Choices and Consequences in Computing

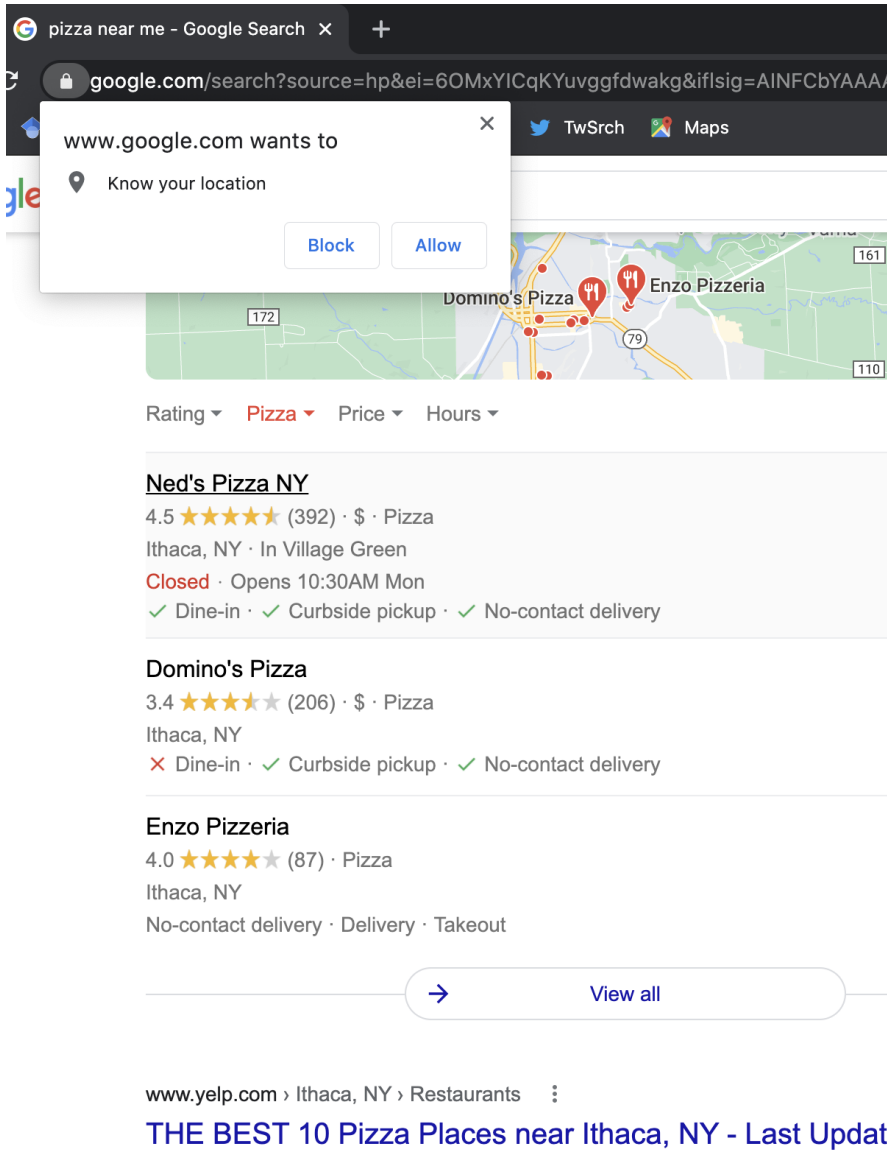
INFO 1260 / CS 1340

Lecture 8: Personalization and Fragmentation

February 7, 2024

Personalization

- We don't all see the same things, even on the same site or in response to the same queries.
- A useful question to ask yourself regularly on-line:
  - "Why am I seeing this?"
- In many situations, it would seem strange not to personalize.
  - Would it make sense to show everyone the same set of search results for the query, "Pizza near me"?



Q: What are some of the basic information sources that Web sites can use to personalize content to you?

Different Preferences

	Source 1	Source 2	Source 3	Source 4
User 1	.4	.4	.2	0
User 2	.2	.2	.1	0
User 3	.2	.3	.3	.2
User 4	0	.1	.2	.2
User 5	0	.2	.4	.4

Often, preferences can't be explained using just activities u_i and appeals s_j.

- What's going on in the table above?
- Are Sources 1 and 4 popular or not?

Maybe there are two kinds of users: those who like Sources 1 and 2, and those who like Sources 3 and 4.

