Introduction to Social Media Analytics (Lec 8)

Hao PENG
Department of Data Science
City University of Hong Kong
https://haoopeng.github.io/

Agenda for this week

- Engagement analysis & prediction
- Macro-level user studies
  - Individual & collective behavior
  - adoption, retention, churn prediction
- Community, polarization
  - Echo Chambers
  - Power of diverse / polarized teams
- Personalization & Recommendation

Action analytics

- **Social media actions analytics** deals with extraction, analysis, and interpretation of insights contained in actions performed by social media users.
- Actions on SM are easy and fast way to **express feelings**, unlike written reaction. They can be considered as **social expressions and symbolic reactions** to social media content.
- Actions are the **cash cow/flow** of social media companies.

Can be used to measure **popularity and influence (not quality)** of a product, service, or idea on social media.

Common social media actions

- **Like buttons**
- **Dislike**
- **Share**
- **Visitors, Visits, Revisits**
- **Views**
- **Clicks**
- **Tagging**

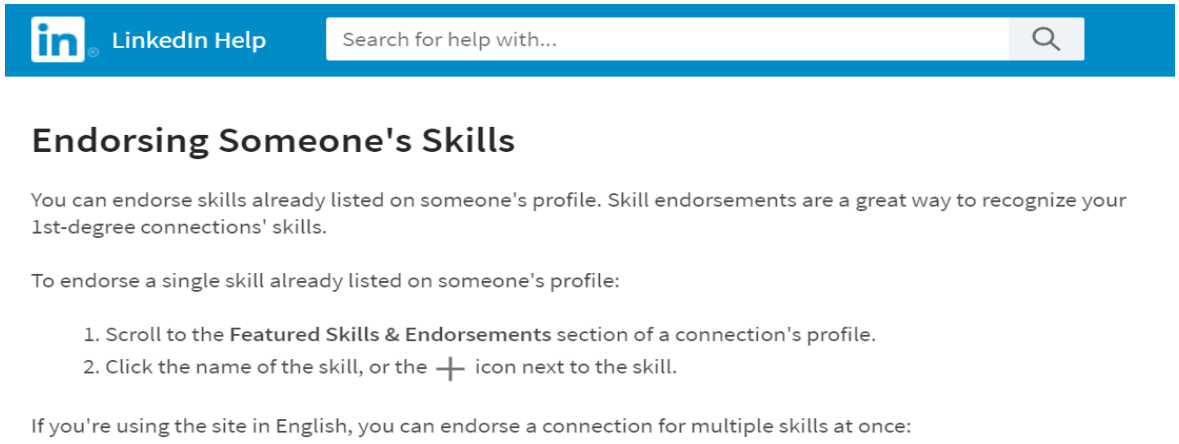
- **Mentions**
- **Hovering**
- **Check-in**
- **Pinning**
- **Embeds**
- **Endorsement**
- **Downloading**

Common social media actions

- **Mentions**
  - **Mentions** or social mentions are the occurrence of a person, place, or thing over social media by name.
  - **Mentions indicate popularity of person, product, place.**
- **Check-in**
  - **Check-in** is a feature that allows users to announce and share their arrival at a location, such as a city, hotel.
  - Rich check-in data can be mined to offer location-based services and products.

Common social media actions

- **Endorsement**
  - **Endorsement** is a features that lets people to endorse, rate, and approve other people, products, and services.
  - **Common types of endorsement:** product reviews



Common social media actions

- **Tagging**
  - **Tagging** is the act of assigning or linking extra pieces of information to social media content for **identification**, **classification**, and **search purposes.**

- **Pinning**
  - **Pinning** is an action performed by social media users to pin and share interesting content (such as product, idea, services, and information) using a virtual pin-board.

Case: predicting papers' news coverage

Media coverage is one of the most important *alternative* metrics (besides traditional metrics such as citations) for measuring the popularity of academic papers.

Case: predicting papers' news coverage

Media coverage is one of the most important *alternative* metrics (besides traditional metrics such as **citations**) for measuring the popularity of academic papers.

- Both metrics are field-dependent (e.g., math < biomedical).
- Both metrics only partly reflect (not equal to) quality/merits.
- Timing difference: citation counts are often delayed (sleeping beauty effect), whereas news often happens in a few weeks.
- Audience difference: papers are mostly cited by academics, whereas news reports are often consumed by the public.
- **Relative to citations, news coverage is much less understood.**

Case: predicting papers' news coverage

Motivation & significance

- Media reports amplify a paper's impact beyond the academy & disseminate the latest breakthroughs & findings to the public.
- We know little about how linguistic features alone can affect a paper's chance of news coverage beyond its content quality.
- Understanding this question can avoid the situation where truly newsworthy papers lose their attention to less-deserving ones.
  - Institution reputation (e.g., Harvard vs. CityU)
  - Author seniority (different ranks of professor, etc.)
  - Journal impact, content quality, etc.

