# Conditional Language Models for Community-Level Linguistic Variation

**Bill Noble** and **Jean-Philippe Bernardy**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
{bill.noble@, jean.philippe.bernardy@}.gu.se

## Abstract

Community-level linguistic variation is a core concept in sociolinguistics. In this paper, we use conditioned neural language models to learn vector representations for 510 online communities. We use these representations to measure linguistic variation between communities and investigate the degree to which linguistic variation corresponds with social connections between communities. We find that our sociolinguistic embeddings are highly correlated with a social network-based representation that does not use any linguistic input.

## 1 Introduction

Linguistic communication requires that speakers share certain linguistic conventions, such as syntactic structure, word meanings, and patterns of interaction. Speakers assume that these conventions are *common ground* among their interlocutors, based on joint membership in a community (Stalnaker, 2002; Clark, 1996). Such *speech communities* (Gumperz, 1972) range in size from the very small, like members of a friend group, to the very large, like speakers of English. However, as Eckert and McConnell-Ginet (1992) point out, it is *communities of practice* — defined by mutual social engagement in a common activity — that are the primary locus of linguistic variation.

Variation is an important object of study in sociolinguistics, and is naturally amenable to computational analysis (Nguyen et al., 2016). Most previous computational work on linguistic variation has considered variation at the level of macro-social categories, such as gender (Burger et al., 2011; Ciot et al., 2013; Bamman et al., 2014b), age (Nguyen et al., 2013), and geographic location (Eisenstein et al., 2010; Bamman et al., 2014a).In the present work, however, we investigate linguistic variation across online communities in the social media website Reddit.

For this purpose, we introduce (section 2) various Community-Conditioned Language Models (CCLMs for short). These models are conditioned on a vector representation (or embedding), which varies by community. Hence, they learn *community embeddings*. We report which architectures make best use of the community information (section 3), however our our primary purpose is not to improve language models in terms of perplexity, but rather to extract community embeddings that capture linguistic similarities between communities and test how the resulting embeddings correspond to the social structure of subreddits. To that end, we test the how well the community embeddings correlate with a social network-based representation of communities (section 4).

The contributions of this work are twofold. First, we develop a language model-based community embedding that we show is correlated with (but still different from) an embedding based on community membership alone. Second, the method we describe for testing the correlation between two embeddings from different models is, to our knowledge, novel to computational linguistics.

## 2 Community-conditioned language models (CCLMs)

We experiment with two kinds of model architecture: simple unidirectional LSTM (Hochreiter and Schmidhuber, 1997) and a masked Transformer (Vaswani et al., 2017). Although Transformer-based language models are considered state-of-the-art, they achieve dominance partly thanks to the availability of very large data sets (e.g., Devlin et al., 2019; Brown et al., 2020), which are not available to us.[1] Thus the LSTM is a worthy

[1] Fine-tuning existing models is not compatible with our methodology, because we fundamentally change the structure of the network by concatenating community embeddings with hidden states at various levels.



*Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 59 - 78

model to test for us.

In either case, the model is organised as a standard 3-layer neural sequence encoder, where the input for the $t$th timestep of the $n+1$st layer is the $t$th hidden state of the $n$th layer. As usual, the input to the first layer, is a sequence of tokens, encoded with a trainable embedding layer over a predetermined vocabulary. At the other end, word tokens are predicted using a softmax projection layer. What we have described so far does not take community into account and as such we call them *unconditioned models*, but the same encoder architecture also forms the core of our conditioned models.

In the CCLMs, we add a *community embedding* parameter, which varies depending on the community of origin of the input sample. This parameter is concatenated (at each time step) with the hidden layer of the sequence encoder, at some layer $l_c \leq n$, and passed through a linear layer which projects the resulting vector back to the original hidden layer size. For $l_c = n$, the output of this linear layer is passed directly to the softmax function, just as the final hidden layer of the sequence encoder is in other models. For $l_c = 0$, the community embedding is concatenated with the token embedding. For this reason, we set the hidden size of the sequence encoder and the size of the token embedding to be equal for all models.

### 2.1 Data sets

We investigate linguistic variation across various communities from the social media website Reddit.[2] Reddit is divided into forums called *subreddits*, which are typically organised around a topic of interest. Users create *posts*, which consist of a link, image, or text, along with a *comment* section. Comments are threaded: a comment can be made directly on a post, or appear as a reply to another comment. Hereafter we refer to such comments as "messages", matching our convention in mathematical formulas: the letter $c$ stands for a community, and $m$ stands for a message.

Our dataset includes messages from 510 subreddits, the set of all subreddits with at least 5000 messages per month for each month of the year 2015. Ignoring empty and deleted comments, we randomly sampled 42 000 messages from 2015 for each community. We reserved 1000 messages from each community for development and testing, leaving a total of 20.4M messages for training.

Using `langid.py` (Lui and Baldwin, 2012), we observe that a majority of the overall messages are classified as English (95% of the test set) and 498 of 510 communities have more than half of their messages classified as English. Given the small amount of non-English data, we decided that the bias introduced by attempting to filter message by language outweighted the potential benefits.[3]

Messages were preprocessed as follows: we excluded the content of block quotes, code blocks, and tables and removed markup (formatting) commands, extracting only rendered text. Messages were tokenized using the default English model for the SpaCy tokenizer Version 2.2.3 (Honnibal and Montani, 2017).

### 2.2 Training scheme

Models used a vocabulary of 40 000 tokens (including a special out-of-vocabulary token), consisting of the most frequent tokens across all communities.

We trained the models on a simple autoregressive language modeling task with cross-entropy loss. Because the Transformer operates on all tokens in the sequence at once, the inputs to the model were masked and incrementally unmasked. We used the AdamW (Loshchilov and Hutter, 2019) optimisation algorithm, with an initial learning rate of 0.001 and no extra control on the decay of learning rate. The batch size was 256 and the maximum sequence length set to 64 tokens, truncating longer messages (16.8% of messages were longer than 64 tokens). During training, a dropout rate of 0.1 was applied between encoder layers and after each linear layer.

All experiments use models with 3 encoder layers, each with hidden (and token embedding) size of 256. The Transformer models had 8 attention heads per layer.[4] The conditional models were given a community embedding with 16 dimensions. We experimented with every possible value for $l_c$, the depth of the community embedding, in a three-layer model ($l_c \in \{0, 1, 2, 3\}$).

We trained the models until the validation loss stopped decreasing for two epochs in a row, and used the weights from the epoch with the small-

[2] Comments were obtained from the archive at `https://pushshift.io/`. (Baumgartner et al., 2020). Code for reproducing our dataset, as well as our pre-trained community embeddings are available at URL.

[3] See section 7 for futher discussion.

[4] This number of attention heads was chosen to give the LSTM and Transformer models a comparable number of parameters (22 171 203 and 21 779 523, respectively).

est validation loss for testing. Each training epoch took approximately 1.5 hours of GPU time.

## 3 CCLM Performance

In this section, we report the performance of the conditioned and un-conditioned models on the held out test set. First, we define two performance metrics: perplexity and information gain. In the following, we use $M$ to refer to messages in the combined test set, and $M_j$ for the partition of the test set originating from community $c_j$.

### 3.1 Perplexity

For a given model, let $H(m)$ be the model's cross-entropy loss, averaged over tokens in $m$. We define the perplexity on a set of messages, $M$, to be the exponential of the model's average cross-entropy loss:

$$\mathsf{Ppl}_M = e^{\text{average}_{m \in M}\, H(m)}$$

**CCLM Information Gain** We also consider the average information gain per token of the CCLM over its baseline un-conditioned counterpart, with the same sequence encoder architecture. For a given message, information gain is defined as the difference between the cross-entropy of the unconditioned model and the conditioned model:

$$H_{\text{LM}}(m) - H_{\text{CCLM}}(m)$$

For a set of messages, $M$, we consider the average information gain in exponential space (as a ratio of perplexities):

$$\mathsf{IG}_M = \frac{e^{\text{average}_{m \in M}(H_{\text{LM}}(m))}}{e^{\text{average}_{m \in M}(H_{\text{CCLM}}(m))}}$$

$$\mathsf{IG}_M = e^{\text{average}_{m \in M}(H_{\text{LM}}(m) - H_{\text{CCLM}}(m))}$$

Unsurprisingly, the conditioned models mostly have lower perplexity than their respective unconditioned baseline models, (i.e., $\mathsf{IG}_M > 1$, table 1). While the absolute performance ($\mathsf{Ppl}_M$) of the LSTM models is better, the best Transformer models have somewhat higher information gain than their LSTM counterparts.

The effect of $l_c$, the depth of the community embedding, is also different across architectures. For the LSTM encoder, the best model concatenates the community embedding after the first encoder layer ($l_c = 1$), but all of the conditioned models perform similarly well. For the Transformer, the best model incorporates the community information first, concatenating it directly to the word vectors ($l_c = 0$). It performs similarly to the model that only integrates the community information after all all the Transformer layers ($l_c = 3$), but the two middle-layer models actually perform worse than the unconditioned model (with $\mathsf{IG}_M < 1$).

| | $l_c$ | test epoch | $\mathsf{Ppl}_M$ | $\mathsf{IG}_M$ |
|---|---|---|---|---|
| LSTM | - | 12 | 68.74 | - |
| | 0 | 13 | 66.16 | 1.039 |
| | 1 | 7 | **66.01** | **1.041** |
| | 2 | 4 | 66.19 | 1.039 |
| | 3 | 4 | 66.35 | 1.036 |
| Transformer | - | 4 | 79.13 | - |
| | 0 | 4 | **75.66** | **1.046** |
| | 1 | 4 | 82.12 | 0.964 |
| | 2 | 7 | 83.53 | 0.947 |
| | 3 | 3 | 75.90 | 1.043 |

Table 1: Performance of baseline (first row for each encoder architecture) and CCLM models. The scope of perplexity and information gain ($M$) is the entire test set, i.e. $5000 \times 510$ messages; 5000 for each community.

We also consider performance stratified by community; that is, $\mathsf{Ppl}_{M_j}$ and $\mathsf{IG}_{M_j}$, where $M_j$ is the set of messages originating from community $c_j$ (fig. 1). We observe a lot of variation in baseline perplexity across communities, with $\mathsf{Ppl}_{M_j}$ ranging from 3.67 to 93.58 for the best conditional LSTM model (fig. 1; also see appendix B for detailed community-level results). The conditioned models also perform differently across different communities — even among the best models, some communities have $\mathsf{IG}_{M_j} < 1$, meaning that the CCLM performs worse than the unconditioned baseline for messages from that community. For other communities $\mathsf{IG}_{M_j}$ is much higher, meaning that the CCLM performs better (fig. 1).[5]

We observe that across all the models we tested, communities where conditioning has the least effect tend to be organised around more general interest topics, such as `/r/relationships` and `/r/advice`, where the subject matter is rele-

[5]Some of the communities with consistently high $IG_{M_j}$ across all models are primarily non-English, but surprisingly, not the three most extreme outliers. There are `/r/counting`, `/r/friendsafari`, and `/r/Fireteams`, the later two of which are places where people coordinate to play video games together. The messages in these communites adhere to highly regular formats, which are presumably conventional to the community.

