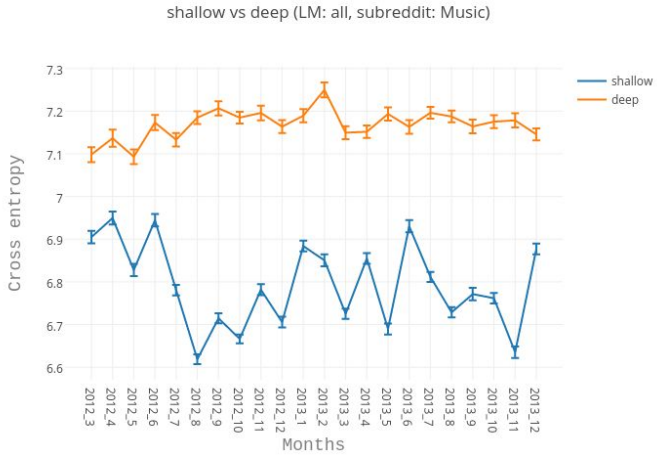
At different depth





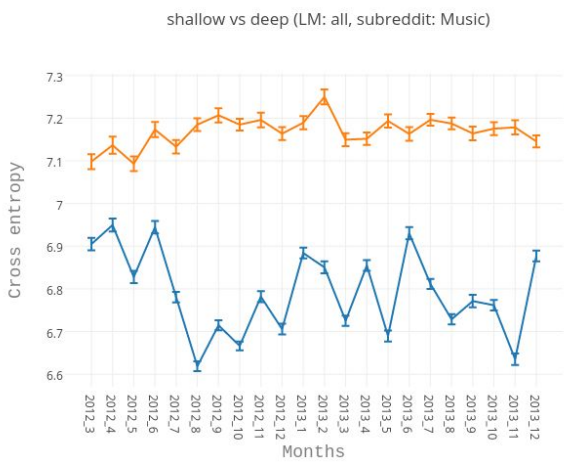


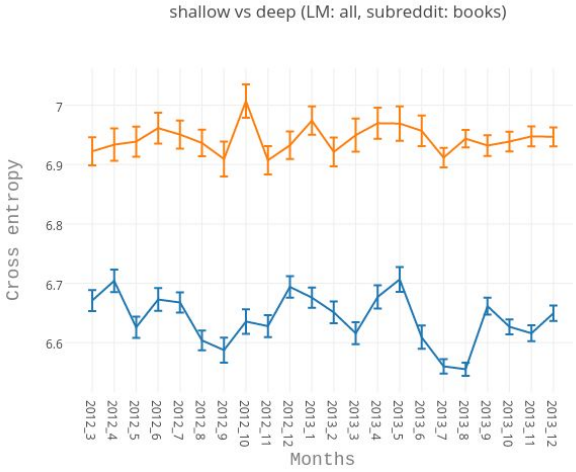
Cross Entropy - Shallow vs. Deep

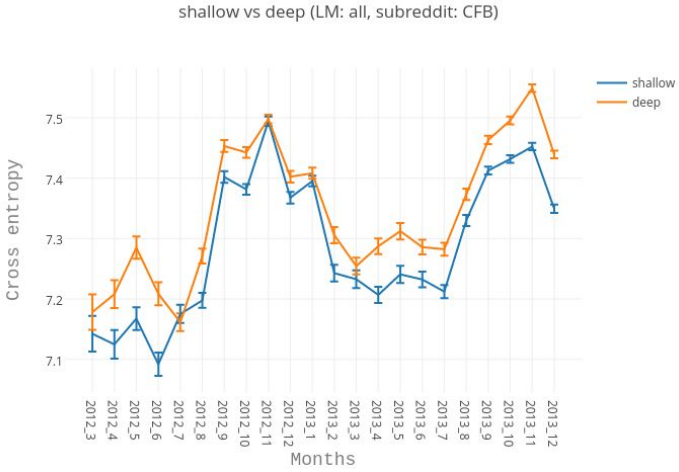


Cross Entropy - Shallow vs. Deep

Different subreddits?