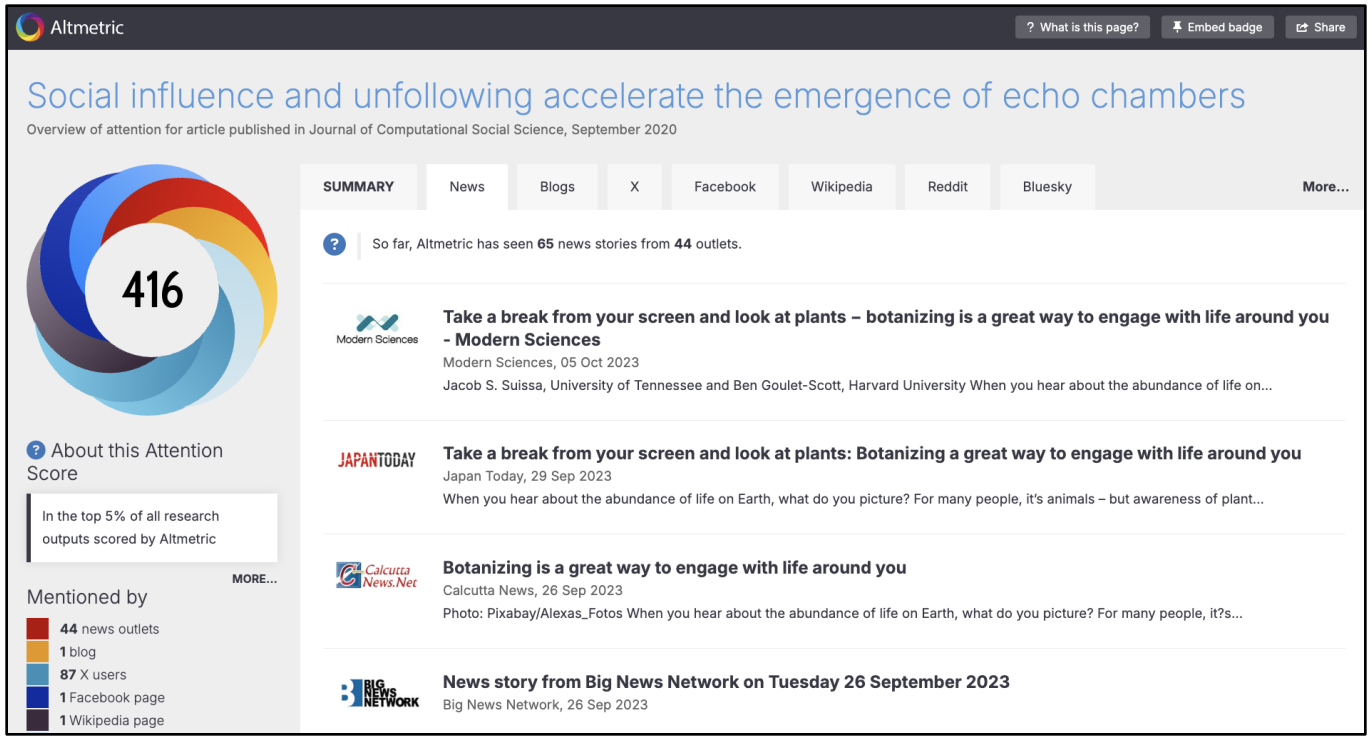
Research Q: To what degree can simple linguistic features predict news coverage above and beyond an idea's content quality?

Case: predicting papers' news coverage

Data & methods

https://www.altmetric.com/details/60364782/news

- Complete news mentions of 2.8M papers from Altmetric in 6y.
- Detailed paper metadata (such as abstract) from OpenAlex.



Case: predicting papers' news coverage

Regression models to predict coverage:

- DV1 = coverage or not (logistic regression)
- DV2 = num. of news coverage (linear regression)
- Key IV / predictors (computed based on the abstract):
  - Sentiment? Emotion?
  - LIWC? (such as first-person plural pronouns "WE")
  - Promotional words?
  - Hedge words? (e.g., "high GPA may not predict future career success")
- Control variables:
  - Content-level factors (other linguistic variables: readability)
  - Author-level factors (reputation, productivity, impact, etc.)
  - Fixed-effects factors (year, research topics)

Case: predicting papers' news coverage

Key findings & takeaways

- "We" words frequency (we|us|our) predicts coverage
- Negative emotion is associated with more coverage
- Hype words (e.g., "groundbreaking") predicts coverage
- Hedge words density increases likelihood of coverage

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-10.5480	0.2397	0.0000
hedges_proportion	2.4727	0.1857	0.0000
promotion_ratio	3.5407	0.3206	0.0000
subjectivity	-0.0109	0.0261	0.6768
concrete	1.0811	0.0149	0.0000
TypeTokenRatio	5.1617	0.2377	0.0000
negemo_percentage	6.5930	0.2167	0.0000
words_per_sentence	0.0079	0.0010	0.0000
we_percentage	8.8234	0.3288	0.0000
readability	-0.0133	0.0015	0.0000

Case: predicting papers' news coverage

Implications

- The pen is as important as the sword: Linguistic "packaging" does affect a paper's marketing success beyond its content.
- Pay close attention to your presentation / communication style as they truly matter in today's attention economy.
  - Sometimes, the quality perceived by your audience matters more than the inherent quality of your content.
  - Valuable lesson for content creators, marketers, influencers.

Limitations

- Correlational analysis does not prove causality (recall the "brevity" paper)
- Focused on English-language news outlets (effect may differ for other lang.)

Case: predicting papers' news coverage

Future work

- Experimentally test the causal effect of language on attention
  - We can tightly control for (**unobservable**) author/content effect
  - Can you give an example of unobserved factors?
- What are the causal (if any) cognitive mechanisms behind the power of language in news coverage?
  - "We" words -> signal teamwork -> perceived credibility
  - Hype words, negative emotions -> capture attention (which is scarce)
  - Uncertain words -> adds accuracy to findings -> perceived quality

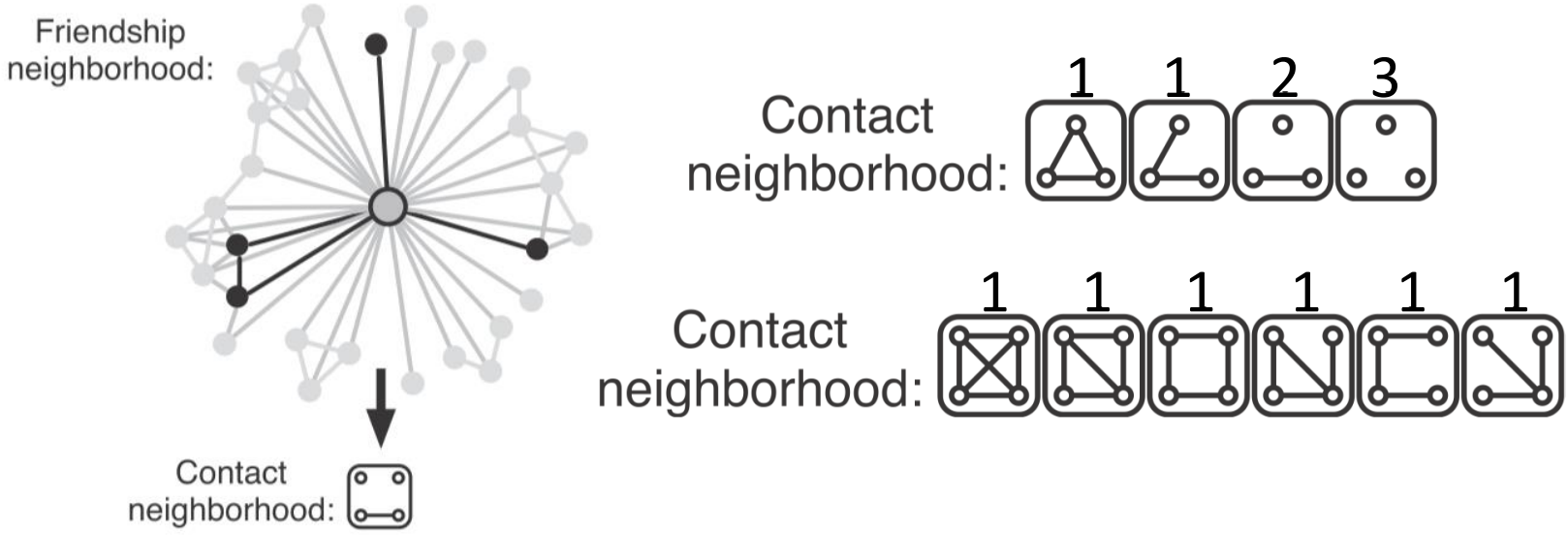
Important practice question: How can we design an experiment to (i) examine causal effect and (ii) explore cognitive mechanisms?