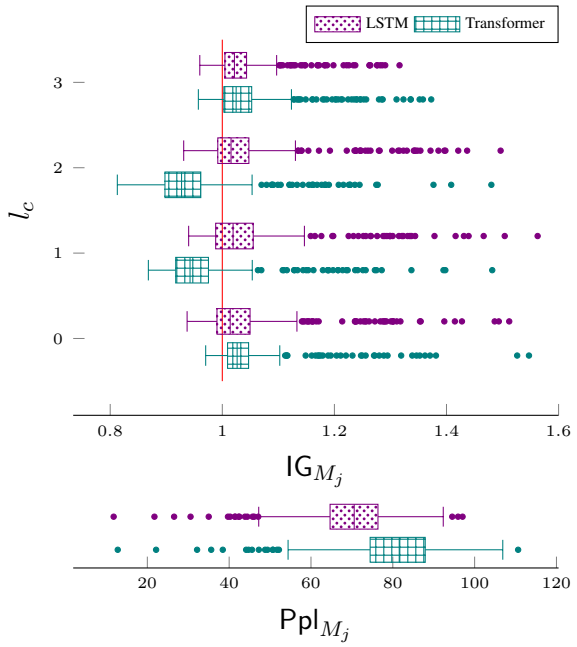


Figure 1: Average model performance by community. The boxes indicate upper and lower quartiles, while the whiskers are placed at the upper and lower maximum, with communities more than $1.5 \times IQR$ (inter-quartile range) above the upper quartile considered outliers (represented as dots). The three most extreme outliers are excluded from this view.

vant to a broad range of people. Conditioning the model on community appears to have the most benefit for narrower special-interest subreddits, such as those organised around a certain videogame, sports team, or subculture. These empirical observations corroborate the idea that communities of practice are the primary locus of linguistic variation.

## 4 Comparison of CCLM community embeddings with a social network embedding

In this section we investigate the degree to which CCLM community embeddings correlate with the social network structure of Reddit.

To this end, we compare the CCLM-learned community embeddings[6] with the community embedding created by Kumar et al. (2018),[7] which were generated using using a negative-sampling optimization algorithm, with the author-community co-occurrence matrix as ground truth, using data from January 2014 to April 2017. We refer the reader to Kumar et al. (2018) for details, but the important point is that no linguistic information is used to create these embeddings: they only reflect the social relationship between communities via community membership. In contrast, CCLM community embeddings depend in no way on which user is the author of any given message: we only use the contents of messages, not authorship data.

### 4.1 Comparing embeddings: cosine similarities

When comparing social embeddings and linguistic embeddings, a difficulty is that they range over completely unrelated spaces. Thus one cannot use the usual cosine similarity metric *between* these spaces. One can, however, use cosine similarity between *pairs* of communities, and verify that the similarities are correlated between linguistic and social embeddings. This gives a way of characterizing the differences between the two kinds of community representation. To get a more concrete sense of what this method yields, we first survey some of the most salient community pairs. We stress that this survey is not meant as a rigorous statistical analysis, as we shall see. Rather it is meant to give a flavor of discrepancies and similarities existing between linguistic and social relations.

We consider communities from three different selection criteria: Those with high linguistic *and* social similarity (where the sum of the two is highest), those with high linguistic and low social similarity (where social similarity is below the median and linguistic similarity is highest), and those with low linguistic and high social similarity (where linguistic similarity is below these median and linguistic similarity is highest).[8][9] We do not consider pairs of communities that are different in both ways, since these don't offer much in the way of understanding the respective embeddings.

Unsuprisingly, the first category (fig. 2, left) yields communities that are qualitatively very similar. The `/r/SSBPM` and `/r/darksouls` communities are focused around discussion of a par-

[6] In this section, we only consider the embeddings from the *best* (highest information gain) CCLM from each architecture family; that is, the LSTM with $l_c = 1$ and the Transformer with $l_c = 0$, however we observed similar results for other values of $l_c$.

[7] Available at `https://snap.stanford.edu/data/web-RedditEmbeddings.html`

[8] We use the LSTM ($l_c = 1$) community vectors for these purposes, but results attain with the best Transformer model.

[9] Median similarity among pairs of communities was 0.177 for the social embedding and 0.010 and 0.012 for the LSTM and Transformer linguistic embeddings, respectively.

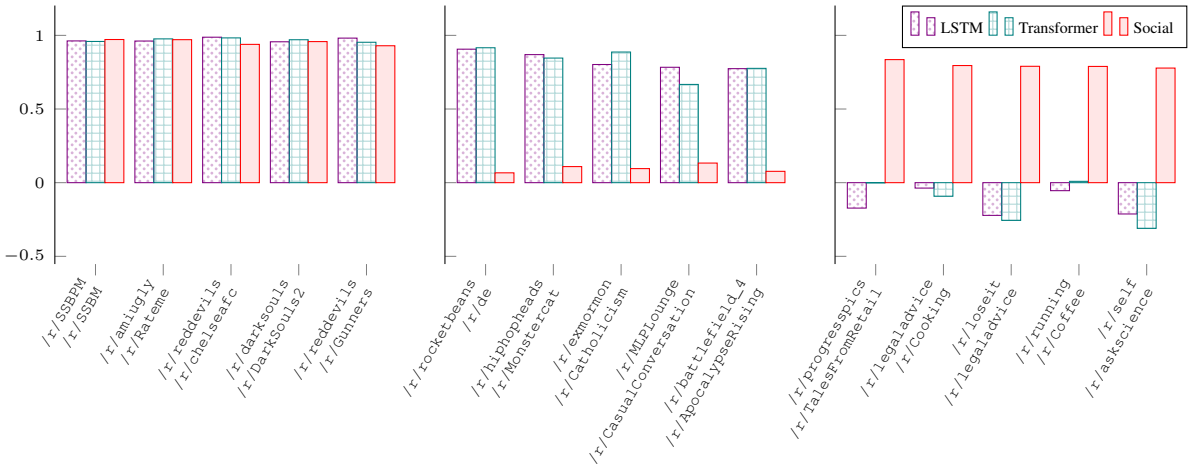


Figure 2: Cosine similarity between pairs of communities, computed for vectors from the best CCLM embeddings (LSTM: $l_c = 1$, Transformer: $l_c = 0$) and the social embedding from Kumar et al. (2018). Communities with high linguistic and social similarity **(left)**, high linguistic but low social similarity **(center)**, and low linguistic but high social similarity **(right)**. See text for details on the selection criteria.

ticular videogame, and are paired with communities that discuss a variation of the same game. The `/r/amiugly` and `/r/Rateme` communities are both forums where the posts are selfies and the comments are mostly comments on the person's appearance. The two communities paired with `/r/reddevils` are likewise comprised of fans of a particular English football club.

Communities with similar linguistic embeddings but dissimilar social embeddings (fig. 2, left) tend to share a similar topic, mode of interaction, or language variety, but in all cases we looked at, there is some reason to expect that they might nevertheless attract different members. For example, `/r/hiphopheads` and `/r/Monstercat` are both topically related to music, but the music genres are different, and the later has a more geographically local focus (Monstercat is an independent electronic music label based in Vancouver). The interactions in both `/r/MLPLounge` and `/r/CasualConversation` could be described as casual conversation, the former is intended specifically for members of a niche internet sub-culture. The `/r/exmormon` and `/r/Catholicism` communities discuss the Mormon and Catholic churches, although their members have different relationships towards those organizations — the former is intended for former members of the church, whereas the later is geared towards practicing Catholics. Finally, both `/r/rocketbeans` and `/r/de` are primarily German-language subreddits, but the former is comprised of fans of a computer gaming YouTube channel, while the later is more general-interest.

Differences at the other end of the spectrum (fig. 2, right) are somewhat harder to interpret. It is mostly easy to see why these communities would have different linguistic embeddings — in all cases the topics are quite different. The reason they have similar social embeddings is less obvious, but we can discern some trends in how the communities are premised. The `/r/progresspics` and `/r/TalesFromRetail` are premised, in part, on seeking support from other people with similar experiences; `/r/legaladvice`, `/r/Cooking`, and `/r/loseit` all involve sharing knowledge on a particular topic; `/r/running` and `/r/Coffee` are hobby-focused; and `/r/self` (often) and `/r/askscience` (by premise) are places people ask and answer questions. It may be that there are different patterns in the *social function* that people attribute to this particular social media website — people who use Reddit in one way are more likely to belong to communities that are premised on the same kind of social function, even if the topics (and indeed language) of those communities are quite different. Testing this hypothesis would require a more focused study design and ideally consider communities from multiple social networks (online or otherwise).

In sum, empirical observation simultaneously reveals examples of high and low correlation between social and linguistic embeddings. To quantify correlation and extract the general trends, we must resort to statistical tools, as we do below.

A straightforward (but ultimately flawed) way to measure how similar the two spaces are would be to generalise the above method, by consider each pair of communities $(i, j)$, and compute the correlation between the cosine similarities of both embeddings.

That is, we can compute the Pearson correlation factor of the data set:

$$C = \{(x = L_i \cdot L_j, y = S_i \cdot S_j) \text{ for } i, j \in [1, 510]\}$$

where $L_i$ and $S_i$ are the linguistic and social embeddings for community $i$. (Thus $L$ is the matrix of (normed) linguistic embeddings and $S$ the matrix of (normed) social embeddings.)

The analysis shows positive correlation for both the LSTM ($r = 0.438$) and Transformer ($r = 0.452$) linguistic embeddings.[10] The correlations are significant with $p < 0.001$ in all cases. However, we note that the number of pairs grows with the square of the number of communities (with 510 communities, we have 129795) pairs), meaning that standard statistical tests on Pearson correlation will assure us of statistical significance in all but the weakest of correlations. A further flaw is that the data points in $C$ are *not* distributed independently — far from it in fact, since each data point is generated from 2 of 510 independent variables. We consider this last flaw fatal, and take a different approach for computing the correlation between community embeddings in the next section.

### 4.2 Comparing embeddings: Procrustes method

In this section, we propose a systematic approach with which we can quantify the correlation between social proximity and linguistic proximity, and measure its statistical significance.