- But there's some spillover: User 2 mainly likes Sources 1 and 2, but sometimes looks at Sources 3 and 4.
- And what about User 3, who seems to like everything?

Multiple Dimensions

	Source 1	Source 2	Source 3	Source 4
User 1	.4	.4	.2	0
User 2	.2	.2	.1	0
User 3	.2	.3	.3	.2
User 4	0	.1	.2	.2
User 5	0	.2	.4	.4

A more general model.

- The sources are organized into two *genres*.
- When you select according to one of the genres, you sample at random using a probability distribution specific to that genre.

For example:

- Probabilities for Genre 1 might be (.4, .4, .2, 0).
- Probabilities for Genre 2 might be (0, .2, .4, .4).
- Note: some items can be chosen under either genre.
- $s_j[1]$ will denote the probability Source j is chosen when selecting according to Genre 1; $s_j[2]$ for Genre 2.

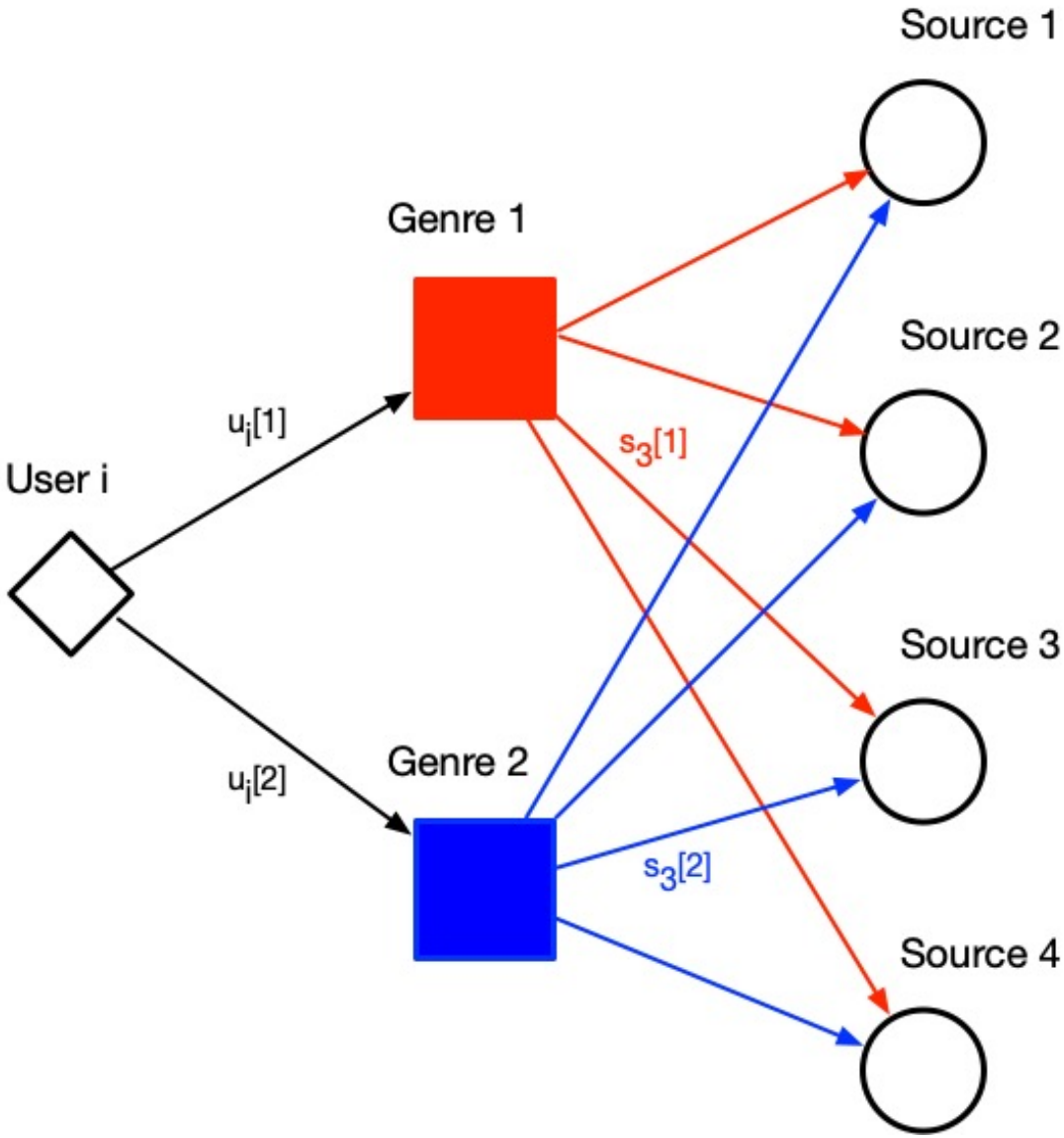
Multiple Dimensions

	Source 1	Source 2	Source 3	Source 4
User 1	.4	.4	.2	0
User 2	.2	.2	.1	0
User 3	.2	.3	.3	.2
User 4	0	.1	.2	.2
User 5	0	.2	.4	.4

How about users?

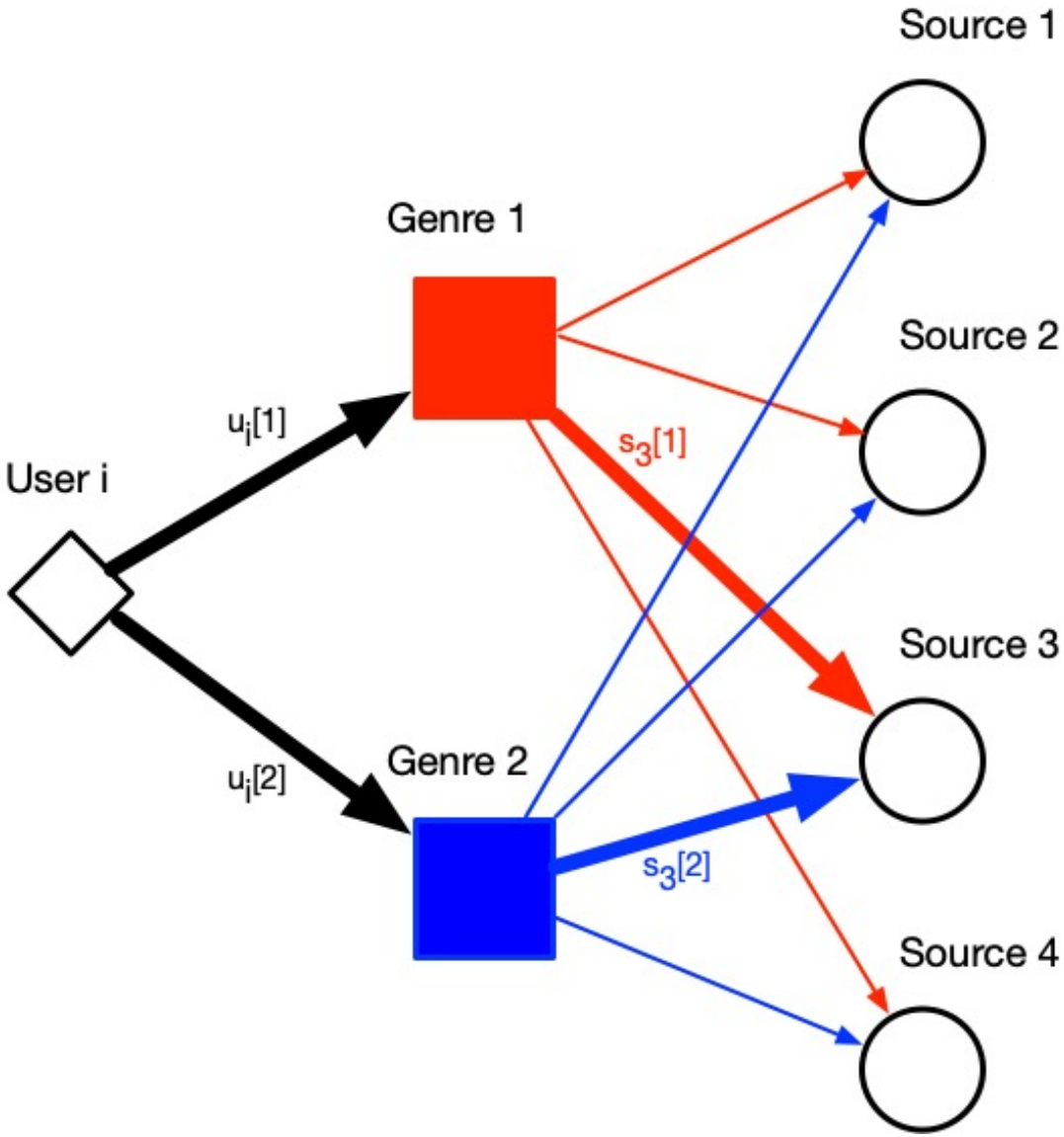
- When User i wants to select a source, they first choose genre at random.
- Chooses Genre 1 with probability $u_i[1]$, Genre 2 with $u_i[2]$.
- Then select a source according to the probabilities in that genre.
- $u_i[1] + u_i[2]$ can be less than 1; they choose nothing with remaining prob.