Macro-level user analysis

The study of large-scale patterns, trends, and dynamics in user behavior, interests, and communities across online platforms.

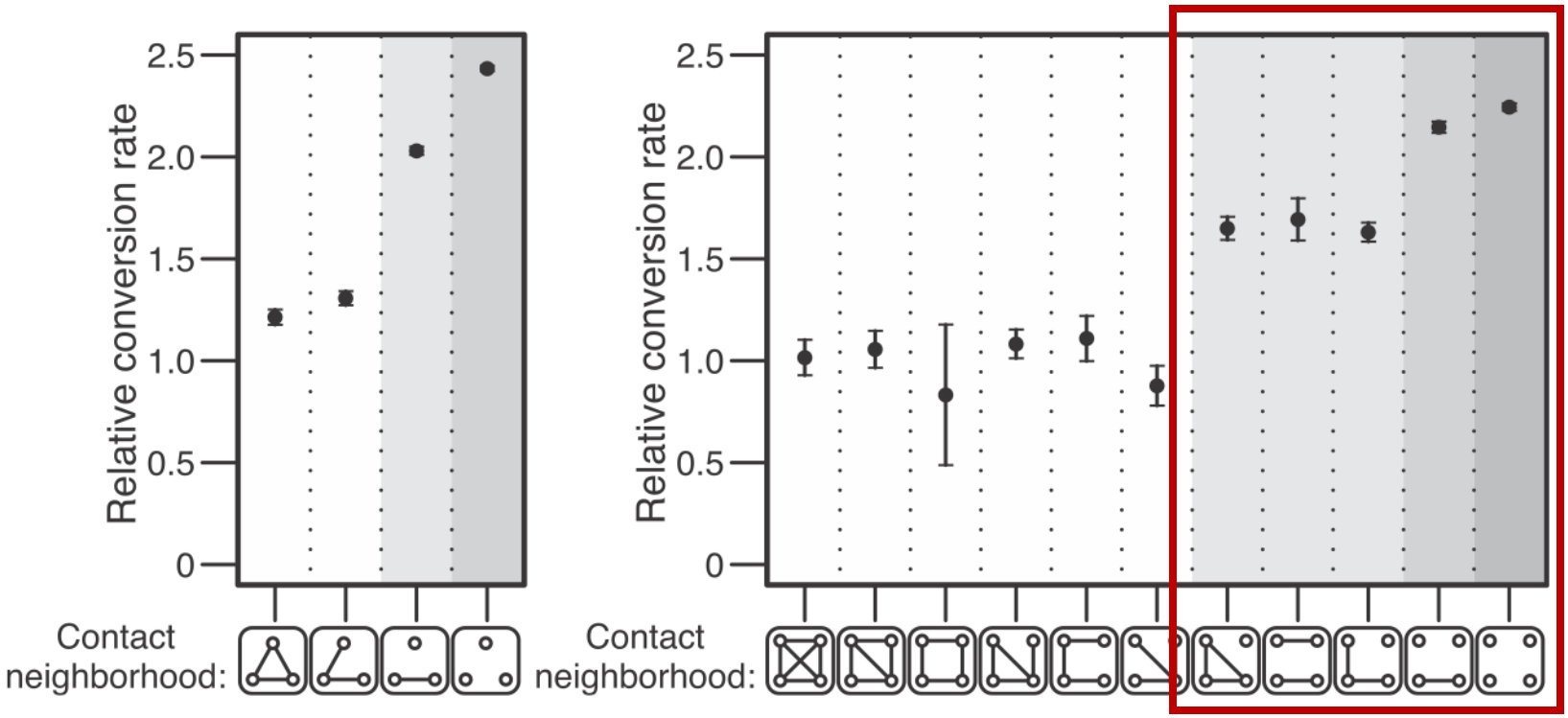
- Examines aggregated data to understand individual or collective behaviors, their evolving interests, and group dynamics within populations.
- It leverages big data techniques to identify trends, predict shifts, and inform strategies at a platform or societal level.
- **Individual user behavior**: adoption/churn prediction, self-promotion, etc.
- **Evolution of user interests**: Tracks how user interests (e.g., topics, hashtags) & behavior evolve over time, influenced by events, trends, or cultural shifts.
- **Community dynamics**: Analyzes the formation, growth, and evolution of user communities based on shared interests, affiliations, or interactions. It examines how users cluster into groups and how these groups shift.

How to increase adoption on FB?



- Grey links: based on email contacts of FB users (only visible to FB)
- Black links: invitations sent by target's FB friends (visible to target)
- Black nodes: Num of invitations vs Num of connected components
  [Which matters more?] Can you make predictions?

Connectivity diversity > network size



Network structure of how one's neighbors are connected predicts the adoption rate more accurately than the size of neighbors.

Ugander et al. "Structural diversity in social contagion." *PNAS* 2012.