Instead of comparing embedding pairs, as in section 4.1, we will compare embeddings community by community. A naive approach would be to calculate the distance between two embeddings index-wise, which is equal to the Frobenius distance between $L$ and $S$:

$$||L - S||_F = \sum_i (L_i - S_i)$$

The problem with the above metric is that even if several dimensions of $L$ and $S$ are correlated, they will not coincide in the *representation* of embeddings. That is, re-aligning the embeddings by applying a simple rotation (orthogonal transformation) on either matrix widely changes the $||L - S||_F$ correlation metric.

To make the metric independent of the representation (up to orthogonal transformations, which preserve cosine similarities), we compute the *minimum* distance between $L_i$ and $S_i$, for any orthogonal matrix $\Omega$ applied to $L$:

$$d(L, S) = \mathsf{argmin}_\Omega ||\Omega L - S||_F$$

Here, the orthogonal matrix $\Omega$ gives a map from linguistic embeddings to social embeddings. The problem of computing $d(L, S)$ is known as the orthogonal Procrustes problem (Gower and Dijksterhuis, 2004).[11] The solution is

$$d(L, S) = n - \mathsf{Tr}(\Sigma)$$

where the matrix $\Sigma$ is obtained by the singular value decomposition (SVD) $U^T \Sigma V = L S^T$. The vectors of $U$ and $V$ give the directions of correlation respectively of $L$ and $S$. That is, each singular value $\sigma_i$ (the elements of the diagonal matrix $\Sigma$), gives a measure of how much correlation there is between the directions $U_i$ and $V_i$.

As is common when doing SVD, we arrange $U$, $V$ and $\Sigma$ such that $\sigma_i > \sigma_j$ iff $i < j$. Doing so, the largest singular value $\sigma_0$ corresponds to the principal directions of correlation ($U_0$, $V_0$), $\sigma_1$ to the second principal direction, etc.

The $d(L, S)$ metric ranges from 0 (corresponding to perfect correlation, obtained for example if $L = S$) to $n$ (corresponding to perfect orthogonality), where $n = 510$ is the number of communities considered.

Now, to test if $d(L, S)$ corresponds to a significant correlation, it suffices to check if its value is significantly larger than the same value for random linguistic embeddings $L'$. The distribution of $d(L', S)$ for random embeddings is difficult to compute analytically, but we can instead evaluate it using a Monte Carlo method.

Doing so, we observed that $d(L', S)$ exhibits a mean of $\mu_d = 431.39$ and a (Bessel's-corrected) standard deviation $s_d = 2.90$ in their distance from the social embedding, $S$.

Thus if the real $d(L, S)$ is below the mean by several standard deviations, we can safely assume

[10] By comparison, the correlation between the two linguistic embeddings is 0.759.

[11] This approach has also been used to compare word embeddings across representations (e.g., Hamilton et al., 2016).

| | | LSTM | Transformer |
|---|---|---|---|
| $l_c$ | 0 | 254.06 (61.21) | 239.41 (66.79) |
| | 1 | 245.14 (64.29) | **232.18** (68.54) |
| | 2 | 249.17 (62.90) | 233.47 (68.32) |
| | 3 | **241.13** (65.67) | 237.74 (66.84) |

Table 2: Distance between CCLM embeddings and the social network-based embedding of Kumar et al. (2018), as measured by $d(L, S)$. In parentheses is the number of standard deviations from the mean distance of our random embedding samples.

that there is statistically significant correlation between $L$ and $S$. A 4-sigma difference has less than one percent chance of occurring randomly. In our case, we observe a difference of between 61 and 68 standard deviations (table 2). This definitely indicates a significant correlation. Furthermore, by coming back to the definition of $d(L, S)$, we know that, on average, the cosine similarity between $\Omega L$ and $S$ is $0.45 = (510 - 232.18/510)$. It further means that if we obtain a linguistic embedding $L_k$ for a new community $k$, we can estimate its social embedding by $\Omega L_k$, and the cosine similarity with its true social embedding $S_k$ is expected to be $0.39 = (431.39 - 232.18)/510$ — accounting for over-fitting effects by taking the average distance rather than the maximum. In sum, it is clear that the CCLM embeddings predict some aspect the social-network embeddings — but far from all of it.

To finish, we also give a sense of *how* the correlation is manifested overall, by analysis of the two principal components of correlation in the linguistic embeddings, $U_0$ and $U_1$. To do so we plot the projection of each embedding along their first two principle components which, together with the corresponding singular values, gives an idea of how much and in what way they differ (see fig. 4).

## 5 Related work

We have presented results using conditional neural language models to model variation between speech communities. The architecture of these models concatenates a vector representation of the conditioned variable to the input of the sequence model. This approach has been applied in various conditioned text generation domains such as image captioning (Vinyals et al., 2015), machine translation (Kalchbrenner and Blunsom, 2013), but it has not, to our knowledge, been used extensively to study linguistic variation.

There are, however, related applications of conditional neural language models. Lau et al. (2017) presents a neural language model that jointly learns to predict words on the sentence-level and represent topics on the document level. The topic representation is then fed back into the language model, improving its performance on next word prediction. This is similar to how our model experiences improved performance by learning community representations. Unlike our model, topics are inferred in an unsupervised way, raising the question of whether communities could be identified from unlabeled data as well.

A piece of work with similar goals as ours is that of O'Connor et al. (2010), which uses a Bayesian generative model to infer communities from variation in text data. In contrast to our work, this model treats words as independent events, ignoring the structure (and variation) in the construction of sequences. It does further suggest, however, that community-level variation can be modeled in an unsupervised way.

Del Tredici and Fernández (2017), use a modified skip-gram model to community-level linguistic variation. They show that lexical semantic variation occurs even across different communities organised around the same topic. Their approach does not result in community level representations, however.

There are several other recent studies that aim to measure *linguistic distinctiveness* at the level of speech community (O'Connor et al., 2010; Zhang et al., 2017; Lucy and Bamman, 2021). Distinctiveness is one possible interpretation of the community-stratified information gain of the CCLM over its unconditioned counterpart (section 3.1). Whereas the metrics in previous work are based on lexical frequency (and in the case of Lucy and Bamman (2021), word sense distributions), CCLM information gain is capable of capturing distinctiveness at multiple levels of linguistic analysis. However, further work is needed to investigate exactly what kinds of variation are captured.

While the focus of this paper is sociolinguistic aspects, computational models of variation can also support robust, equitable language technology. Previous work has shown that speaker demographics can improve performance on standard NLP tasks (Hovy, 2015; Yang and Eisenstein,

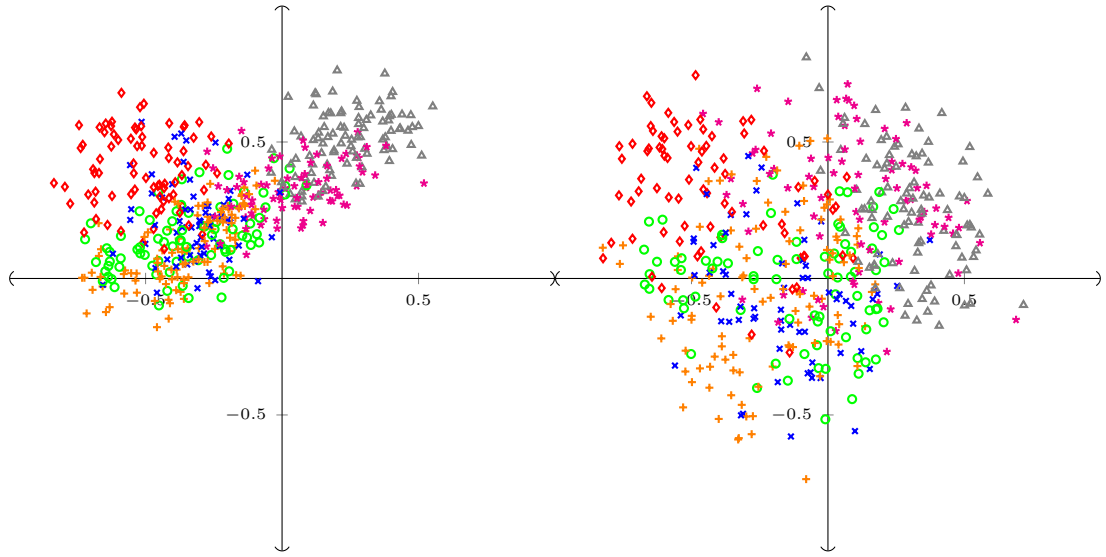


Figure 3: First two components of the aligned social (left) and linguistic (right) embeddings, where the lingusitic embedding is taken from the LTSM with $l_c = 1$. Correlation between these directions is given by $\sigma_0 = 53.4$ and $\sigma_1 = 35.6$. Colors are assigned by k-means clustering of the social embedding. This figure is reproduced in the supplementary materials with a legend that helps to characterise the clusters.

2017).

## 6 Discussion and Conclusion

To sum up our findings, we have defined community-conditioned language models (CCLMs). These models are generally able to attune to community-specific language, as witnessed by the information gain that they exhibit over baseline unconditioned models.

We find that the layer depth of the community embedding ($l_c$) has a weak effect on the information gain and the perplexity of the CCLMs.

For LSTM models, the perplexity per word, averaged over messages from all communities, was between 66.01 and 66.35 (with 68.74 for the unconditioned model). For Transformer models, it varies a bit more, between 75.66 and 83.53, but this seems to be mainly due to the poor performance of the models where the community embedding is inserted between Transformer layers ($l_c = 2$ and 3 both test above the unconditioned Transformer's average perplexity of 79.13).

The pattern of information gain by community is similar across architectures; communities that benefit most from the conditioned model behave that way for both the LSTM and Transformer. However, there are some differences. For example, many of the communities with the biggest difference in information gain between the $l_c = 0$ and $l_c = 3$ LSTMs are organised around trading collectables or organising virtual meetups (e.g., `/r/Pokemongiveaway`, `/r/ACTrade`, and `/r/SVExchange`). These communities tended to have highly conventionalized ways negotiating trades and coordinating meetups. It would be interesting to investigate these differences further in future work, since it could reveal differences in the kind of linguistic variation the different model architectures capture.

Our main result is that community representations learned by CCLMs are positively correlation with user co-occurrence patterns. Even though such *homophilic* correlation is a core hypothesis of sociolinguistics (see Kovacs and Kleinbaum (2020), for example), we believe that this study is the first to test it at the level of communities of practice using computational methods. Furthermore, it appears that our method (correlating linguistic embeddings and social embeddings) is novel. Indeed, even though the Procrustes method has been used to correlate two sets of linguistic embeddings *for the same model*, we find no evidence of the method being applied to embeddings for widely different models, as we have done.

