Stanford
GRADUATE SCHOOL OF
EDUCATION

Modeling Language Games

Chris Proctor cproctor@stanford.edu
Veronica Lin vronlin@stanford.edu

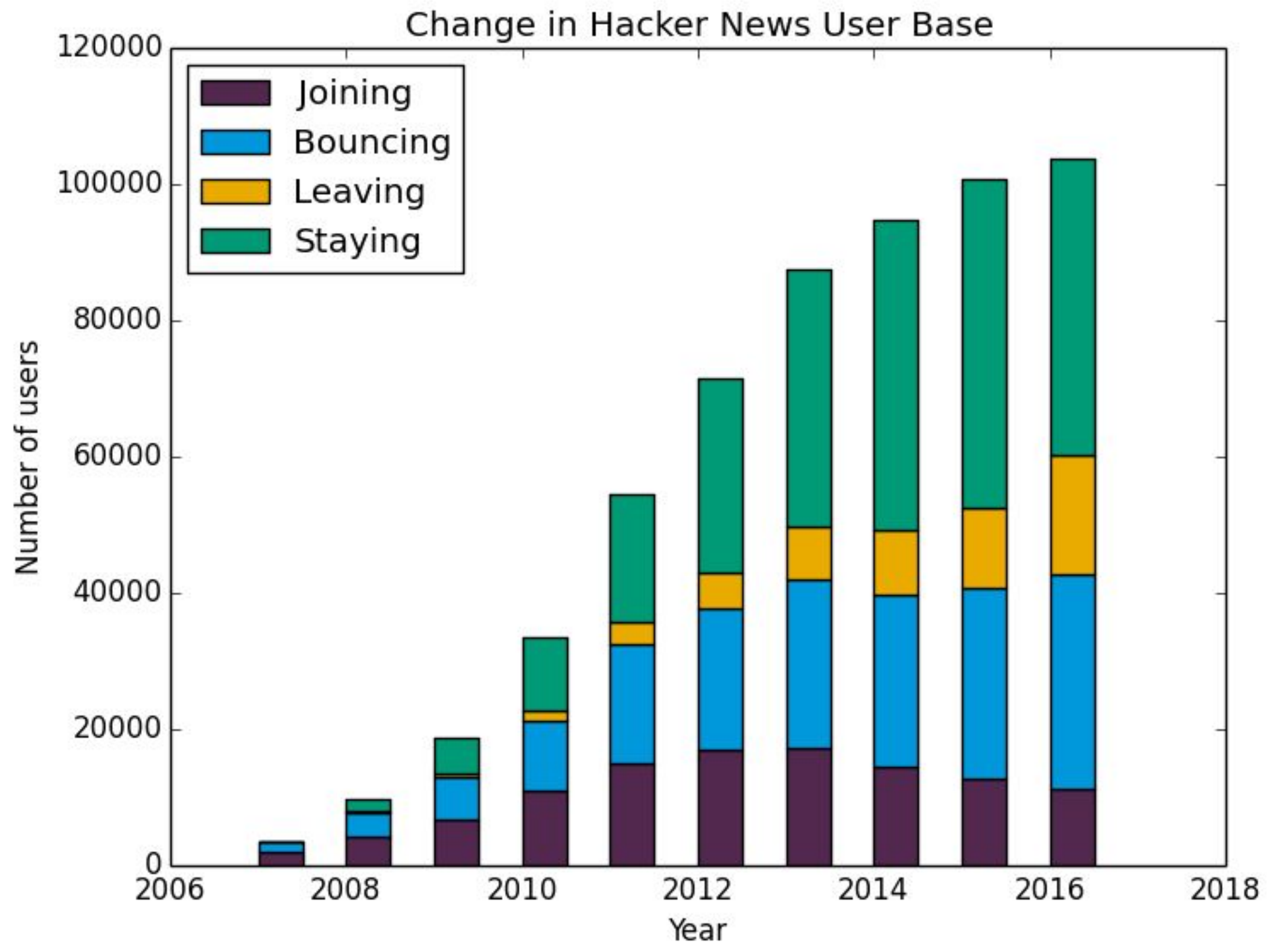
Background

A central belief of the learning sciences is that learning is situated within communities of practice (Lave & Wenger, 1991). Word embedding models, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), frame word meaning as fixed and global; however, learning theories view word meaning as dynamically shaped. This project proposes a method for modeling the semantic content of: 1) participants' trajectories of participation and 2) change in the community's language norms.