Case study: scholarly self-promotion

Motivation & Significance

- Enhances research visibility: Social media platforms can amplify the reach of academic work, making it accessible to the broader audiences beyond traditional academic circles.
- Builds professional networks: Self-promotion on social media can facilitate collaborations with peers and institutions.
- Influences funding & visibility: Online presence can attract grants and recognition by showcasing impact to stakeholders.
- Scholarly marketing companies can reach potential customers by identifying scholars who are interested in boosting visibility.
  - E.g. https://paperspotlight.com

Case study: scholarly self-promotion

Research Questions

- How often do scientists involve in self-promotion?
- Does it vary by year, discipline, and author impact?
  - Authorship position (first, middle, last author)
  - Author career stage (num. of publications)
  - Journal prestige and institution reputation
- Are there gender differences in self-promotion?
  - Difference in self-promotion rates?
  - Gap in engagement with self-promotion?
  - Why is the gender gap important for us to understand?

Case study: scholarly self-promotion

Data & methods https://www.altmetric.com/details/60364782/twitter

- Complete Twitter mentions of 2.8M papers from Altmetric.
- Detailed paper metadata (such as authors) from OpenAlex.

Case study: scholarly self-promotion

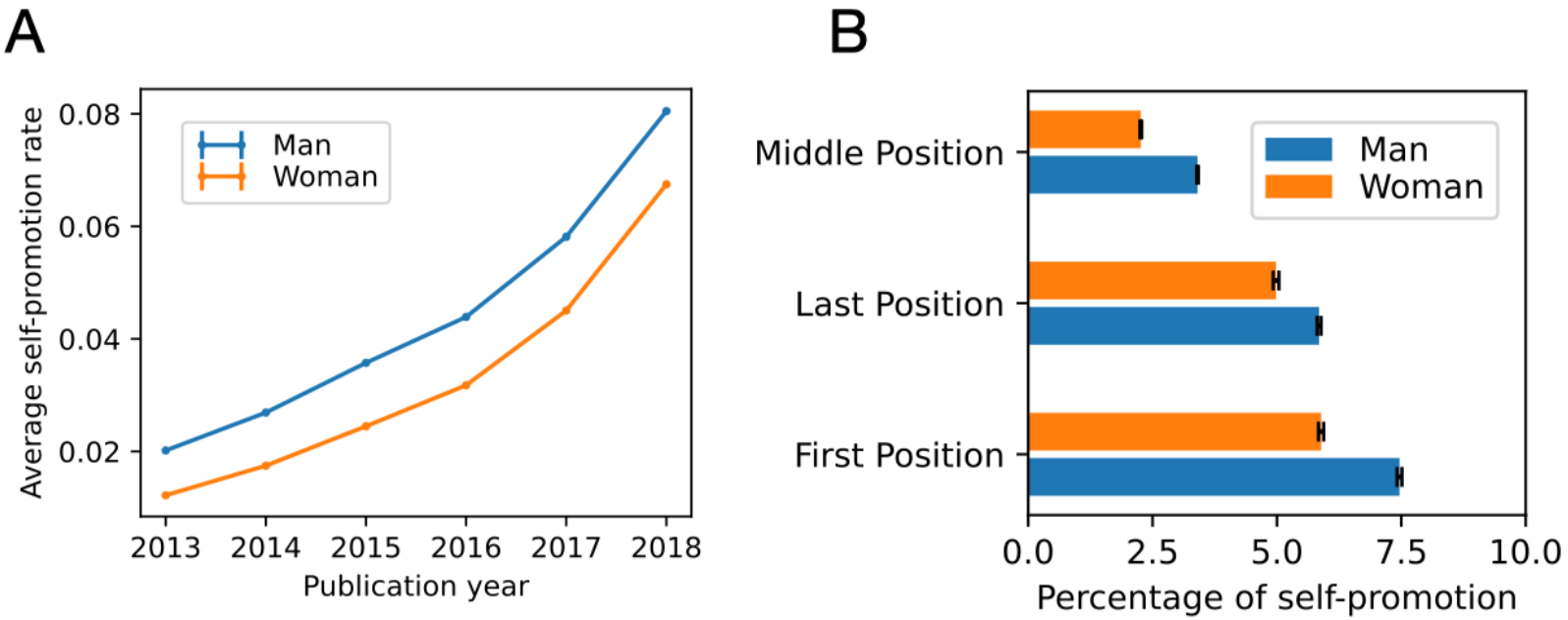
Data & methods

- How to detect if an author self-promoted their paper?
  - Token-matching: first or last name should be matched to the tokens of tweet names (split by space or underscore).
  - Containment-matching: only if the tweet names are single-token strings and the first/last name has >= four characters.
- How to predict author gender?
  - Should we use self-identities or inferred ones?
  - How to estimate error rates for name-based classifiers?
  - Does the error rate differ by ethnicity?
- How to use regressions to examine the gender gap?
  - DV, IV, Controls

Case study: scholarly self-promotion

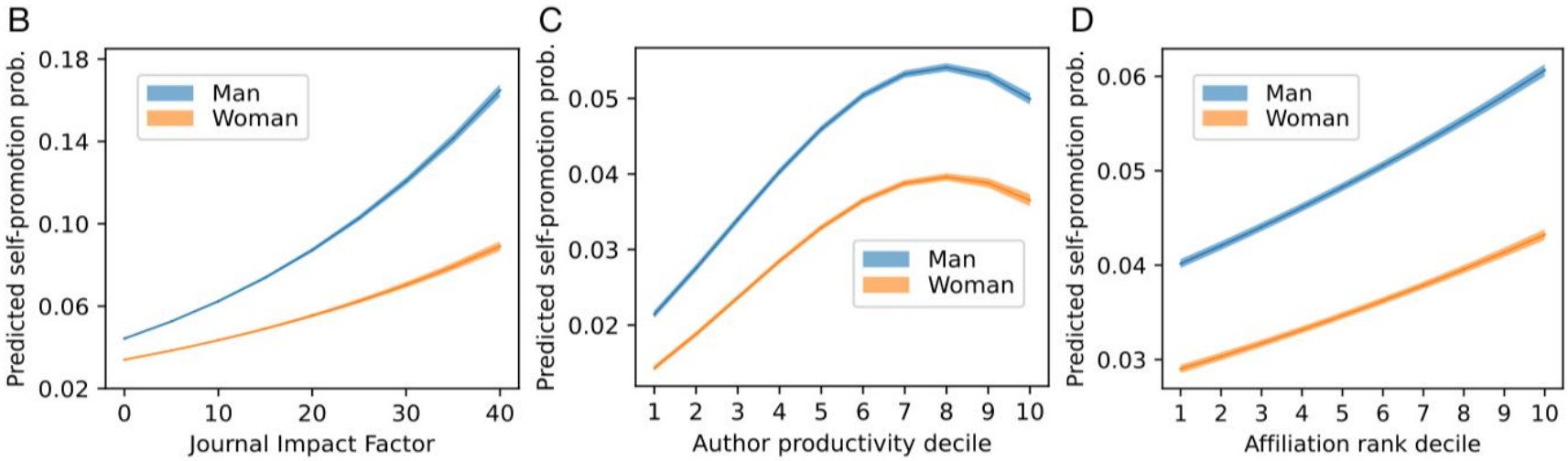
Key findings (raw differences w/o. any control)