## 7 Ethical considerations

**Data privacy** Our work uses publicly available data from Reddit, collected from the API made available by Baumgartner et al. (2020). Additional considerations apply, however (see Gliniecka et al.

(2021) for discussion). Reddit users are not, in general, aware of the possibility that their data will be used for research purposes, and deleted posts can persist in archive formats. We do not release any data, since the it is already publicly available and duplicating the dataset increases the likelihood that deleted posts will persist.

The paper does not include any text that could be linked back to personally identifiable information. We do release our trained community embeddings, but they have low dimensionality and pose a low risk for exposing personally identifiable information.

**Language identification** As mentioned in section section 2.1, we decided not to filter our data for non-English comments. Although our focus in this paper is intra-language variation, language identification has the potential to introduce bias by reinforcing hegemonic language classes and the boundaries between them. In our case, filtering out messages classified as non-English would introduce bias by disproportionately removing messages in non-standard and code-switched language varieties, which are of interest in the current work.

Nevertheless, the representations learned by our model are (necessarily) relative to the other communities in the dataset. Thus the learned representations for non-English communities tend to be more similar to each other than to other communities that use mostly English, even if their predominant language is not the same. This would probably not be the case if the distribution of messages was more varied across hegemonic language classes; our work cannot be used to conclude, for example, that there is more variation within English than between Dutch and German.

**Subjective analysis** In the qualitative discussion offered in section 4.1, our comparative characterization of the topic, mode of interaction, and language varieties used in the pairs of communities were formed by reading comments from the data our language models were trained on. This included Googling words and phrases that were unfamiliar. Where we make claims about the how the community is "premised" or what kinds of members it is "geared towards" or "intended for", these are based on the text of the sidebar on the community's Reddit page. While we believe this methodology, aggregated over many pairs of communities, is appropriate for making a qualitative comparison of the community features encoded by different representations, to make conclusions about *particular* communities based on such an analysis would be dubious and potentially harmful.

## Acknowledgements

This work was supported by grant 2014-39 from the Swedish Research Council (VR) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

## References

David Bamman, Chris Dyer, and Noah A. Smith. 2014a. Distributed Representations of Geographically Situated Language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *arXiv:2001.08435 [cs]*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Seattle, Washington, USA. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Marco Del Tredici and Raquel Fernández. 2017. Semantic Variation in Online Communities of Practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long Papers*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Penelope Eckert and Sally McConnell-Ginet. 1992. Communities of practice: Where language, gender, and power all live. *Locating Power, Proceedings of the 1992 Berkeley Women and Language Conference*, pages 89–99.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.

Martyna Gliniecka, Joseph Reagle, Nicholas Proferes, Casey Fiesler, Sarah Gilbert, Naiyan Jones, Michael Zimmer, Huichuan Xia, Connie Moon Sehat, Tarunima Prabhakar, and Aleksei Kaminski. 2021. AoIR ethics panel 2: Platform challenges. *AoIR Selected Papers of Internet Research*.

J. C. Gower and Garmt B. Dijksterhuis. 2004. *Procrustes Problems*. Number 30 in Oxford Statistical Science Series. Oxford University Press, Oxford ; New York.

J Gumperz. 1972. The Speech Community. In Pier Paolo Giglioli, editor, *Language and Social Context: Selected Readings*. Harmondsworth : Penguin.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Explosion.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, page 10, Seattle, Washington.

Balazs Kovacs and Adam M. Kleinbaum. 2020. Language-Style Similarity and Social Networks. *Psychological Science*, 31(2):202–213.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 933–943, Lyon, France. ACM Press.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically Driven Neural Language Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]*.

Li Lucy and David Bamman. 2021. Characterizing English Variation across Social Media Communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Marco Lui and Timothy Baldwin. 2012. Langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. How Old Do You Think I Am?: A Study of Language and Age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, page 10.

Brendan O'Connor, Jacob Eisenstein, Eric P Xing, and Noah A Smith. 2010. A mixture model of demographic lexical variation. In *In Proceedings of NIPS Workshop on Machine Learning for Social Computing*, page 6, Vancouver, BC, Canada. 2010.

Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy*, 25(5-6):701–721.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. *arXiv:1411.4555 [cs]*.

Yi Yang and Jacob Eisenstein. 2017. Overcoming Language Variation in Sentiment Analysis with Social Attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community Identity and User Engagement in a Multi-Community Landscape. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 377–386. International AAAI Conference on Weblogs and Social Media.

# A Projection of aligned embeddings

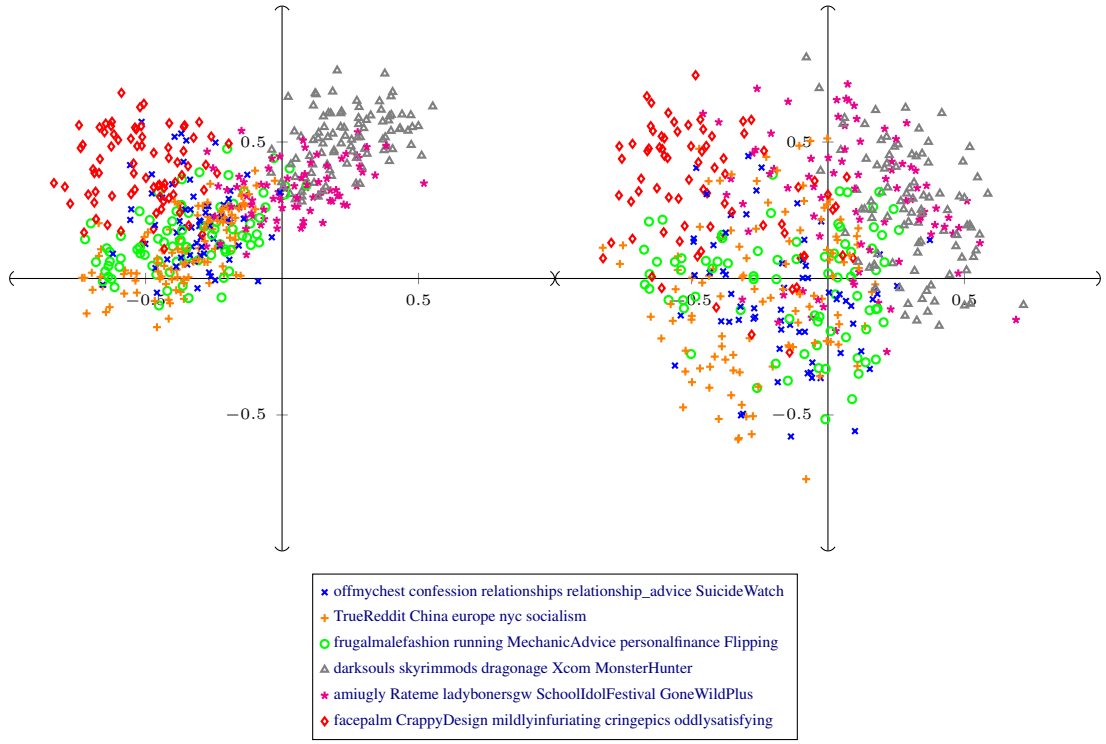


Figure 4: First two components of the aligned social (top) and linguistic (bottom) embeddings, where the lingusitic embedding is taken from the LTSM with $l_c = 1$. Correlation between these directions is given by $\sigma_0 = 53.4$ and $\sigma_1 = 35.6$. Colors are assigned by k-means clustering of the social embedding. The legend shows the closest 5 communites to each cluster centroid. The legend shows the closest 5 communites to each cluster centroid. The cluster of each community is also available in appendix B

## B Community-level results

The following table shows results at the community level. The baseline $\mathsf{Ppl}_{M_j}$ is computed from the unconditioned LSTM and the CCLM results ($\mathsf{Ppl}_{M_j}$, $\mathsf{IG}_{M_j}$, and $\mathsf{Ind}_{M_j}$ use the LSTM with $l_c = 1$). "Social cluster" is determined by k-means clustering of the social embedding.