No need to limit the model to two genres (though our examples will).



User first selects genre, then source according to genre.

- Probability User i chooses Source j is $u_i[1]s_j[1] + u_i[2]s_j[2]$.



User first selects genre, then source according to genre.

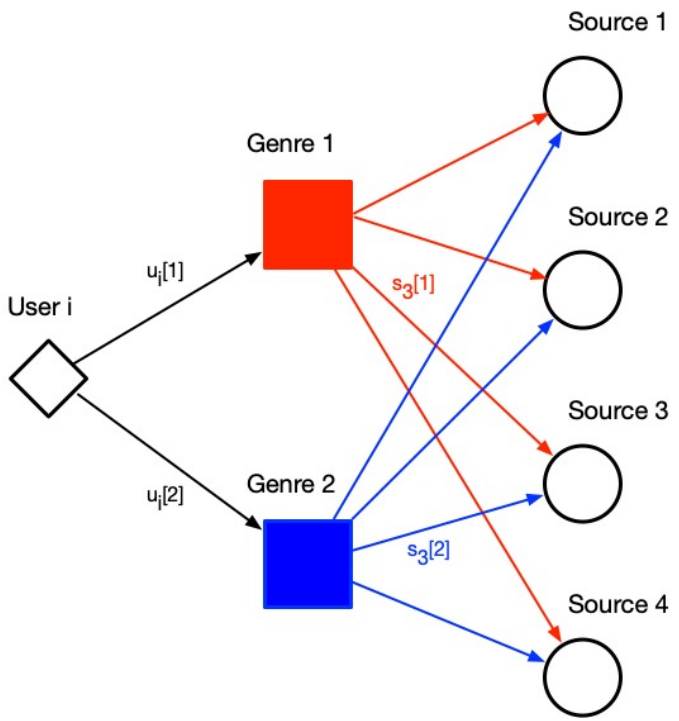
- Probability User i chooses Source j is $u_i[1]s_j[1] + u_i[2]s_j[2]$.

Multiple Dimensions

	Source 1	Source 2	Source 3	Source 4
User 1	.4	.4	.2	0
User 2	.2	.2	.1	0
User 3	.2	.3	.3	.2
User 4	0	.1	.2	.2
User 5	0	.2	.4	.4

In our example:

- Genre probabilities are (.4, .4, .2, 0) and (0, .2, .4, .4).
- User probabilities are (1,0), (.5,0), (.5,.5), (0, .5), (0,1).

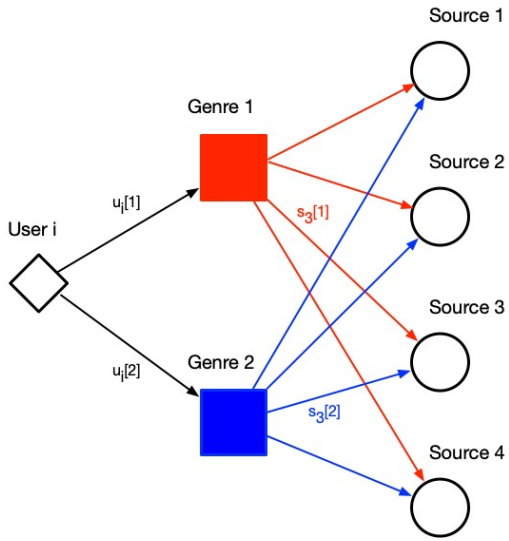


Latent factor models

	Source 1	Source 2	Source 3	Source 4
User 1	.4	.4	.2	0
User 2	.2	.2	.1	0
User 3	.2	.3	.3	.2
User 4	0	.1	.2	.2
User 5	0	.2	.4	.4

Multiple terms for this type of model.