- The increasing gender gap in (exploded) self-promotion rates
- Consistent gaps across different author roles and disciplines.



Peng et al. The Gender Gap in Scholarly Self-Promotion on Social Media. ***Nat. Commun.*** 2025.

Case study: scholarly self-promotion



Key findings (gender gaps adjusted for confounds)

- Gap increases with higher performance and achievements.
- Larger gender gaps for women with higher academic status.

Peng et al. The Gender Gap in Scholarly Self-Promotion on Social Media. ***Nat. Commun.*** 2025.

Case study: scholarly self-promotion

- Are there differential return by gender with self-promotion?
- DV: predicting a paper's total num. of tweets:

  $Y = \beta_0 + \beta_1 \text{ women} + \beta_2 \text{ promotion} + \beta_3 (\text{women} \cdot \text{promotion}) + \Sigma$

  Rewrite the equation as:

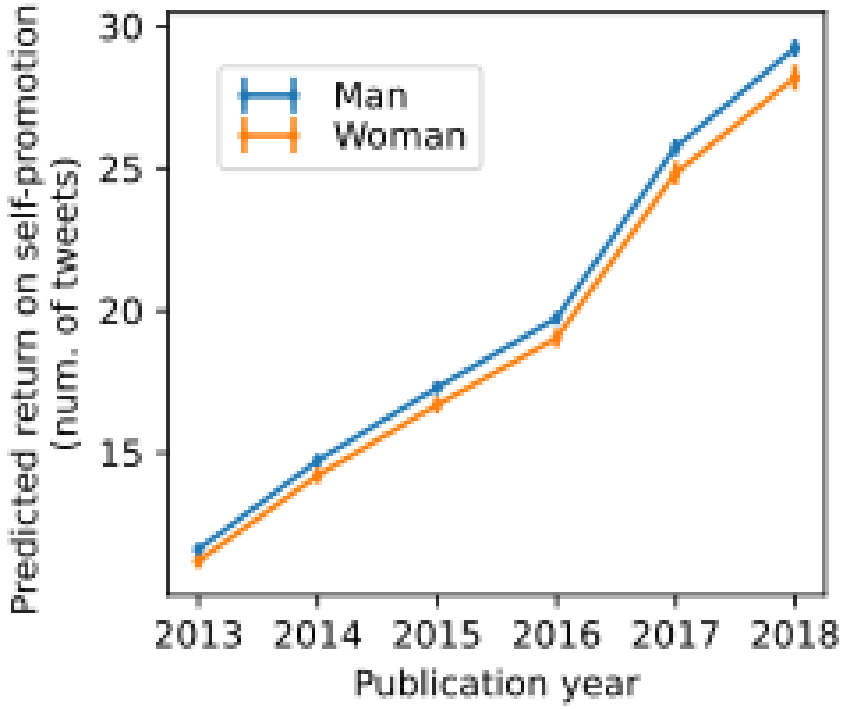
  $Y = \beta_0 + \beta_1 \text{ women} + (\beta_2 + \beta_3 \text{ women}) \text{ promotion} + \Sigma$

- How do we interpret β_3?
  - Equation for men: $Y = \beta_0 + \beta_2 \text{ promotion} + \Sigma$
  - Increase in Y if men self-promotes: β_2
  - Equation for women: $Y = \beta_0 + \beta_1 + (\beta_2 + \beta_3) \text{ promotion} + \Sigma$
  - Increase in Y if women self-promotes: $\beta_2 + \beta_3$
  - β_3: Women's return on self-promotion relative to men's

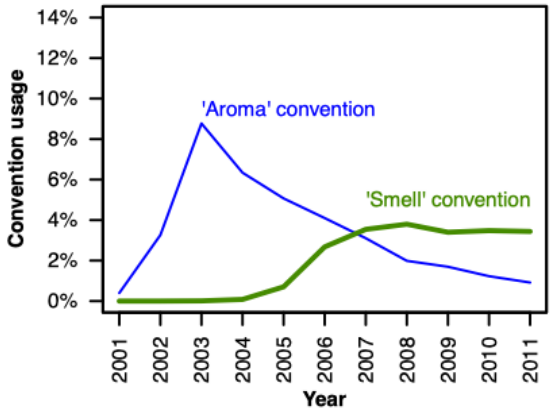
Case study: scholarly self-promotion

Key findings (differential "return" by gender)

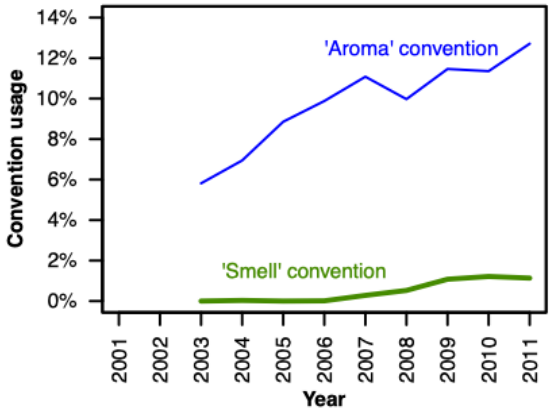
- Self-promotion increases engagement for both genders.
- But the boost for women is smaller than men's boost.