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| ukraina | 21.74 | 15.19 | 1.43 | 0.006 | 4 |
| france | 68.34 | 50.85 | 1.34 | 0.008 | 1 |
| brasil | 64.13 | 57.68 | 1.11 | 0.008 | 1 |
| podemos | 55.71 | 42.89 | 1.3 | 0.008 | 4 |
| Denmark | 64.2 | 54.56 | 1.18 | 0.009 | 1 |
| de | 71.06 | 54.49 | 1.3 | 0.01 | 1 |
| rocketbeans | 95.95 | 74.17 | 1.29 | 0.011 | 4 |
| thenetherlands | 69.16 | 53.61 | 1.29 | 0.011 | 1 |
| italy | 59.56 | 44.62 | 1.33 | 0.011 | 1 |
| argentina | 70.14 | 53.01 | 1.32 | 0.012 | 4 |
| Romania | 58.34 | 43.84 | 1.33 | 0.012 | 1 |
| sweden | 53.21 | 43.12 | 1.23 | 0.013 | 1 |
| friendsafari | 26.55 | 9.76 | 2.72 | 0.026 | 4 |
| Fireteams | 43.85 | 20.01 | 2.19 | 0.039 | 4 |
| SVExchange | 44.44 | 30.88 | 1.44 | 0.062 | 4 |
| summonerschool | 83.28 | 75.35 | 1.11 | 0.082 | 4 |
| EDH | 76.55 | 58.21 | 1.32 | 0.085 | 3 |
| buildapcforme | 74.75 | 69.45 | 1.08 | 0.098 | 3 |
| Pokemongiveaway | 47.1 | 32.12 | 1.47 | 0.099 | 4 |
| summonerswar | 90.06 | 81.71 | 1.1 | 0.108 | 4 |
| ACTrade | 47.82 | 33.77 | 1.42 | 0.121 | 4 |
| makeupexchange | 53.03 | 40.85 | 1.3 | 0.136 | 4 |
| SkincareAddiction | 58.2 | 56.43 | 1.03 | 0.153 | 0 |
| listentothis | 35 | 32.07 | 1.09 | 0.157 | 5 |
| pokemontrades | 52.37 | 41.69 | 1.26 | 0.175 | 4 |
| AsianBeauty | 70.35 | 67.43 | 1.04 | 0.177 | 4 |
| MechanicAdvice | 82.64 | 78.14 | 1.06 | 0.179 | 2 |
| amiugly | 49.01 | 43.24 | 1.13 | 0.179 | 0 |
| ClashOfClans | 78.73 | 71.12 | 1.11 | 0.184 | 2 |
| dndnext | 94.53 | 91.07 | 1.04 | 0.186 | 3 |
| Homebrewing | 79.8 | 74.67 | 1.07 | 0.187 | 2 |
| fountainpens | 66.31 | 64.16 | 1.03 | 0.19 | 2 |
| buildapc | 66.77 | 62.5 | 1.07 | 0.192 | 3 |
| Pathfinder_RPG | 97.05 | 93.58 | 1.04 | 0.196 | 3 |
| Rateme | 60.74 | 45.41 | 1.34 | 0.199 | 0 |
| Coffee | 70.76 | 66.47 | 1.06 | 0.201 | 2 |
| MakeupAddiction | 69.2 | 64.46 | 1.07 | 0.213 | 0 |
| Vaping | 73.15 | 66.63 | 1.1 | 0.216 | 2 |
| makinghiphop | 70.47 | 62.22 | 1.13 | 0.218 | 2 |
| SSBM | 84.02 | 77.93 | 1.08 | 0.218 | 3 |
| PuzzleAndDragons | 79.77 | 74.57 | 1.07 | 0.222 | 4 |
| Aquariums | 68.74 | 63.47 | 1.08 | 0.232 | 2 |
| gameswap | 69.99 | 50.77 | 1.38 | 0.236 | 3 |
| dogs | 67.25 | 65.98 | 1.02 | 0.247 | 2 |
| bodyweightfitness | 72.5 | 71.38 | 1.02 | 0.247 | 2 |
| Indiemakeupandmore | 73.19 | 69.11 | 1.06 | 0.257 | 4 |
| vaporents | 69.66 | 64.79 | 1.08 | 0.264 | 2 |
| churning | 75.02 | 72.01 | 1.04 | 0.264 | 2 |
| Animesuggest | 75.84 | 72.22 | 1.05 | 0.272 | 3 |
| HomeImprovement | 78.86 | 76.24 | 1.03 | 0.275 | 2 |
| edmproduction | 70.49 | 67.59 | 1.04 | 0.28 | 0 |
| poker | 80.61 | 74.09 | 1.09 | 0.289 | 2 |
| learnprogramming | 68.05 | 66.95 | 1.02 | 0.29 | 2 |
| yugioh | 90.35 | 83.49 | 1.08 | 0.292 | 3 |
| eu4 | 81.78 | 77.56 | 1.05 | 0.292 | 3 |
| femalefashionadvice | 66.48 | 65.45 | 1.02 | 0.292 | 0 |
| beyondthebump | 69.56 | 68.34 | 1.02 | 0.294 | 4 |
| Watches | 61.97 | 58.26 | 1.06 | 0.297 | 2 |
| DebateReligion | 76.39 | 76.91 | 0.99 | 0.298 | 0 |
| 3Dprinting | 73.61 | 69.77 | 1.06 | 0.299 | 2 |
| headphones | 65.24 | 61.43 | 1.06 | 0.301 | 2 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| frugalmalefashion | 68.61 | 62.57 | 1.1 | 0.305 | 2 |
| ecigclassifieds | 51.95 | 40.7 | 1.28 | 0.316 | 4 |
| Multicopter | 73.95 | 69.85 | 1.06 | 0.316 | 2 |
| goodyearwelt | 66.06 | 63.71 | 1.04 | 0.324 | 2 |
| steroids | 78.23 | 73.67 | 1.06 | 0.326 | 4 |
| WeAreTheMusicMakers | 70.03 | 68.65 | 1.02 | 0.326 | 0 |
| bravefrontier | 86.87 | 80.39 | 1.08 | 0.328 | 4 |
| techsupport | 69.66 | 66.14 | 1.05 | 0.33 | 3 |
| xxfitness | 70.02 | 70.48 | 0.99 | 0.331 | 0 |
| math | 75.52 | 73.93 | 1.02 | 0.335 | 2 |
| rawdenim | 72.29 | 69.83 | 1.04 | 0.335 | 2 |
| weddingplanning | 66.62 | 64.88 | 1.03 | 0.34 | 0 |
| Guitar | 73.45 | 71.53 | 1.03 | 0.34 | 0 |
| worldpowers | 50.61 | 46.38 | 1.09 | 0.342 | 4 |
| jailbreak | 60.47 | 54.55 | 1.11 | 0.345 | 4 |
| csgobetting | 80.89 | 66.01 | 1.23 | 0.346 | 4 |
| DnD | 87.77 | 87.84 | 1 | 0.35 | 3 |
| networking | 81.33 | 79.74 | 1.02 | 0.35 | 2 |
| keto | 68.32 | 66.66 | 1.02 | 0.354 | 0 |
| counting | 11.77 | 3.67 | 3.21 | 0.355 | 5 |
| hardwareswap | 57.3 | 43.39 | 1.32 | 0.355 | 3 |
| electronic_cigarette | 66.98 | 62.44 | 1.07 | 0.357 | 2 |
| magicTCG | 81.65 | 74.11 | 1.1 | 0.36 | 3 |
| hearthstone | 78.36 | 74.29 | 1.05 | 0.361 | 3 |
| pathofexile | 91 | 85.54 | 1.06 | 0.367 | 3 |
| photography | 69.07 | 68.28 | 1.01 | 0.368 | 2 |
| MMORPG | 79.01 | 77.84 | 1.02 | 0.369 | 3 |
| randomactsofcsgo | 41.37 | 26.47 | 1.56 | 0.369 | 4 |
| Boxing | 73.85 | 69.41 | 1.06 | 0.37 | 1 |
| malefashionadvice | 68.62 | 65.47 | 1.05 | 0.377 | 2 |
| Cooking | 82.3 | 77.61 | 1.06 | 0.378 | 2 |
| Diablo | 85.14 | 81.71 | 1.04 | 0.379 | 3 |
| askscience | 40.25 | 35.11 | 1.15 | 0.381 | 5 |
| relationship_advice | 53.94 | 54.56 | 0.99 | 0.382 | 0 |
| loseit | 59.5 | 59.16 | 1.01 | 0.384 | 0 |
| skyrimmods | 75.03 | 71.98 | 1.04 | 0.386 | 3 |
| SSBPM | 81.88 | 77.59 | 1.06 | 0.386 | 3 |
| golf | 77.53 | 74.76 | 1.04 | 0.387 | 2 |
| ar15 | 73.38 | 70.38 | 1.04 | 0.387 | 5 |
| investing | 81.32 | 80.7 | 1.01 | 0.387 | 2 |
| supremeclothing | 85.59 | 67.65 | 1.27 | 0.388 | 4 |
| ADHD | 62.9 | 64.03 | 0.98 | 0.39 | 0 |
| Fitness | 64.81 | 64.27 | 1.01 | 0.39 | 2 |
| chelseafc | 69.95 | 64.93 | 1.08 | 0.39 | 1 |
| Xcom | 92.33 | 89.68 | 1.03 | 0.392 | 3 |
| DeadBedrooms | 62.03 | 63.7 | 0.97 | 0.392 | 0 |
| millionairemakers | 42.31 | 35.32 | 1.2 | 0.392 | 5 |
| heroesofthestorm | 80.23 | 78.53 | 1.02 | 0.398 | 3 |
| photoshopbattles | 30.56 | 26.92 | 1.14 | 0.404 | 5 |
| BabyBumps | 67.72 | 67.66 | 1 | 0.404 | 4 |
| DarkSouls2 | 78.64 | 75 | 1.05 | 0.405 | 3 |
| NHLHUT | 60.6 | 53.07 | 1.14 | 0.406 | 4 |
| buildapcsales | 66.81 | 63.34 | 1.05 | 0.409 | 3 |
| reddevils | 72.75 | 67.73 | 1.07 | 0.409 | 1 |
| woodworking | 73.29 | 70.82 | 1.03 | 0.41 | 2 |
| MechanicalKeyboards | 65.95 | 61.4 | 1.07 | 0.41 | 3 |
| civ | 87.7 | 84.83 | 1.03 | 0.411 | 3 |
| discgolf | 76.52 | 74.18 | 1.03 | 0.412 | 5 |
| LSD | 68.01 | 65.36 | 1.04 | 0.412 | 0 |
| progresspics | 51.02 | 46.98 | 1.09 | 0.415 | 0 |
| stopdrinking | 55.44 | 53.91 | 1.03 | 0.418 | 0 |
| dbz | 70.71 | 69.7 | 1.01 | 0.419 | 3 |
| Twitch | 66.02 | 64.59 | 1.02 | 0.419 | 3 |
| Sneakers | 72.98 | 63.04 | 1.16 | 0.421 | 4 |
| beer | 71.65 | 68.97 | 1.04 | 0.421 | 2 |
| Surface | 70.56 | 69.52 | 1.01 | 0.423 | 2 |
| CrusaderKings | 76.55 | 74.28 | 1.03 | 0.426 | 3 |
| Gunners | 68.4 | 64.23 | 1.06 | 0.428 | 1 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| WorldofTanks | 82.8 | 81.15 | 1.02 | 0.429 | 3 |
| personalfinance | 61.2 | 62.06 | 0.99 | 0.429 | 2 |
| Bitcoin | 75.17 | 73.26 | 1.03 | 0.429 | 1 |
| LiverpoolFC | 72.85 | 68.22 | 1.07 | 0.43 | 1 |
| webdev | 73.46 | 72.34 | 1.02 | 0.432 | 2 |
| Smite | 84.11 | 79.08 | 1.06 | 0.433 | 4 |
| running | 74 | 75.24 | 0.98 | 0.433 | 2 |
| feedthebeast | 81.55 | 77.04 | 1.06 | 0.433 | 3 |
| windowsphone | 77.28 | 74.32 | 1.04 | 0.434 | 2 |
| elderscrollsonline | 80.75 | 79.12 | 1.02 | 0.436 | 3 |
| cscareerquestions | 69.79 | 70.57 | 0.99 | 0.436 | 2 |
| GoneWildPlus | 61.76 | 47.5 | 1.3 | 0.441 | 4 |
| rpg | 87.28 | 88.46 | 0.99 | 0.443 | 3 |
| Naruto | 66.46 | 63.63 | 1.04 | 0.446 | 3 |
| smashbros | 78.81 | 73.38 | 1.07 | 0.447 | 3 |
| philosophy | 71.98 | 73.55 | 0.98 | 0.448 | 1 |
| RandomActsOfGaming | 39.71 | 26.4 | 1.5 | 0.448 | 3 |
| FIFA | 65.16 | 61.11 | 1.07 | 0.451 | 4 |
| eagles | 69.93 | 66.27 | 1.06 | 0.454 | 1 |
| programming | 85.82 | 84.6 | 1.01 | 0.457 | 1 |
| bjj | 74.59 | 74.23 | 1 | 0.457 | 4 |
| vinyl | 69.04 | 67.42 | 1.02 | 0.46 | 2 |
| subaru | 72.5 | 67.33 | 1.08 | 0.461 | 2 |
| MaddenUltimateTeam | 72.05 | 63.72 | 1.13 | 0.462 | 4 |
| asktrp | 70.4 | 71.31 | 0.99 | 0.464 | 0 |