- Latent factor model, mixture model, topic model
- Closely related: principal components analysis, multidimensional scaling.
- Linear algebra provides the main set of methods for fitting the model (outside this course).



Subtle aspect of the model: Who exactly chose the genres?
They emerge from the data and from the answer to the following question:

- Choose two numbers for each user, and two numbers of each source, so the resulting selection process matches the table as well as possible.
- And again: In general, we'd look for more than two genres. (Perhaps 20-50 to fit table with millions of users, thousands of sources.)

Latent factor models

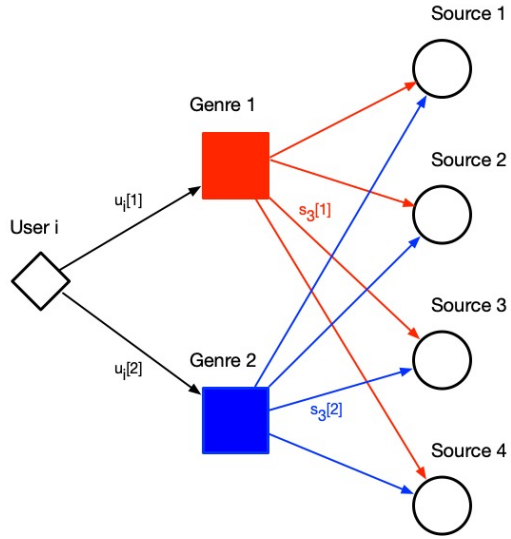
Q: Should we think of the platform as having chosen the genres or not?

Not clear-cut:

- The engineers working for the platform can't necessarily explain where the genres came from.
- But they made a set of decisions along the way:
  - Choosing to use this model at all;
  - Choosing the number of genres: parsimony versus fit.
- An interpretability problem: if you're working with content producers, you might need to explain to them what the different genres represent.
- A feedback loop problem: once system is running, the data you're using to build your model came from recommendations made by your model.

Each user and piece of content described by a small sequence of numbers.

- Recommending entertainment
- Prioritizing search results
- Prioritizing product displays
- Personalizing streams of news



Personalizing content can help users focus on what's most important to them.

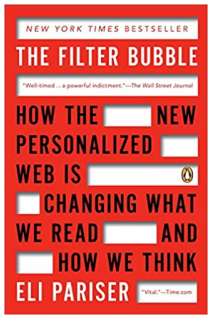
- Without personalization, we might all just see what's globally popular.
- Personalization similarly helps creators of niche content get seen.

Q: Counterbalancing this, what are some of the risks in personalizing?

- Loss of opportunities for serendipitous discovery.
  (Foster and Ford, "Serendipity and information seeking," 2003.)
- The model might be wrong about you;
  some of your preferences might not be visible in the data;
  your preferences might change over time.
- Tension with the democracy theory of free speech.

Echo Chambers and Filter Bubbles

Concerns about effect on democratic debate coalesced around the ideas of *echo chambers* and (more recently) *filter bubbles*



A well-functioning system of free expression must meet two requirements.

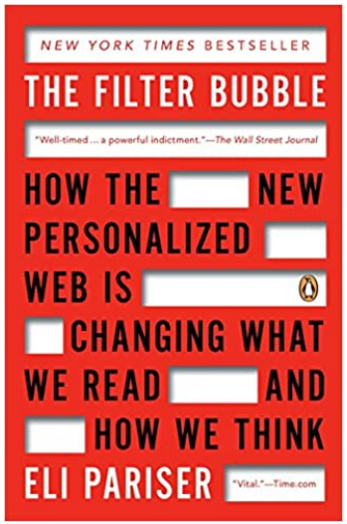
- First, people should be exposed to materials that they would not have chosen in advance. Unplanned, unanticipated encounters are central to democracy itself. ....
- Second, many or most citizens should have a range of common experiences. Without shared experiences, a heterogeneous society will have a much more difficult time in addressing social problems. ...

— Cass Sunstein, *Republic.com*, 2001

In the filter bubble, there's less room for the chance encounters that bring insight and learning.

— Eli Pariser, *The Filter Bubble*, 2011

Echo Chambers and Filter Bubbles



Concerns about effect on democratic debate coalesced around the ideas of *echo chambers* and (more recently) *filter bubbles*

- "In the filter bubble, there's less room for the chance encounters that bring insight and learning" (Pariser 2011).

These are arguments that this could happen in principle.
It leaves open the empirical question of how much it's happening (but more on this later).