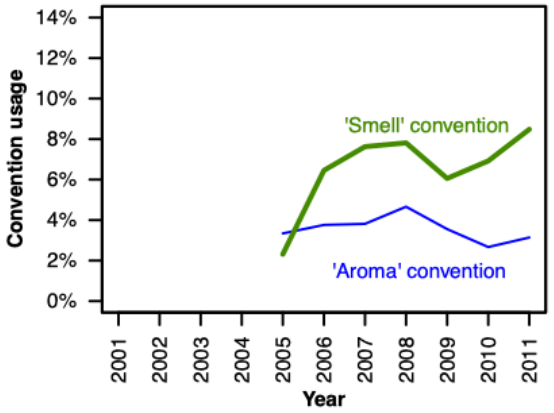
Churn prediction using linguistic change



(a) 'Aroma' was the dominant convention by 2003, but it was supplanted by 'S' (for 'Smell') around 2007.



(b) Users who joined in 2003 hung on to the 'Aroma' convention of their youth.



(c) Users who joined in 2005 were more receptive to the emerging 'S' norm.

Figure 1: Example of community and user evolution in BeerAdvocate: one norm for referring to the smell of a beer gave way to another, with different effects on different users depending on when they joined the community.

- **There are linguistic norms in online communities, which is evolving.**
- New users often adapt to current norms upon joining a community.
- As they get older, they become "lazy" to sync with new norms/rules.

Cristian et al. User lifecycle & linguistic change in online communities. ***WWW***. 2013.

Churn prediction using linguistic change

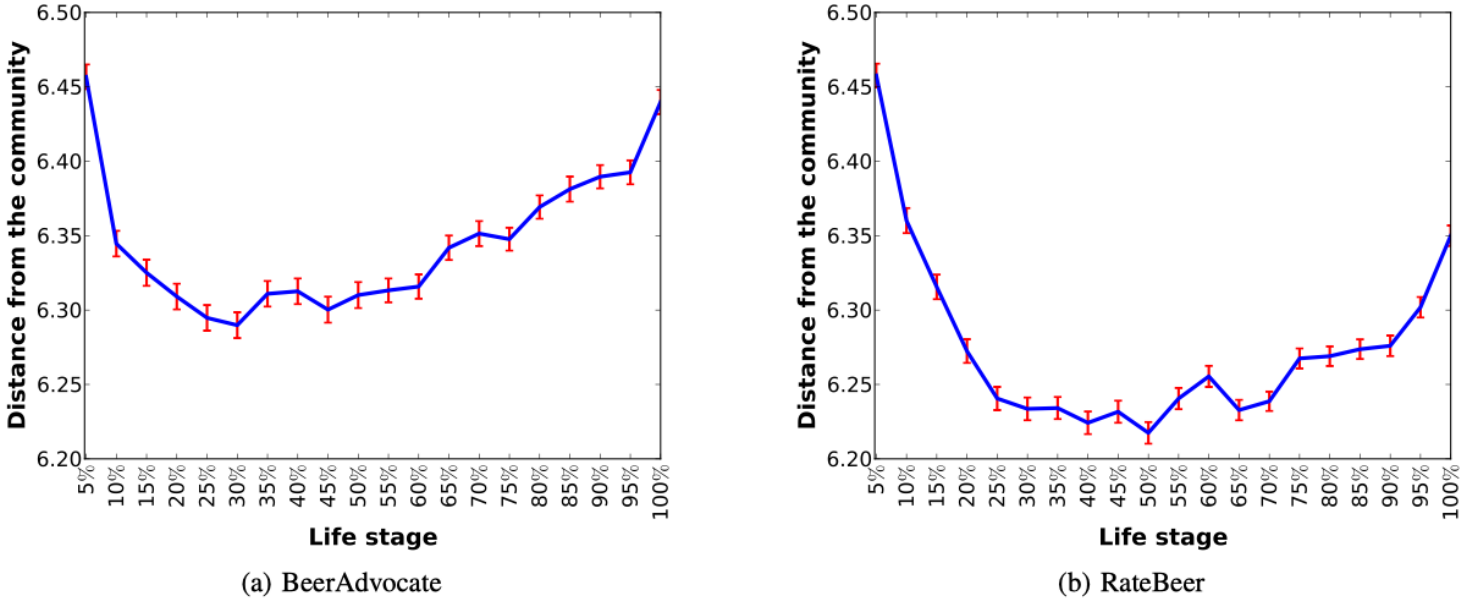
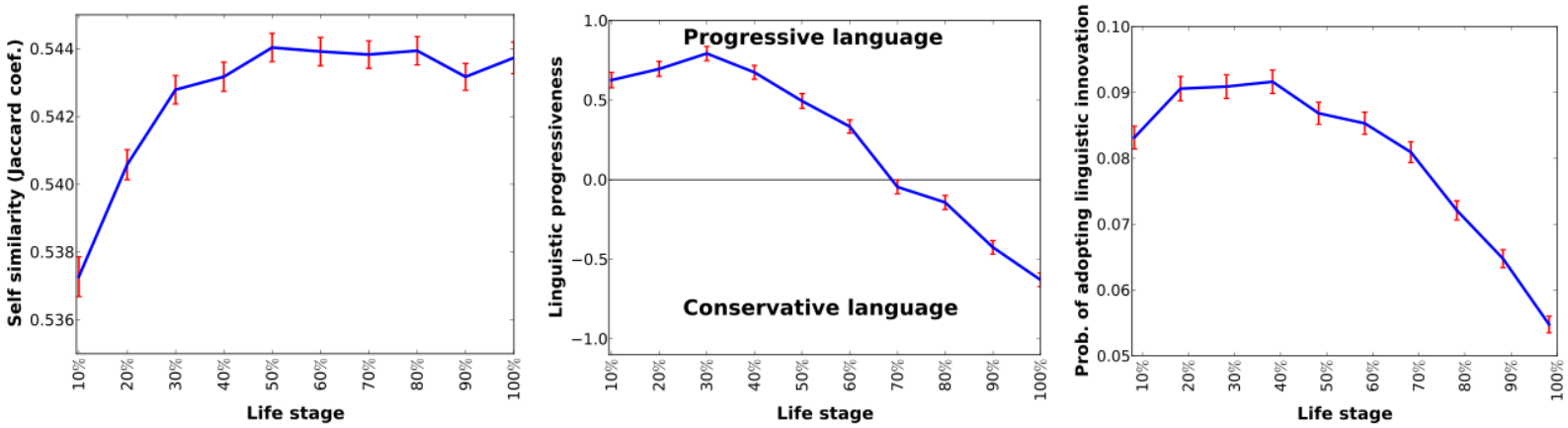


Figure 6: Lifecycle: Distance from the language of the community at each life-stage, calculated as the cross-entropy of each post according to the snapshot language models of the post's month (0% is birth, 100% is death). Lower values mean "closer to the community". (a) BeerAdvocate; (b) RateBeer.