| linux | 79.71 | 77.57 | 1.03 | 0.466 | 1 |
| SchoolIdolFestival | 77.47 | 72.93 | 1.06 | 0.468 | 4 |
| longboarding | 75.86 | 69.68 | 1.09 | 0.468 | 2 |
| darksouls | 74.81 | 72.79 | 1.03 | 0.468 | 3 |
| socialism | 76.18 | 77.14 | 0.99 | 0.469 | 1 |
| zen | 71.52 | 68.15 | 1.05 | 0.47 | 0 |
| gonewild | 62.73 | 50.47 | 1.24 | 0.47 | 4 |
| starbucks | 78.55 | 77.12 | 1.02 | 0.471 | 0 |
| wiiu | 69.32 | 67.77 | 1.02 | 0.472 | 3 |
| gonewildcurvy | 61.23 | 48.92 | 1.25 | 0.473 | 4 |
| vita | 71.64 | 71.51 | 1 | 0.474 | 3 |
| wow | 86.05 | 84.34 | 1.02 | 0.475 | 3 |
| Drugs | 64.08 | 64.16 | 1 | 0.476 | 0 |
| CCW | 69.49 | 70.23 | 0.99 | 0.477 | 5 |
| OnePiece | 70.44 | 68.17 | 1.03 | 0.477 | 3 |
| PoliticalDiscussion | 79.6 | 82.62 | 0.96 | 0.478 | 1 |
| PurplePillDebate | 75.96 | 77.38 | 0.98 | 0.479 | 0 |
| nintendo | 73.99 | 71.66 | 1.03 | 0.48 | 3 |
| gonewildaudio | 59.05 | 51.82 | 1.14 | 0.481 | 4 |
| MonsterHunter | 80.89 | 80.07 | 1.01 | 0.485 | 3 |
| Warthunder | 85.85 | 84.04 | 1.02 | 0.486 | 3 |
| streetwear | 78.19 | 65.36 | 1.2 | 0.487 | 4 |
| relationships | 53.37 | 54.64 | 0.98 | 0.488 | 0 |
| KerbalSpaceProgram | 74.81 | 72.6 | 1.03 | 0.489 | 3 |
| CanadaPolitics | 81.79 | 83.99 | 0.97 | 0.49 | 1 |
| Warhammer40k | 86.64 | 85.07 | 1.02 | 0.49 | 3 |
| iphone | 69 | 66.68 | 1.03 | 0.492 | 2 |
| Economics | 90.45 | 91.44 | 0.99 | 0.492 | 1 |
| coys | 67.54 | 64.03 | 1.05 | 0.493 | 1 |
| vegan | 68.74 | 68.82 | 1 | 0.493 | 0 |
| manga | 72.97 | 69.65 | 1.05 | 0.495 | 4 |
| Metal | 75.65 | 73.47 | 1.03 | 0.495 | 4 |
| leagueoflegends | 76.76 | 72.42 | 1.06 | 0.495 | 4 |
| islam | 69.7 | 69.48 | 1 | 0.496 | 1 |
| Christianity | 72.59 | 73.29 | 0.99 | 0.496 | 1 |
| depression | 47.83 | 49.21 | 0.97 | 0.497 | 0 |
| knifeclub | 60.88 | 58.21 | 1.05 | 0.499 | 4 |
| Music | 67.94 | 65.58 | 1.04 | 0.502 | 5 |
| playrust | 78.58 | 74.51 | 1.05 | 0.504 | 3 |
| SuicideWatch | 41.51 | 42.05 | 0.99 | 0.505 | 0 |
| serialpodcast | 71.76 | 72.58 | 0.99 | 0.505 | 1 |
| NoFap | 62.15 | 61.51 | 1.01 | 0.505 | 0 |
| jobs | 55.59 | 56.72 | 0.98 | 0.505 | 2 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| russia | 74.41 | 73.52 | 1.01 | 0.505 | 1 |
| cars | 67.48 | 66.17 | 1.02 | 0.506 | 2 |
| Philippines | 78.73 | 74.85 | 1.05 | 0.506 | 1 |
| Parenting | 67.68 | 69.58 | 0.97 | 0.51 | 2 |
| Bad_Cop_No_Donut | 78.04 | 77.35 | 1.01 | 0.511 | 1 |
| syriancivilwar | 77.33 | 77.4 | 1 | 0.512 | 1 |
| h1z1 | 77.21 | 73.44 | 1.05 | 0.513 | 3 |
| seduction | 61.67 | 62.26 | 0.99 | 0.517 | 0 |
| truegaming | 73.32 | 75.89 | 0.97 | 0.517 | 3 |
| 3DS | 68 | 67.06 | 1.01 | 0.518 | 3 |
| flying | 77.44 | 77.61 | 1 | 0.522 | 2 |
| apple | 74.05 | 73.43 | 1.01 | 0.523 | 2 |
| exmuslim | 74 | 74.63 | 0.99 | 0.523 | 1 |
| swtor | 80.56 | 81.2 | 0.99 | 0.524 | 3 |
| ffxiv | 79.74 | 80.22 | 0.99 | 0.525 | 3 |
| whowouldwin | 89.85 | 88 | 1.02 | 0.525 | 4 |
| OutreachHPG | 85.19 | 84.9 | 1 | 0.526 | 4 |
| Fantasy | 69.35 | 69.73 | 0.99 | 0.527 | 1 |
| halo | 75.8 | 75.12 | 1.01 | 0.528 | 3 |
| WritingPrompts | 46.13 | 41.44 | 1.11 | 0.528 | 0 |
| ladybonersgw | 55.83 | 44.16 | 1.26 | 0.529 | 4 |
| sex | 56.07 | 58.34 | 0.96 | 0.53 | 0 |
| airsoft | 71.09 | 68.82 | 1.03 | 0.53 | 3 |
| Warframe | 87.6 | 85.95 | 1.02 | 0.53 | 3 |
| nfl | 75.86 | 71.07 | 1.07 | 0.533 | 1 |
| ukpolitics | 79.49 | 80.44 | 0.99 | 0.533 | 1 |
| DCcomics | 73.92 | 73.51 | 1.01 | 0.535 | 1 |
| rugbyunion | 83.39 | 79.28 | 1.05 | 0.535 | 1 |
| motorcycles | 74.3 | 73.9 | 1.01 | 0.536 | 2 |
| CoDCompetitive | 76.44 | 71.77 | 1.07 | 0.536 | 4 |
| indieheads | 84.46 | 82.78 | 1.02 | 0.537 | 1 |
| cordcutters | 74.68 | 73.6 | 1.01 | 0.538 | 2 |
| paradoxplaza | 75.21 | 74.63 | 1.01 | 0.54 | 3 |
| Android | 76.03 | 74.05 | 1.03 | 0.541 | 2 |
| letsplay | 68.08 | 68.02 | 1 | 0.544 | 3 |
| Guildwars2 | 81.19 | 80.72 | 1.01 | 0.544 | 3 |
| sto | 83.26 | 84.11 | 0.99 | 0.545 | 3 |
| Cricket | 89.17 | 82.26 | 1.08 | 0.545 | 1 |
| Anarcho_Capitalism | 83.27 | 84.21 | 0.99 | 0.545 | 1 |
| bodybuilding | 77.16 | 74.22 | 1.04 | 0.546 | 2 |
| minnesotavikings | 71.49 | 69.89 | 1.02 | 0.546 | 1 |
| hiphopheads | 78.14 | 71.33 | 1.1 | 0.547 | 1 |
| soccer | 74.33 | 71.16 | 1.04 | 0.547 | 1 |
| guns | 68.94 | 67.21 | 1.03 | 0.549 | 5 |
| DestinyTheGame | 81.85 | 81.09 | 1.01 | 0.55 | 3 |
| boardgames | 73.89 | 74.86 | 0.99 | 0.551 | 3 |
| formula1 | 75.47 | 72.7 | 1.04 | 0.551 | 1 |
| kpop | 72.65 | 69.91 | 1.04 | 0.553 | 4 |
| sysadmin | 80.94 | 81.57 | 0.99 | 0.554 | 2 |
| AskHistorians | 53.96 | 52.75 | 1.02 | 0.556 | 0 |
| horror | 73.1 | 72.75 | 1 | 0.556 | 1 |
| Justrolledintotheshop | 82.13 | 78.16 | 1.05 | 0.559 | 5 |
| bicycling | 72.55 | 71.81 | 1.01 | 0.559 | 2 |
| cats | 59.55 | 55.74 | 1.07 | 0.561 | 5 |
| politics | 83.72 | 84.83 | 0.99 | 0.561 | 1 |
| Flipping | 70.57 | 69.43 | 1.02 | 0.561 | 2 |
| MMA | 69.95 | 66.04 | 1.06 | 0.562 | 1 |
| Libertarian | 77.87 | 78.87 | 0.99 | 0.563 | 1 |
| neopets | 67.31 | 64.27 | 1.05 | 0.564 | 4 |
| Marvel | 77.97 | 76.78 | 1.02 | 0.57 | 1 |
| DotA2 | 84.24 | 79.15 | 1.06 | 0.573 | 4 |
| survivor | 66.58 | 64.4 | 1.03 | 0.573 | 4 |
| Games | 65.99 | 67.57 | 0.98 | 0.574 | 3 |
| Catholicism | 75.28 | 78.56 | 0.96 | 0.577 | 1 |
| battlefield_4 | 75.42 | 73.96 | 1.02 | 0.577 | 3 |
| DarkNetMarkets | 72.35 | 69.67 | 1.04 | 0.578 | 0 |
| marvelstudios | 77.5 | 76.66 | 1.01 | 0.579 | 1 |
| breakingmom | 68.42 | 69.76 | 0.98 | 0.582 | 4 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| EliteDangerous | 82.32 | 82.86 | 0.99 | 0.584 | 3 |
| dragonage | 73.97 | 76.41 | 0.97 | 0.584 | 3 |
| raisedbynarcissists | 61.4 | 63.01 | 0.97 | 0.585 | 0 |
| starcraft | 79.73 | 77.33 | 1.03 | 0.585 | 3 |
| opiates | 69.82 | 68.43 | 1.02 | 0.586 | 0 |
| amiibo | 67.55 | 63.9 | 1.06 | 0.588 | 4 |
| space | 54.14 | 51.6 | 1.05 | 0.588 | 1 |
| gamedev | 72.03 | 73.98 | 0.97 | 0.589 | 3 |
| EDC | 67.23 | 66.72 | 1.01 | 0.589 | 5 |
| comicbooks | 78.94 | 78.29 | 1.01 | 0.589 | 1 |
| legaladvice | 63.01 | 62.68 | 1.01 | 0.592 | 0 |
| nba | 74.9 | 69.47 | 1.08 | 0.592 | 1 |
| Patriots | 70.71 | 69.9 | 1.01 | 0.593 | 1 |
| worldpolitics | 82.57 | 83.91 | 0.98 | 0.593 | 1 |
| changemyview | 74.27 | 78.29 | 0.95 | 0.594 | 0 |
| Planetside | 80.3 | 79.45 | 1.01 | 0.595 | 3 |
| MensRights | 74.53 | 76.96 | 0.97 | 0.596 | 1 |
| dayz | 69.58 | 68.07 | 1.02 | 0.597 | 3 |
| asktransgender | 60.84 | 63.19 | 0.96 | 0.597 | 0 |
| runescape | 73.83 | 71.15 | 1.04 | 0.598 | 4 |
| books | 65.13 | 66.44 | 0.98 | 0.598 | 1 |
| GameDeals | 63.24 | 62.81 | 1.01 | 0.6 | 3 |
| travel | 61.46 | 61.99 | 0.99 | 0.601 | 2 |
| oculus | 81.27 | 82.68 | 0.98 | 0.603 | 3 |
| DIY | 63.8 | 63.71 | 1 | 0.604 | 2 |
| battlestations | 63.3 | 60.03 | 1.05 | 0.605 | 3 |
| worldbuilding | 83.56 | 85.48 | 0.98 | 0.607 | 0 |
| TheRedPill | 74.87 | 77.57 | 0.97 | 0.608 | 0 |
| anime | 71.45 | 69.63 | 1.03 | 0.608 | 3 |
| bindingofisaac | 77.83 | 73.4 | 1.06 | 0.609 | 3 |
| aviation | 72.42 | 71.11 | 1.02 | 0.61 | 1 |
| osugame | 68.02 | 60.33 | 1.13 | 0.612 | 4 |
| Minecraft | 71.11 | 67.91 | 1.05 | 0.613 | 3 |
| Conservative | 65.03 | 65.92 | 0.99 | 0.615 | 1 |
| pcgaming | 73.15 | 74.15 | 0.99 | 0.616 | 3 |
| Advice | 53.12 | 55.05 | 0.97 | 0.617 | 0 |
| MLS | 75.07 | 72.32 | 1.04 | 0.618 | 1 |
| writing | 70.1 | 73.25 | 0.96 | 0.619 | 0 |
| Filmmakers | 66.7 | 67.55 | 0.99 | 0.619 | 2 |
| xboxone | 69.04 | 68.37 | 1.01 | 0.619 | 3 |
| 2007scape | 74.44 | 69.96 | 1.06 | 0.621 | 4 |
| TrueReddit | 80.07 | 82.89 | 0.97 | 0.629 | 1 |
| Monstercat | 70.18 | 62.12 | 1.13 | 0.631 | 4 |
| skyrim | 71.84 | 70.53 | 1.02 | 0.633 | 3 |
| Eve | 84.15 | 80 | 1.05 | 0.635 | 3 |
| rupaulsdragrace | 73.74 | 69.88 | 1.06 | 0.635 | 4 |
| europe | 81.43 | 83.28 | 0.98 | 0.637 | 1 |
| GlobalOffensive | 73.65 | 69.85 | 1.05 | 0.637 | 4 |