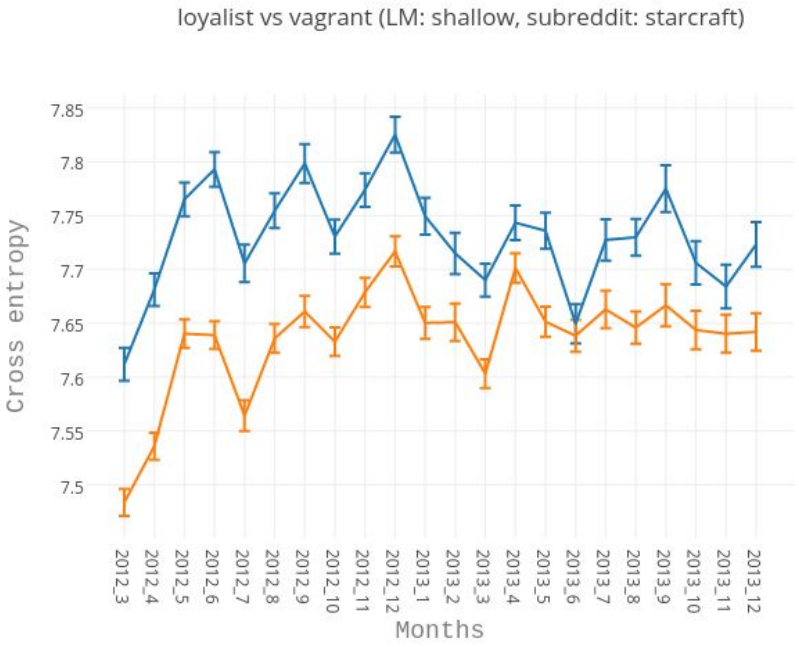


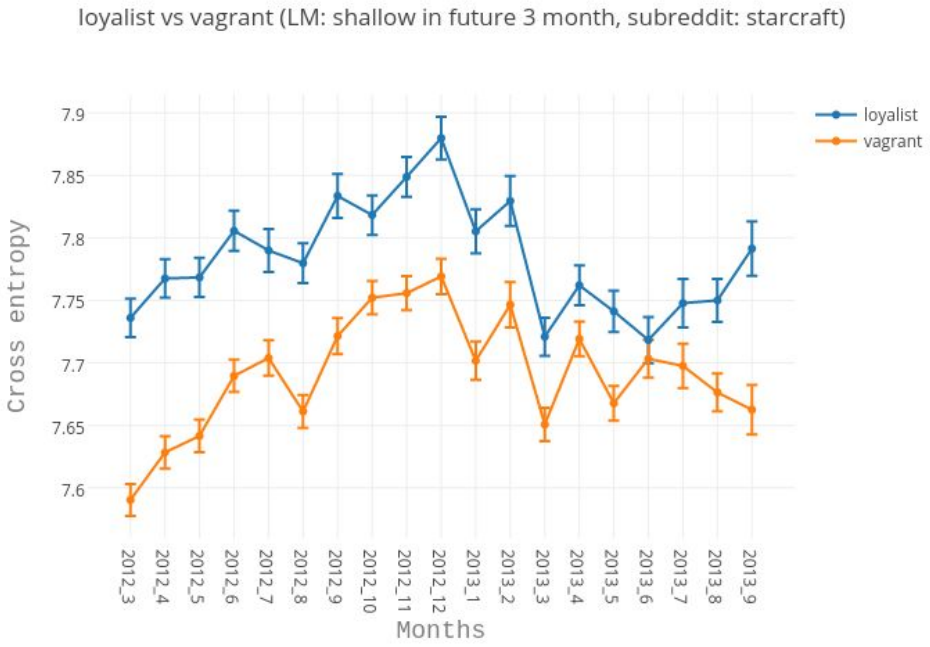




Whose language defines the future?