A two-stage lifecycle of user susceptibility to linguistic norm change:

- Learning phase (1/3): users learn community's innovative language
- Conservative phase (2/3): users stop tuning with evolving norms

Churn prediction using linguistic change



(a) Language flexibility (b) Linguistic progressiveness (c) Adoption of lexical innovations

Figure 7: Lifecycle: (a) User-language flexibility at each life-stage, computed as the Jaccard coefficient between each post and a window of 10 previous posts written by the same user. (First and last 10 reviews of each user are not represented.) Users' language rigidifies after their linguistic adolescence. (b) Linguistic progressiveness at each life-stage. Positive values indicate future-leaning language, while negative values indicate past-leaning language. (c) Probability of adopting lexical innovations at each life-stage (0% is birth, 100% is death). (BeerAdvocate; same trends hold for RateBeer.)

- Adolescent users learn and adapt to new and progressive language.
- Senior users stop conforming and drift away from latest trends.

Can we leverage these patterns/features to predict user retention?

Churn prediction using linguistic change

Community	w	Departed range	Living range	Model	Test set performance Precision	Recall	F1	Test set class sizes Departed	Living
BeerAdvocate	20	20–50	200+	Activity	77.0	41.2	53.6	327 (46%)	387 (54%)
BeerAdvocate	20	20–50	200+	Full	69.6	46.9	**56.0**	327 (46%)	387 (54%)
BeerAdvocate	40	40–100	200+	Activity	74.6	27.3	39.8	218 (36%)	378 (64%)
BeerAdvocate	40	40–100	200+	Full	66.4	31.1	**42.2**	218 (36%)	378 (64%)
RateBeer	20	20–50	200+	Activity	73.7	19.3	30.5	261 (36%)	465 (64%)
RateBeer	20	20–50	200+	Full	64.8	32.3	**42.9**	261 (36%)	465 (64%)
RateBeer	40	40–100	200+	Activity	65.9	19.6	30.0	179 (27%)	470 (73%)
RateBeer	40	40–100	200+	Full	61.3	26.3	**36.7**	179 (27%)	470 (73%)

Table 3: Predicting whether a new user is about to leave the community or will remain as an active user. The number of posts we analyze is denoted by w. The 'full' models uses all of our features, while the 'activity' models uses only activity-based features. The precision, recall, and F1 numbers given are for the target 'departing' class. For all sites and w, the full model significantly improves over the activity-only model according to a paired Wilcoxon signed rank test on the F1 scores ($p < 0.001$).

Using first *w* posts to predict (logit model) churn or retention:

- Departed: users who left before writing m more posts (e.g. [20, 20+30])
- Stayed: users who stayed long enough to write >= 200 posts in total

Churn prediction using linguistic change

Features	F1 $w = 20$	F1 $w = 40$
Activity	30.5	30.0
+ Cross-entropy	37.4	32.2
+ Jaccard self-similarity	38.0	33.5
+ Adoption of lexical innovations	40.9	35.3
+ First-person singular pronouns	41.2	35.0
+ Number of words	42.9	36.7

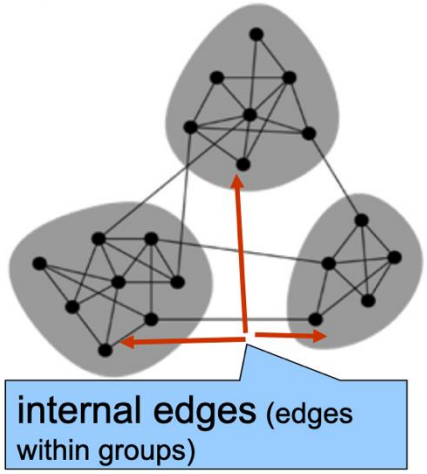
Table 4: Performance improvement resulting from incrementally adding our linguistic change features to the 'activity' model (for RateBeer, our 'test community').

- Churn prediction is a hard problem (baseline F1 = 0.3)
- Linguistic features can improve performance by ~40%!

Polarization on social media

Key polarization metrics

- Structure based:
  - Network modularity *Q*: the fraction of all edges that fall within communities minus the expected fraction if edges were distributed at random.
  - Edge density ratio: num. of edges (within / between)
- Opinion based:
  - Opinion variance among group members
    - Ideological preference
    - Sentiment score
  - Can use Std. or Gini index



Echo chambers

Echo chamber in social media refers to an environment where users are primarily exposed to content, opinions, and information that align with their existing beliefs.

How it forms:

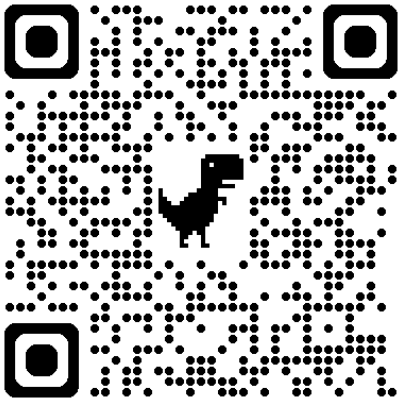
- Algorithms prioritizing content with high engagement
- User behavior: people tend to follow, friend, and join groups with like-minded individuals (homophily).
- Network effects: online communities (e.g., subreddits) amplify shared narratives due to social influence.

Echo chambers

How to break out of Echo Chambers:

- Diversify your feed: Follow credible accounts with opposing views (e.g., @nytimes and @foxnews).
- Use platform tools: Turn off personalized recommendations.
- Seek primary sources: Read original studies instead of opinion posts. Make own judgement with critical thinking.