| PS4 | 67.45 | 67.75 | 1 | 0.64 | 3 |
| tf2 | 76.61 | 73.79 | 1.04 | 0.646 | 3 |
| fatlogic | 70.99 | 72.33 | 0.98 | 0.646 | 0 |
| Scotland | 82.95 | 84.4 | 0.98 | 0.647 | 1 |
| asoiaf | 65.9 | 65.32 | 1.01 | 0.65 | 1 |
| paydaytheheist | 77.23 | 76.49 | 1.01 | 0.651 | 3 |
| Anarchism | 77.17 | 78.87 | 0.98 | 0.651 | 1 |
| pcmasterrace | 63.58 | 62.29 | 1.02 | 0.651 | 3 |
| AskScienceFiction | 88.54 | 90.24 | 0.98 | 0.653 | 0 |
| food | 65.9 | 59.79 | 1.1 | 0.653 | 5 |
| atheism | 72.62 | 75.05 | 0.97 | 0.654 | 5 |
| science | 45.72 | 43.24 | 1.06 | 0.655 | 1 |
| ForeverAlone | 55.78 | 58.2 | 0.96 | 0.656 | 0 |
| Silverbugs | 69.6 | 69.52 | 1 | 0.66 | 4 |
| NASCAR | 70.72 | 67.28 | 1.05 | 0.66 | 1 |
| history | 59.83 | 59.44 | 1.01 | 0.66 | 1 |
| cigars | 65.73 | 63.86 | 1.03 | 0.661 | 4 |
| askgaybros | 59.77 | 62.74 | 0.95 | 0.664 | 0 |
| fireemblem | 67.02 | 65.25 | 1.03 | 0.664 | 3 |
| ProgrammerHumor | 73.31 | 69.89 | 1.05 | 0.665 | 5 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| harrypotter | 65.42 | 66.1 | 0.99 | 0.668 | 5 |
| Shitty_Car_Mods | 70.04 | 65.03 | 1.08 | 0.668 | 5 |
| scifi | 72.28 | 72.43 | 1 | 0.669 | 1 |
| gadgets | 63.79 | 64.06 | 1 | 0.669 | 1 |
| starcitizen | 82.51 | 83.11 | 0.99 | 0.67 | 3 |
| gameofthrones | 58.92 | 57.41 | 1.03 | 0.67 | 5 |
| Weakpots | 74.45 | 70.04 | 1.06 | 0.671 | 4 |
| Steam | 70.07 | 69.55 | 1.01 | 0.671 | 3 |
| confession | 51.66 | 54.61 | 0.95 | 0.673 | 0 |
| offmychest | 51.86 | 54.5 | 0.95 | 0.675 | 0 |
| lego | 68.43 | 67.66 | 1.01 | 0.675 | 5 |
| baseball | 70.74 | 68.4 | 1.03 | 0.675 | 1 |
| CFB | 71.18 | 71.17 | 1 | 0.676 | 1 |
| startrek | 71.6 | 70.55 | 1.01 | 0.678 | 5 |
| TheBluePill | 72.1 | 74.67 | 0.97 | 0.681 | 1 |
| StarWars | 66.21 | 67.48 | 0.98 | 0.684 | 5 |
| SquaredCircle | 77.05 | 75.25 | 1.02 | 0.685 | 4 |
| shittyfoodporn | 68.5 | 60.2 | 1.14 | 0.687 | 5 |
| ApocalypseRising | 71.35 | 62.46 | 1.14 | 0.689 | 4 |
| canada | 75.02 | 77.62 | 0.97 | 0.69 | 1 |
| opieandanthony | 73.76 | 70.47 | 1.05 | 0.694 | 4 |
| Futurology | 77.87 | 78.86 | 0.99 | 0.695 | 1 |
| worldnews | 77.54 | 79.36 | 0.98 | 0.696 | 1 |
| Entrepreneur | 65.54 | 66.72 | 0.98 | 0.696 | 2 |
| TwoXChromosomes | 56.24 | 58.87 | 0.96 | 0.698 | 0 |
| pokemon | 63.22 | 61.43 | 1.03 | 0.701 | 3 |
| hockey | 67.11 | 64.73 | 1.04 | 0.702 | 1 |
| fakeid | 63.34 | 56.86 | 1.11 | 0.709 | 4 |
| Frugal | 70.53 | 72.71 | 0.97 | 0.713 | 2 |
| masseffect | 69.64 | 72.89 | 0.96 | 0.717 | 3 |
| unitedkingdom | 80.59 | 81.94 | 0.98 | 0.719 | 1 |
| movies | 71.46 | 72.31 | 0.99 | 0.719 | 1 |
| news | 72.41 | 74.2 | 0.98 | 0.725 | 1 |
| exmormon | 75.92 | 79.35 | 0.96 | 0.726 | 4 |
| actuallesbians | 57.76 | 59.88 | 0.96 | 0.731 | 0 |
| ShitRedditSays | 66.5 | 65.84 | 1.01 | 0.733 | 1 |
| sports | 65.9 | 65.08 | 1.01 | 0.733 | 1 |
| AskMen | 65.65 | 68.32 | 0.96 | 0.735 | 0 |
| MapPorn | 77.93 | 77.93 | 1 | 0.738 | 1 |
| television | 69.36 | 70.26 | 0.99 | 0.74 | 1 |
| australia | 90.21 | 91.61 | 0.98 | 0.741 | 1 |
| AskWomen | 62.65 | 65.26 | 0.96 | 0.743 | 0 |
| circlejerk | 53.48 | 41.55 | 1.29 | 0.743 | 5 |
| Kappa | 74.12 | 67.21 | 1.1 | 0.744 | 4 |
| vancouver | 77.54 | 79.82 | 0.97 | 0.744 | 1 |
| nsfw | 42.5 | 38.8 | 1.1 | 0.744 | 4 |
| fivenightsatfreddys | 60.84 | 55.88 | 1.09 | 0.746 | 4 |
| aww | 63.96 | 59.87 | 1.07 | 0.746 | 5 |
| conspiracy | 78.47 | 80.21 | 0.98 | 0.747 | 1 |
| ultrahardcore | 64.64 | 55.47 | 1.17 | 0.754 | 4 |
| childfree | 65 | 67.36 | 0.97 | 0.756 | 0 |
| lewronggeneration | 74.67 | 70.38 | 1.06 | 0.756 | 5 |
| GamerGhazi | 76.13 | 77.3 | 0.98 | 0.758 | 1 |
| KotakuInAction | 78.84 | 80.99 | 0.97 | 0.76 | 1 |
| GetMotivated | 53.69 | 54.68 | 0.98 | 0.761 | 2 |
| boston | 75.07 | 77.55 | 0.97 | 0.762 | 2 |
| Seattle | 81.3 | 83.34 | 0.98 | 0.764 | 2 |
| Celebs | 48.89 | 45.1 | 1.08 | 0.764 | 1 |
| washingtondc | 74.04 | 76.07 | 0.97 | 0.765 | 2 |
| technology | 76.7 | 78.91 | 0.97 | 0.766 | 1 |
| GrandTheftAutoV | 66.58 | 65.53 | 1.02 | 0.768 | 3 |
| Civcraft | 72.82 | 68.39 | 1.06 | 0.772 | 4 |
| RealGirls | 52.95 | 47.56 | 1.11 | 0.774 | 4 |
| AirForce | 74.36 | 75.44 | 0.99 | 0.774 | 2 |
| gamegrumps | 64.74 | 63.05 | 1.03 | 0.779 | 3 |
| Fallout | 71.83 | 71.08 | 1.01 | 0.783 | 3 |
| rage | 58.16 | 59.24 | 0.98 | 0.785 | 5 |
| exjw | 75.12 | 79.59 | 0.94 | 0.786 | 4 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| OkCupid | 65.09 | 66.74 | 0.98 | 0.787 | 0 |
| JusticePorn | 57.81 | 57.62 | 1 | 0.788 | 5 |
| Tinder | 66.8 | 62.91 | 1.06 | 0.789 | 5 |
| nyc | 74.19 | 75.98 | 0.98 | 0.79 | 1 |
| China | 80.86 | 82.35 | 0.98 | 0.791 | 1 |
| EarthPorn | 48.93 | 47.65 | 1.03 | 0.791 | 5 |
| ProtectAndServe | 68.97 | 71.76 | 0.96 | 0.791 | 2 |
| TumblrInAction | 72.39 | 73.72 | 0.98 | 0.792 | 5 |
| chicago | 70.98 | 72.54 | 0.98 | 0.794 | 2 |
| Denver | 74.52 | 76.23 | 0.98 | 0.795 | 2 |
| talesfromtechsupport | 74.42 | 74.99 | 0.99 | 0.8 | 5 |
| forwardsfromgrandma | 71.97 | 71.87 | 1 | 0.8 | 5 |
| gaming | 67.38 | 67.86 | 0.99 | 0.803 | 3 |
| trees | 72.34 | 69.43 | 1.04 | 0.803 | 0 |
| Documentaries | 65.16 | 67.01 | 0.97 | 0.803 | 1 |
| metalgearsolid | 73.5 | 74.33 | 0.99 | 0.804 | 3 |
| PublicFreakout | 64.42 | 64.08 | 1.01 | 0.804 | 5 |
| offbeat | 73.4 | 76.73 | 0.96 | 0.804 | 1 |
| TwoBestFriendsPlay | 77.9 | 78.3 | 0.99 | 0.805 | 3 |
| LosAngeles | 72.9 | 74.73 | 0.98 | 0.806 | 2 |
| explainlikeimfive | 73.28 | 76.26 | 0.96 | 0.807 | 5 |
| whatisthisthing | 61.1 | 59.35 | 1.03 | 0.808 | 5 |
| nottheonion | 67.33 | 68.82 | 0.98 | 0.812 | 5 |
| Austin | 78.35 | 81.43 | 0.96 | 0.812 | 2 |
| army | 74.93 | 75.26 | 1 | 0.814 | 2 |
| SubredditDrama | 69.59 | 70.97 | 0.98 | 0.815 | 1 |
| weekendgunnit | 67.8 | 60.63 | 1.12 | 0.816 | 4 |
| HistoryPorn | 59.95 | 59.49 | 1.01 | 0.816 | 1 |
| toronto | 72.25 | 74.51 | 0.97 | 0.817 | 1 |
| dataisbeautiful | 67.53 | 69.83 | 0.97 | 0.817 | 1 |
| polandball | 76.38 | 74.13 | 1.03 | 0.818 | 4 |
| philadelphia | 74.4 | 76.32 | 0.97 | 0.819 | 2 |
| ireland | 81.86 | 82.64 | 0.99 | 0.82 | 1 |
| london | 78.96 | 79.78 | 0.99 | 0.82 | 1 |
| Whatcouldgowrong | 65.73 | 63.96 | 1.03 | 0.82 | 5 |
| india | 82.2 | 81.55 | 1.01 | 0.824 | 1 |
| TrollXChromosomes | 62.66 | 65.22 | 0.96 | 0.825 | 0 |
| furry | 68.32 | 67.13 | 1.02 | 0.827 | 4 |
| sydney | 79.44 | 80.22 | 0.99 | 0.828 | 2 |
| Random_Acts_Of_Amazon | 58.85 | 56.22 | 1.05 | 0.828 | 4 |
| ottawa | 70.3 | 71.88 | 0.98 | 0.828 | 1 |
| watchpeopledie | 61.89 | 60.44 | 1.02 | 0.829 | 5 |
| trashy | 61.95 | 59.75 | 1.04 | 0.829 | 5 |
| BlackPeopleTwitter | 68.08 | 63.08 | 1.08 | 0.831 | 5 |
| Art | 47.23 | 47.1 | 1 | 0.831 | 1 |
| Portland | 79.22 | 81.43 | 0.97 | 0.831 | 2 |
| Atlanta | 68.99 | 70.68 | 0.98 | 0.833 | 2 |
| Calgary | 77.45 | 80.48 | 0.96 | 0.834 | 2 |
| houston | 75.12 | 76.18 | 0.99 | 0.834 | 2 |
| creepyPMs | 53.32 | 53.87 | 0.99 | 0.835 | 0 |
| TalesFromRetail | 61.27 | 63.24 | 0.97 | 0.838 | 0 |
| justneckbeardthings | 68.51 | 66.51 | 1.03 | 0.839 | 5 |
| bestof | 55.58 | 56.8 | 0.98 | 0.842 | 5 |
| Military | 73.37 | 74.32 | 0.99 | 0.843 | 1 |
| self | 60.25 | 63.57 | 0.95 | 0.843 | 0 |
| tipofmytongue | 56.53 | 52.54 | 1.08 | 0.843 | 5 |
| shittyaskscience | 80.42 | 79.24 | 1.01 | 0.845 | 5 |
| cringepics | 53.84 | 53.22 | 1.01 | 0.848 | 5 |
| cringe | 57.16 | 56.38 | 1.01 | 0.848 | 5 |
| Wishlist | 55.21 | 52.58 | 1.05 | 0.853 | 4 |
| 4chan | 68.54 | 61.32 | 1.12 | 0.854 | 5 |
| OldSchoolCool | 58.75 | 57.78 | 1.02 | 0.856 | 5 |
| roosterteeth | 63.23 | 63.99 | 0.99 | 0.856 | 3 |
| UpliftingNews | 58.52 | 60.32 | 0.97 | 0.861 | 1 |
| iamverysmart | 66.58 | 66.41 | 1 | 0.862 | 5 |
| teenagers | 67.5 | 65.1 | 1.04 | 0.862 | 4 |
| fireemblemcasual | 64.14 | 62.75 | 1.02 | 0.865 | 4 |
| melbourne | 75.59 | 76.25 | 0.99 | 0.868 | 2 |