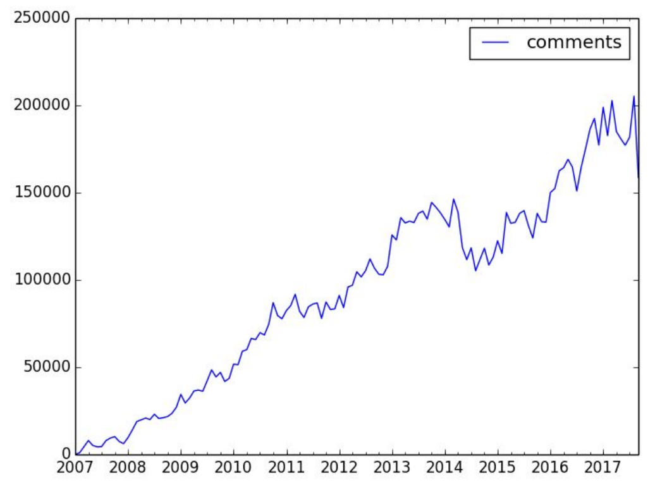
Data

Hacker News is a discussion forum affiliated with the Bay Area startup incubator Y Combinator. The site is organized as a list of posts, each of which has an attached comments thread. Within a thread, users are considered connected.

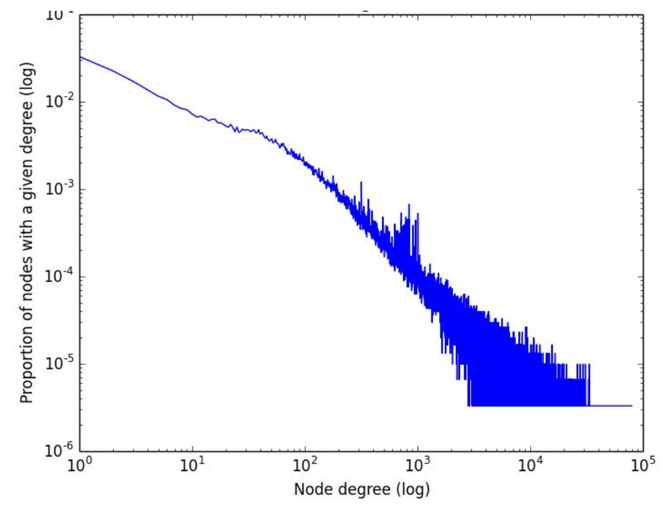
12m posts from 2007 to 2017
315k users (31k with > 50 posts)
Median 2 sentences per post (std: 3)