An D3 interactive visualization of the formation of Echo Chambers:



Power of polarized teams

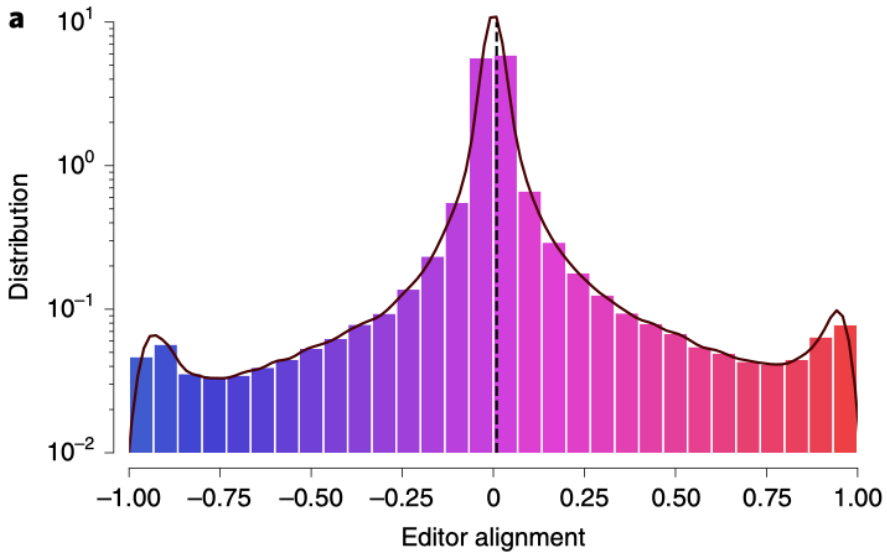
Dataset

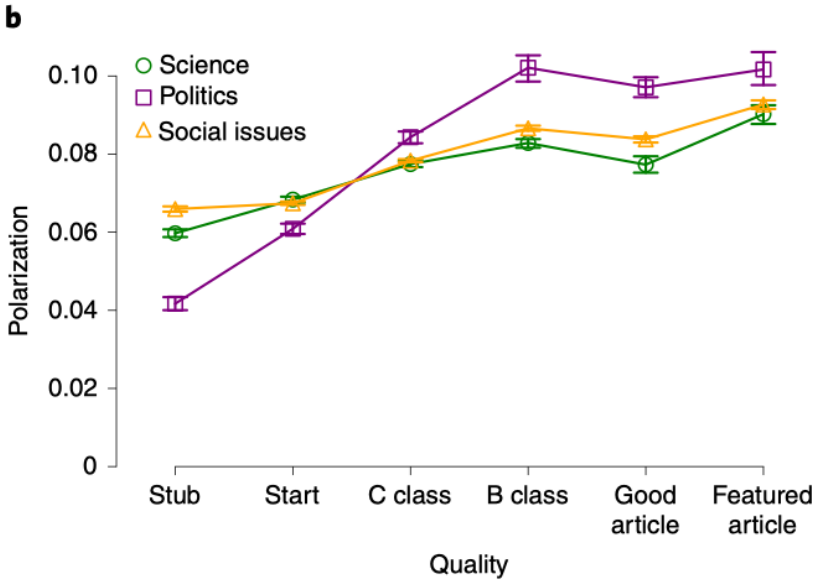
- English Wikipedia database: focused on articles in 3 categories: politics (20,947 articles), social issues (162,085), science (49,530).
- Represents about 5% of all English Wikipedia articles.

Methods

- Measure editor's ideological preference:
  - **Difference in the fraction** of bytes/words they contribute to conservative vs. liberal Wikipedia articles (based on a given corpus); **[-1, +1]**.
- Measure each article's quality:
  - Use the **Wikiclass package** to predict one of six quality categories
  - https://pypi.org/project/wikiclass/
- Measure team polarization of each article:
  - Compute the **variance** of alignments among all its editors

Power of polarized teams





- Center peak: most editors only contributed minor edits (typos)
- Two tail peaks: some editors have contributed much content to either liberal or conservative articles.
- Fig b: What factors can confound the link btw. Polz. and Quality?

Shi et al. The wisdom of polarized crowds. ***Nat. Human Behaviour.*** 2019.

Power of polarized teams

Table 1 | Odds ratios from ordinal logistic regression models predicting article quality

Independent variable	Politics	*P*	Social issues	*P*	Science	*P*
Polarization	18.88	<0.001	2.06	<0.001	1.79	0.006
|alignment|	0.30	<0.001	0.49	<0.001	0.65	0.002
Editing experience	1.05	0.02	1.06	<0.001	1.01	0.30
Number of editors	0.41	<0.001	0.51	<0.001	0.56	<0.001
Article length	33.55	<0.001	47.83	<0.001	56.54	<0.001
Number of edits	3.26	<0.001	1.71	<0.001	1.69	<0.001
N	12,570		161,070		49,995	

Statistical significance levels (*P* values) are derived from two-sided Wald tests. The columns present odds ratios estimated on political, social issues and science articles, separately. *N* denotes sample size.

- Polarized teams consisting of ideologically diverse editors can produce articles of a higher quality than homogeneous teams.
- The effect is more pronounced in Politics articles. Why?

Personalization & Recommendation

Building intelligent systems to recommend content, accounts, and products tailored to individual users.

- **User segmentation**: dividing users into groups based on their demographics, behavior, or preference for marketing strategy.
- **Content-based filtering**: match content to user interests.
- **Collaborative filtering**: user-user & item-item similarities.
  - Find users similar to you (e.g., via Pearson correlation or cosine on rating vectors). Recomm. what those users liked & you haven't seen.
  - Compute item-to-item similarity (based on features). For a user who liked A, recommend similar items B, C, etc.

User segmentation

Linguistic methods:

- Cluster users based on their linguistic features
  - BoW, TF-IDF, LIWC, LDA, etc.
- Obtain user embeddings based on post vectors

Network-based methods:

- Node2vec: do random walks on social networks and train the word2vec model on the sampled trajectories of users (words).

Course notes

- HW2 released
  - Due on Nov 14.
- Social network analysis next week
  - Will use Team-Based Learning (TBL) setting
  - **New location**: CIC G-001 (map)
  - Watch AI-made lecture before class:
    - Canvas -> SDSC3013 -> Panopto Recordings
      -> Social Network Centrality Measure
  - In-class group-based exercises (not graded)
  - Will take our second attendance checking!