The abundance of content available creates two opposing forces:

- There's greater heterogeneity of voices and opinions than ever before.
- It's increasingly easy to consume only things that agree with you.
    - Both when finding things on your own, and when they're personalized.

The interaction of ranking and genres

- Our ranking model tells us to favor items the user is likely to choose.
- Our latent factor model tells us that user selections are guided by genre.
- What happens when we put these together?

A simple thought experiment (Harald Steck, Netflix, 2018)

- Two genres (e.g. action movie and documentary).
- Each movie j has a selection probability $s_j[1]$ in action genre and $s_j[2]$ in documentary genre.
- Imagine a user i with $u_i[1] = .70$ and $u_i[2] = .30$ (70% of the time chooses from action genre, 30% from documentary.)

Consider top 10 likeliest selections in each genre

- Top action movies: largest $s_j[1]$ values. Call these $x_1 > x_2 > \cdots > x_{10}$.
- Top documentaries: largest $s_j[2]$ values. Call these $y_1 > y_2 > \cdots > y_{10}$.
- Safe to assume these two lists of 10 movies have no entries in common. (In fact assume they have negligible probability in the other genre.)
- Empirically: x_1, y_1 similar sizes, x_{10}, y_{10} about half size. ($x_{10} \approx .5 * x_1$ and similarly for y_{10})

The interaction of ranking and genres

- Top action movies: largest $s_j[1]$ values. Call these $x_1 > x_2 > \cdots > x_{10}$.
- Top documentaries: largest $s_j[2]$ values. Call these $y_1 > y_2 > \cdots > y_{10}$.

User i: 70% of the time chooses from action genre, 30% from documentary.

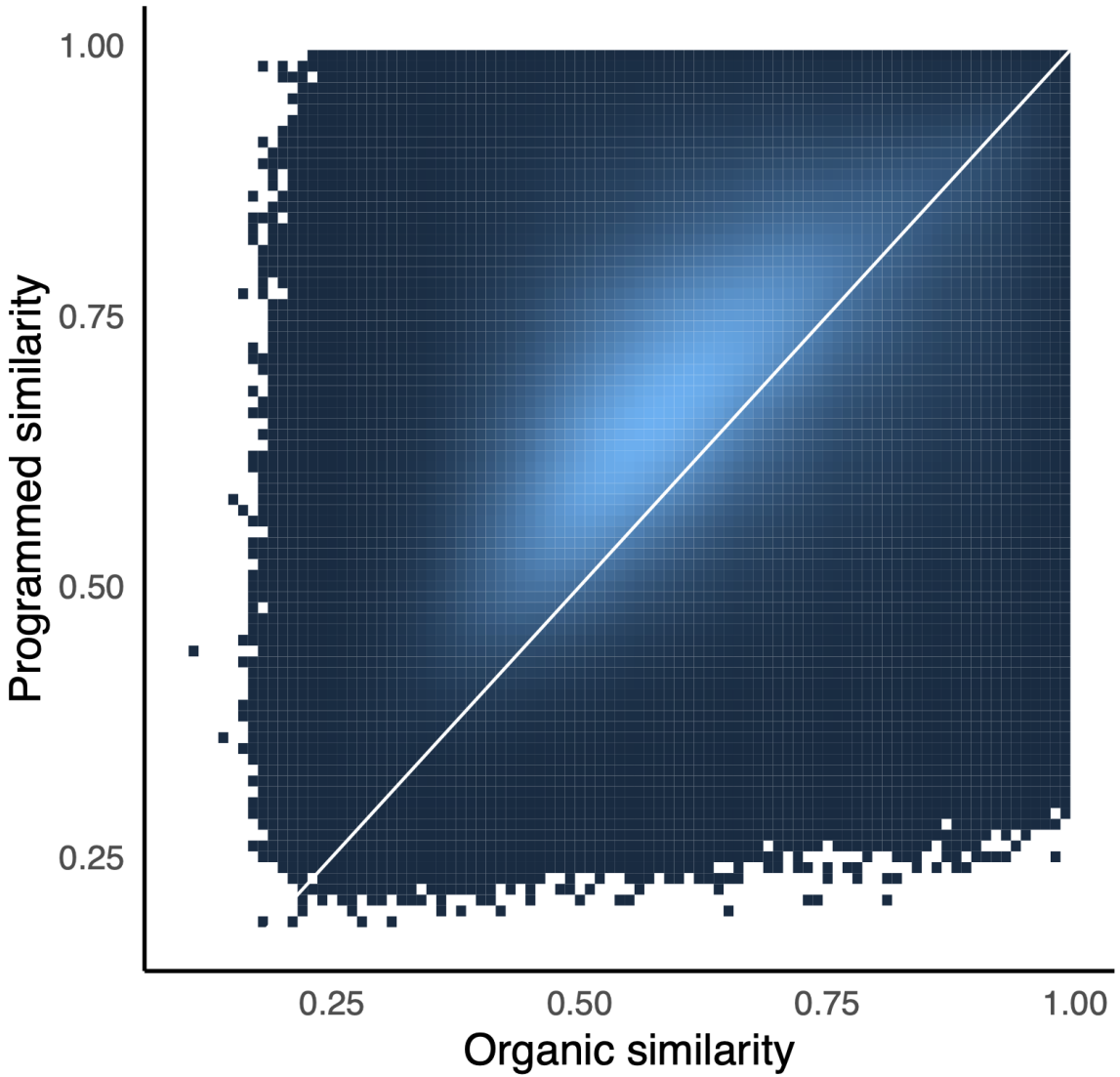
- Action probabilities are $.7x_1 > .7x_2 > \cdots > .7x_{10}$.
- Documentary probabilities $.3y_1 > .3y_2 > \cdots > .3y_{10}$.
- Empirically: x_1, y_1 similar sizes, x_{10}, y_{10} about half size.
- So: $.7x_{10} \approx .7 * .5x_1 = .35x_1 \approx .35y_1 > .3y_1$.