Changes in user base



Monthly comments



User degree distribution

Trajectories of participation

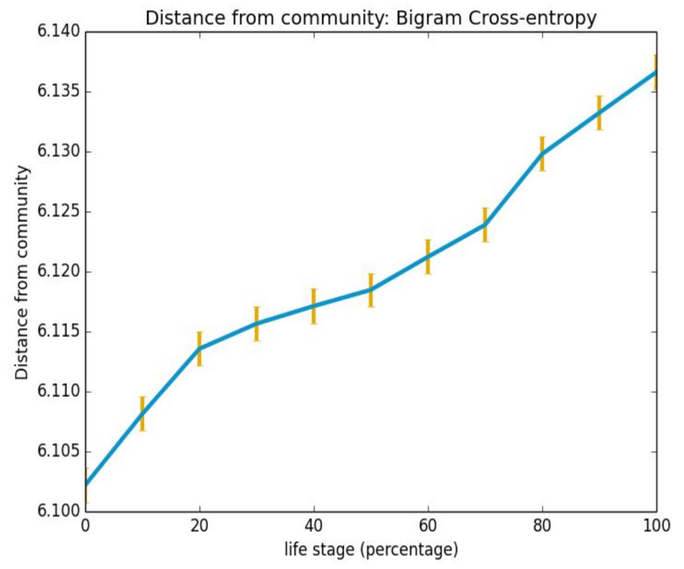
Methods

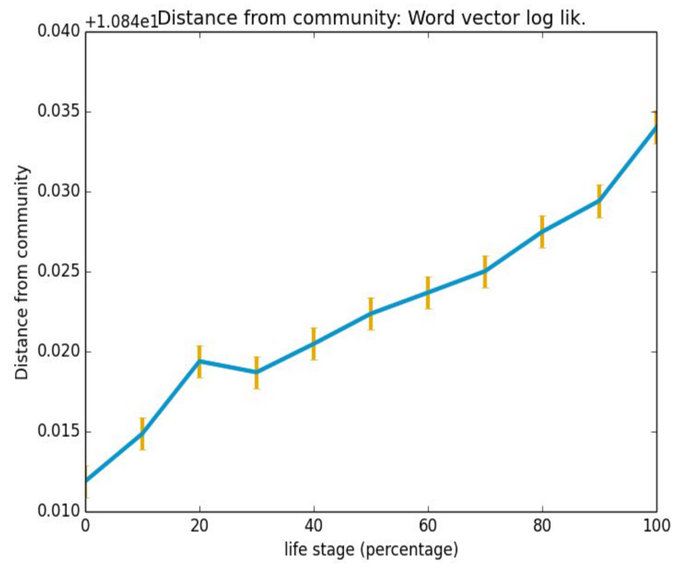
Given a user's first 20 posts, the task is to predict if the user will leave the community (<50 posts) or stick around (>200 posts). Features for the logistic regression model (language models used monthly snapshots):

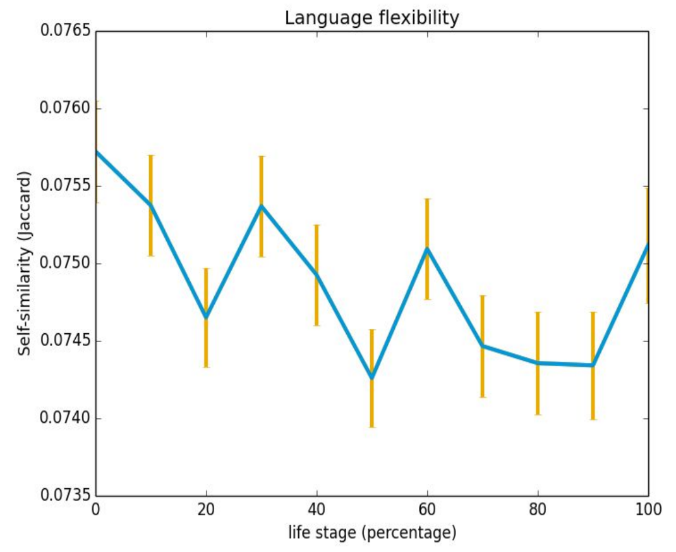
Baseline features (post frequency and month)
Bigram (using cross-entropy to evaluate posts)
Word2Vec (using log-likelihood to evaluate posts)

Results

Data set	Features	Precision	Recall	F_1	Departed	Living
RateBeer	Activity	0.737	0.193	0.305	261	465
RateBeer	Activity + BigramCE			0.374	261	465
HackerNews	Activity	0.769	0.803	0.786	1977	1602
HackerNews	Activity + BigramCE	0.770	0.805	0.787	1977	1602
HackerNews	Activity + WordVectorLL	0.768	0.804	0.785	1977	1602
HackerNews	Activity + DiffLL	0.771	0.807	0.788	1977	1602
HackerNews	Activity + WordVectorLL + DiffLL	0.769	0.805	0.787	1977	1602







Discussion

In contrast to Danescu-Niculescu-Mizil et al. (2013), new Hacker News users do not experience language adaptation; their language appears to have already stabilized.