Patterns are similar





Where does the difference come from?

Content words? (word frequency ranked 1,001-5,000)

vagrants vs. loyalists

Where does the difference come from?

Function words / Stop words? (word frequency ranked within 100)

vagrants vs. loyalists

Behind Users' Activity Change

Users change their activity level:

within a consecutive 6-month window

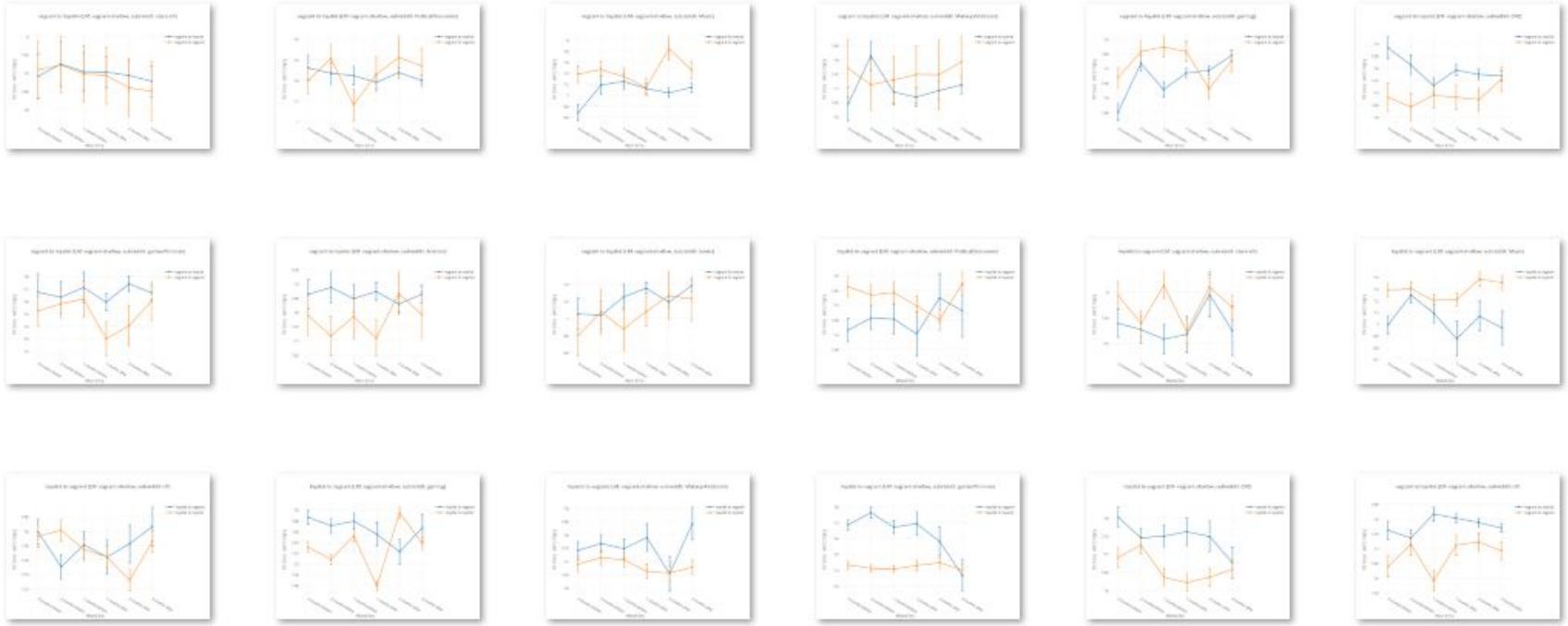
patterns like: *vvv111* or *111vvv*

pair them with: *vvvvvv* *111111*

Control for: subreddit, time, depth (all shallow), length (first 30 words)

Behind Users' Activity Change

Results are noisy, hard to interpret (blue: changing users; orange: non-changing)



Fightiest Words

What words distinguish loyalists from vagrants?

investment, expertise

For each **month and subreddit**, compute fightin' word z-scores for vocabulary split between loyalists and vagrants (we focused on **shallow** comments here).