| Subredddit | baseline $\mathsf{Ppl}_{M_j}$ | CCLM $\mathsf{Ppl}_{M_j}$ | $\mathsf{IG}_{M_j}$ | $\mathsf{Ind}_{M_j}$ | Social embed. cluster |
|---|---|---|---|---|---|
| newzealand | 79.58 | 82.61 | 0.96 | 0.868 | 1 |
| thatHappened | 69.04 | 68.48 | 1.01 | 0.868 | 5 |
| ImGoingToHellForThis | 56.18 | 53.73 | 1.05 | 0.868 | 5 |
| gaybros | 71.02 | 75.55 | 0.94 | 0.872 | 0 |
| RWBY | 69.44 | 70.03 | 0.99 | 0.874 | 4 |
| LifeProTips | 64.54 | 65.85 | 0.98 | 0.875 | 5 |
| OutOfTheLoop | 60.83 | 64.38 | 0.94 | 0.875 | 5 |
| WTF | 69.05 | 67.8 | 1.02 | 0.876 | 5 |
| AMA | 61.41 | 63.51 | 0.97 | 0.876 | 0 |
| Unexpected | 59.3 | 57.25 | 1.04 | 0.876 | 5 |
| nosleep | 57.84 | 58 | 1 | 0.876 | 0 |
| facepalm | 64.59 | 65.56 | 0.99 | 0.877 | 5 |
| todayilearned | 77.6 | 78.75 | 0.99 | 0.878 | 5 |
| rva | 68.82 | 71.47 | 0.96 | 0.879 | 2 |
| CasualConversation | 62.58 | 64.02 | 0.98 | 0.88 | 0 |
| tifu | 64.84 | 65.06 | 1 | 0.88 | 5 |
| oddlysatisfying | 62.89 | 60.87 | 1.03 | 0.882 | 5 |
| mylittlepony | 64.29 | 63.91 | 1.01 | 0.883 | 4 |
| videos | 63.11 | 63.56 | 0.99 | 0.884 | 5 |
| woahdude | 63.53 | 61.9 | 1.03 | 0.885 | 5 |
| gifs | 63.96 | 62.03 | 1.03 | 0.886 | 5 |
| creepy | 61.21 | 59.49 | 1.03 | 0.886 | 5 |
| Jokes | 62.37 | 58.55 | 1.07 | 0.889 | 5 |
| AdviceAnimals | 66.09 | 69.02 | 0.96 | 0.89 | 5 |
| mildlyinteresting | 69.17 | 67.41 | 1.03 | 0.891 | 5 |
| casualiama | 60.5 | 62.03 | 0.98 | 0.891 | 0 |
| NoStupidQuestions | 66.89 | 69.53 | 0.96 | 0.893 | 0 |
| interestingasfuck | 63.16 | 61.9 | 1.02 | 0.897 | 5 |
| CrappyDesign | 66.41 | 66.1 | 1 | 0.901 | 5 |
| pics | 65.78 | 65.55 | 1 | 0.901 | 5 |
| britishproblems | 76.34 | 77.55 | 0.98 | 0.902 | 5 |
| funny | 62.25 | 61.28 | 1.02 | 0.902 | 5 |
| mildlyinfuriating | 67.46 | 67.37 | 1 | 0.906 | 5 |
| CFBOffTopic | 70.55 | 72.98 | 0.97 | 0.908 | 1 |
| reactiongifs | 54.47 | 54.14 | 1.01 | 0.909 | 5 |
| singapore | 81.94 | 85.02 | 0.96 | 0.912 | 2 |
| AskReddit | 74.3 | 75.72 | 0.98 | 0.913 | 5 |
| MLPLounge | 54.02 | 51.92 | 1.04 | 0.913 | 4 |
| InternetIsBeautiful | 64.82 | 65.13 | 1 | 0.914 | 1 |
| Showerthoughts | 71.29 | 69.45 | 1.03 | 0.918 | 5 |
| IAmA | 65.06 | 68.55 | 0.95 | 0.919 | 5 |