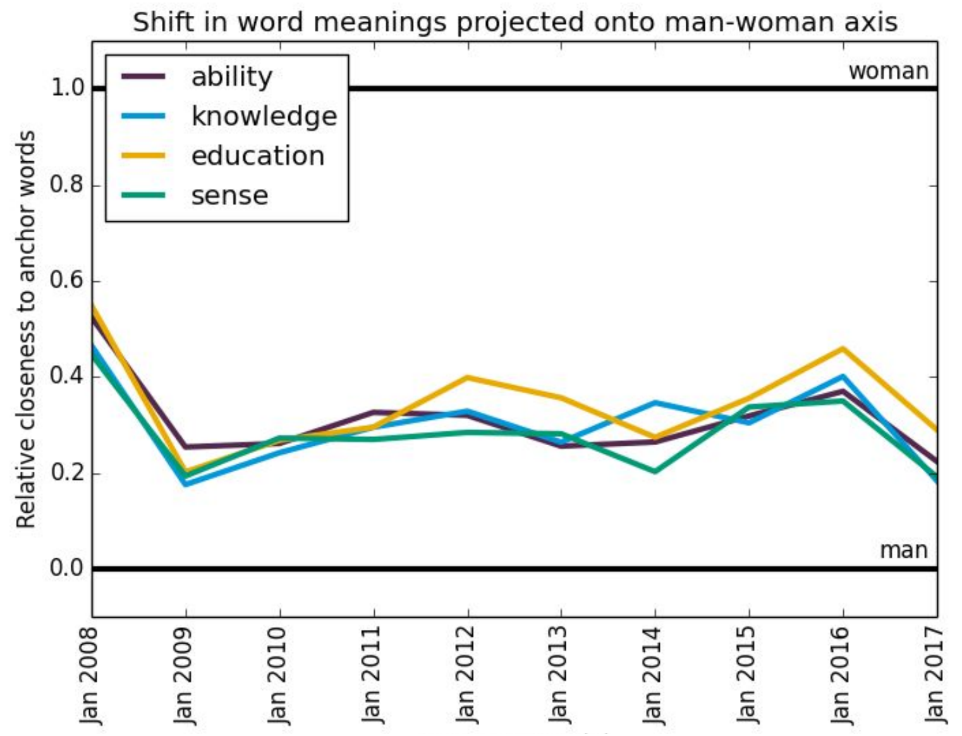
Change in community norms

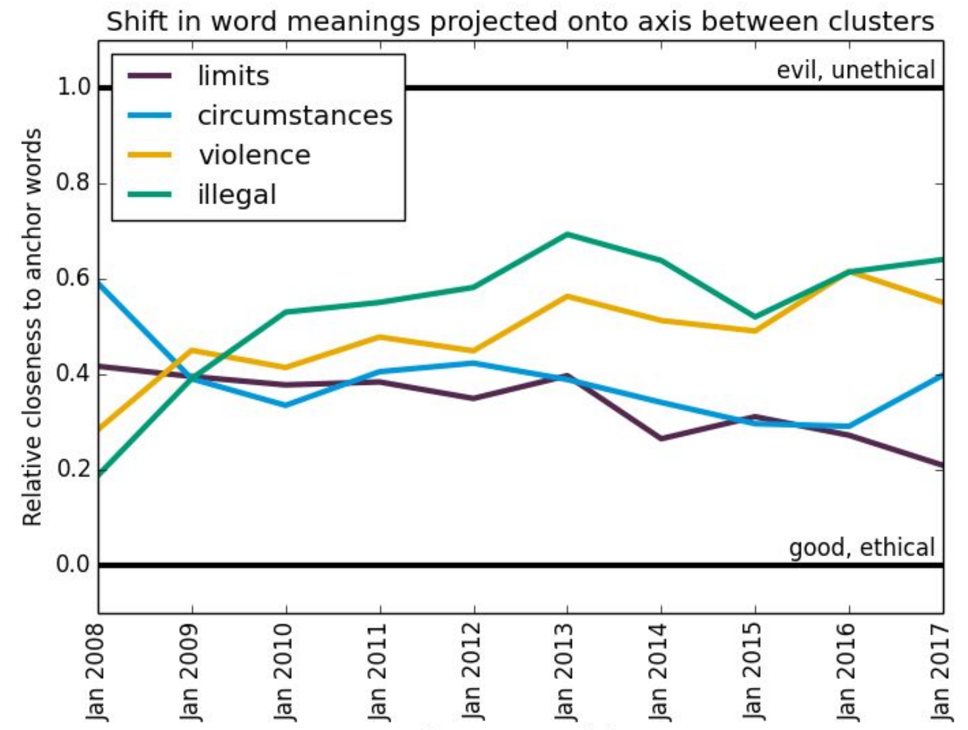
Methods

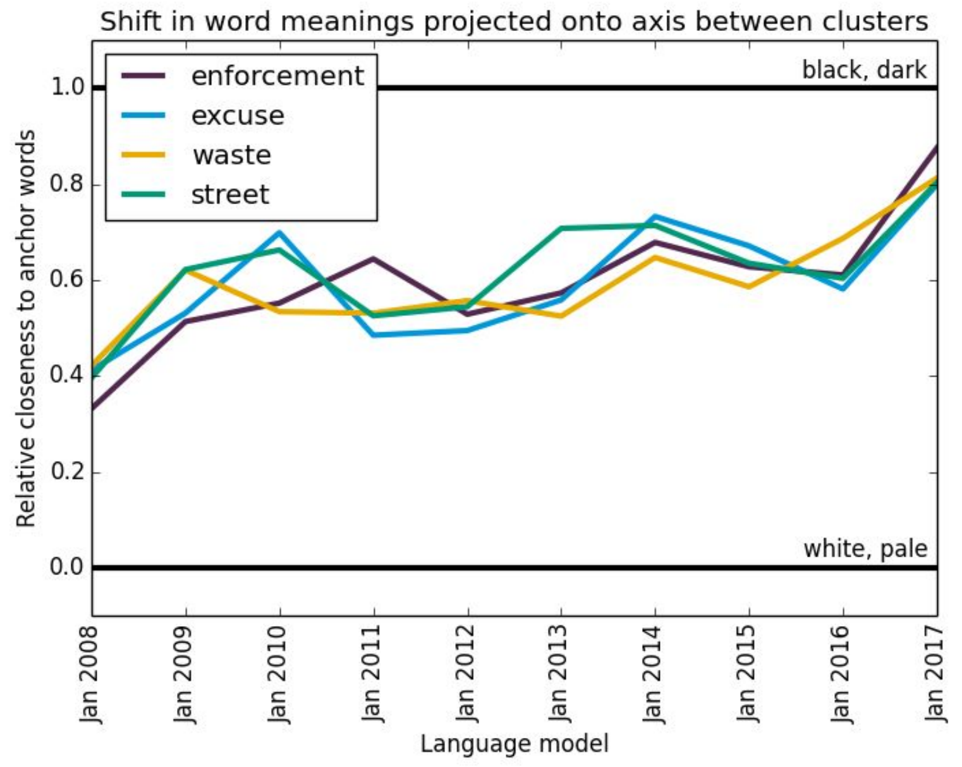
To characterize how word meanings change over time, word vectors are projected onto a normalized relational axis between two points E_0 and E_1:

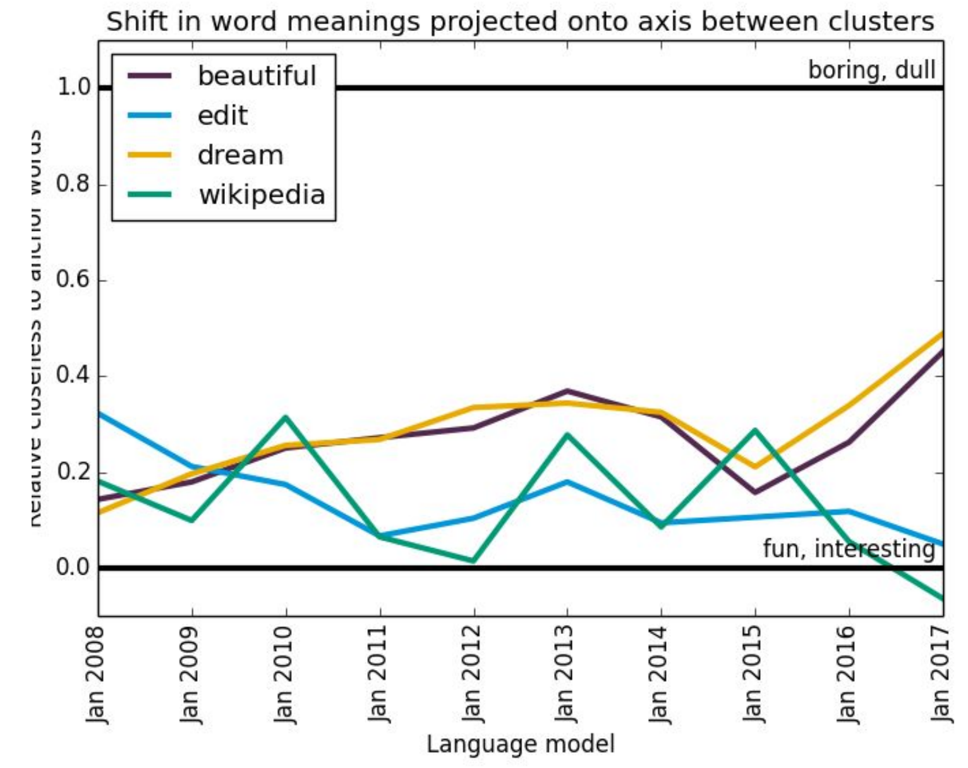
$$|Projection| = \frac{(E_{word} - E_0) \cdot (E_1 - E_0)}{(E_1 - E_0) \cdot (E_1 - E_0)}$$

Results









Discussion

Within the community's word meanings, both community perspectives and racist / sexist biases change over time.

Future Work

Further research could apply these methods to in-person discourse communities, and use the orientation of users' language on relational axes of gender and other identity categories to predict their future participation trajectories.

References

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (May). No country for old members: User lifecycle and linguistic change in online communities. In *Proc. of the 22nd international conference on World Wide Web* (pp. 307-318). ACM.
Lave, J., & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge university press.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).