Fightiest Words: {w: z-score > k}; {w: z-score < -k}

We take *k=3*.

Fightiest Words

Fightiest Words: {w: z-score > 3}; {w: z-score < -3}

Examples:

gameofthrones, March 2013:

loyalist: *asos, chapters, narrative, barristan, stoneheart*

vagrant: *emilia, interview, watching, disney, doctor*

Music, October 2012

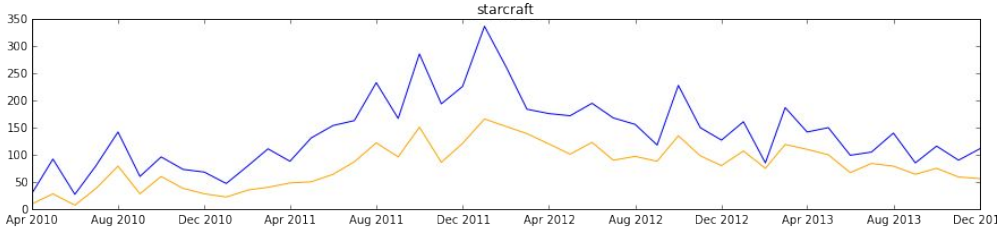
loyalist: *arcade, collective, quadrophenia, pitchfork, underrated*

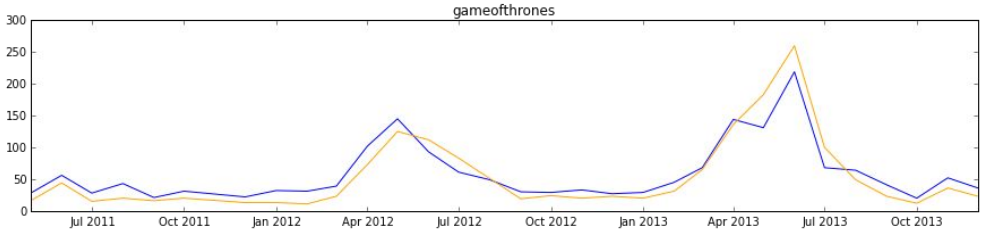
vagrant: *nicki, grooveshark, remix, stupid, videos*

Is loyalty fightier?

Mean number of fighting words per subreddit (over months):

subreddit	loyalists	vagrants
starcraft	**139**	**80**
gameofthrones	58	53
books	**42**	**32**
Android	**95**	**68**
MakeupAddiction	105	87
PoliticalDiscussion	44	34
gaming	316	326





Fightiest Words over time

How consistent are fightin' words over time?

Hypothesis 1 (mine):

loyalist words are more consistent, because loyalists have clearly-defined takes on the subject matter that are stable.

Hypothesis 2:

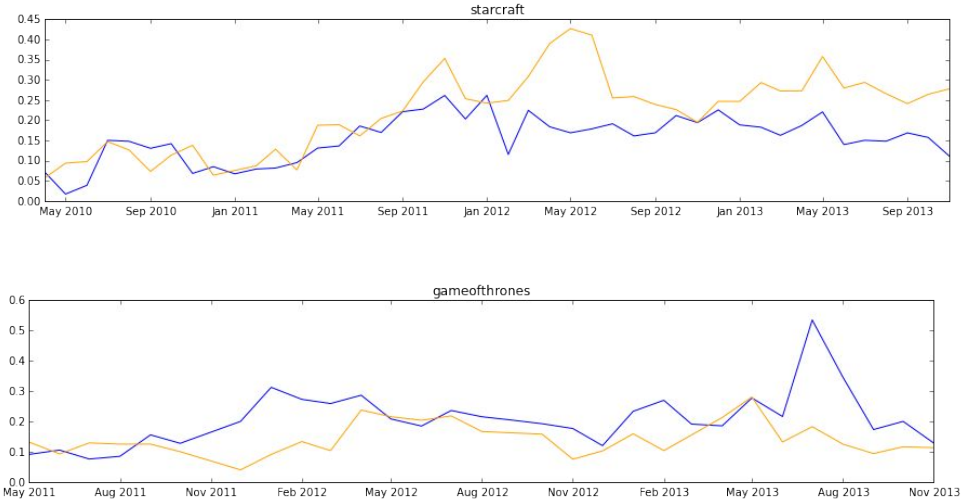
loyalists words are less consistent, because they diligently follow changes in the subject matter.

Fightiest Words over time

For each month *m*, compute the **jaccard similarity** between fightiest words in *m* and *m+dt*.

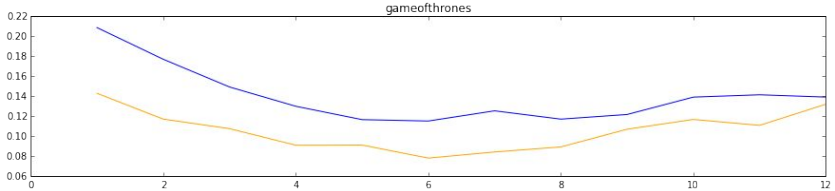
Mean Jaccards, *dt=3*:

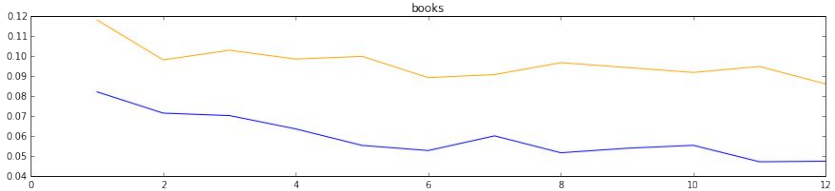
subreddit	loyalists	vagrants
starcraft	**0.15475**	**0.21943**
gameofthrones	**0.20875**	**0.14281**
books	0.08215	0.11820
Android	**0.20075**	**0.16077**
gaming	0.22155	0.19515

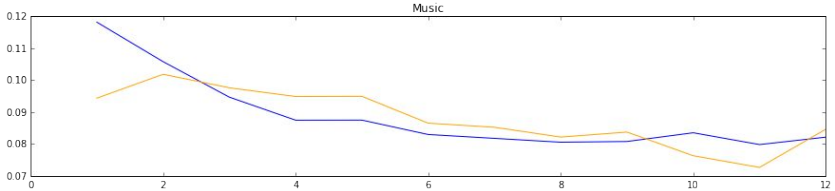


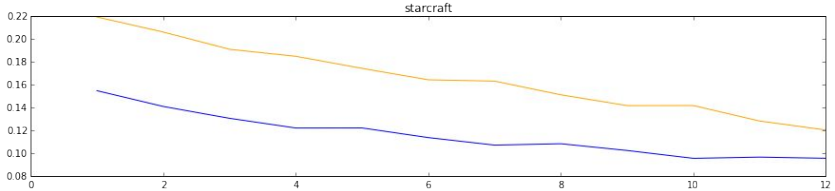
Fightin' Words Have Different Lifespans

Jaccards vs *dt*:





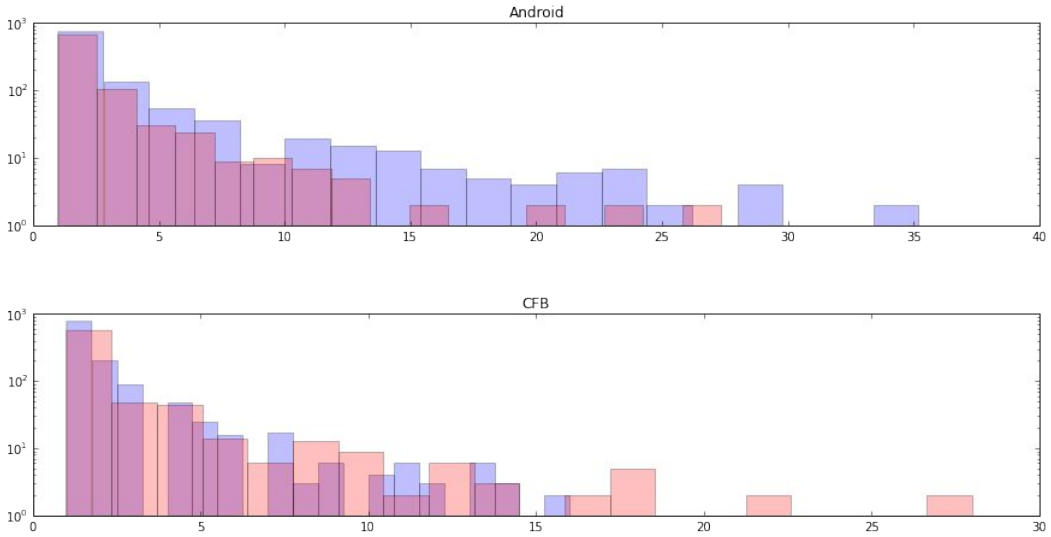




Loyalists are more loyal to their fightin' words

Average number of months a word is fightin':

subreddit	loyalists	vagrants
starcraft	**3.08**	**3.05**
gameofthrones	2.86	2.47
books	**2.33**	**2.19**
Android	**3.44**	**2.83**
nfl	2.64	2.74
CFB	2.01	2.55



Longest-living fightin' words (nsfw: language)

starcraft:

loyalist: teamliquid (33), gsl (26), protoss (22)
vagrant: twitter (28), dota (20), bot (20)

gameofthrones:

loyalist: asos (26), affs (24), speculation (18)
vagrant: fuck (17), hodor (15)

Music:

loyalist: album (44), radiohead (26), beatles (21)
vagrant: fuck (28), shit (14), dubstep (9)

MakeupAddiction:

loyalist: brush (19), palette (18), blush (15)
vagrant: beautiful (22), eyebrows (20), gorgeous (19)

Fine print

Fightiness and usage are slightly different; a word can become a fighting word if the other side uses it *less*. On average they're the same so I thought of them as such, and this is certainly true of the longest-living fightin' words. Letting the two cases be "straightforward" and "gah", vagrant fightin' words were more likely to be "straightforward" than loyalist fightin' words.

In light of this, what the fighting words analysis really tries to say is that one side's language use is more/less distinctive than the others, and how it is distinctive is preserved over time in different ways.

It's probably worth coupling FW with actual word frequencies. I tried and didn't find anything interesting, probably cause I wasn't asking the right questions. Here's a query that yielded small p values: is the lifespan of a FW related to the magnitude of the change in frequency between the month it became a FW and the month previous? Yes for loyalists, for some subreddits; not really, for almost all vagrants.

Also note that while relatively infrequent, a word can be fighty for loyalists and vagrants at different points in time. On average it wasn't clear to me who was introducing such words (loyalists or vagrants) so I didn't pursue this analysis further.

Who introduces such words, anyways? I have no idea about this and plenty of ideas about confounds.

Conclusions and future work

Loyalists and vagrants behave differently, but precisely how depends on the subreddit.

We can make up nice anecdotes to explain the effect direction in individual subreddits. **Can we characterize subreddits by the behaviour of their loyalists and vagrants?**

Loyalists and vagrants both seem to define some community roles in their language use. **Can we couple an analysis of community network structure with loyalist and vagrant roles?**