This means if we show a list of 10 items ranked by probability,
all 10 will be action movies.

- User (correctly) thinks of their viewing as 70% action, 30% documentary.
- But a ranked list of their most likely selection is 100% action movies.
- Whatever's gone wrong is subtle:

> The ranked list really does solve the problem of: list a set of 10 movies to maximize chance of containing user's intended selection.

Increasing homogeneity in practice: Case Study of Spotify



Anderson-Maystre-Mehrotra-Anderson-Lalmas 2020

Empirical Studies

Challenging to evaluate the question of whether personalization in on-line platforms is leading to increased divergence.

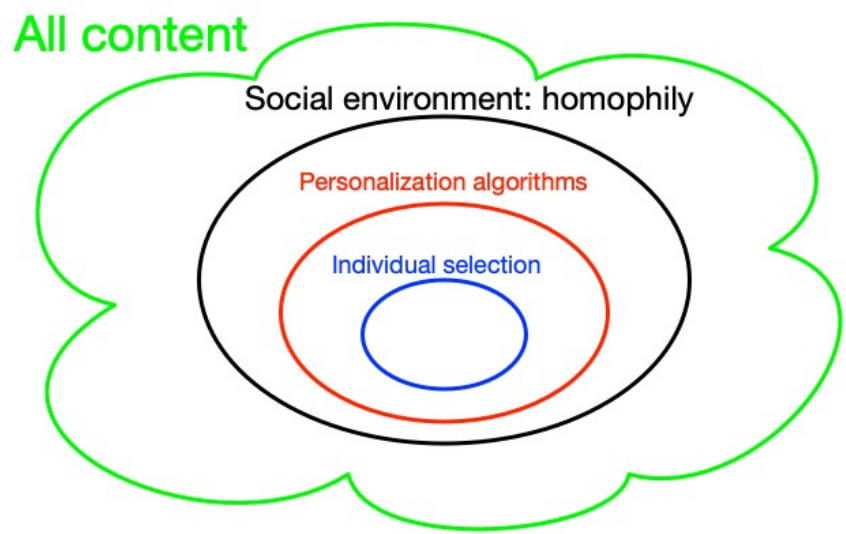
- Hard to measure opinions from decades ago with same fidelity as now.
- Hard to separate effect of on-line tools from broader environment.

Bakshy-Messing-Adamic 2015

- Study performed at Facebook, published in *Science*.
- Analyzed the effect of News Feed's algorithmic ranking of content.
- Cross-cutting content: sources different from user's ideological leaning.

Findings: A cascade of three filters

- Similarity to friends: principle of *homophily*.
- Exposure by News Feed algorithm.
- Selection by the user.



Empirical Studies

Challenging to evaluate the question of whether personalization in on-line platforms is leading to increased divergence.

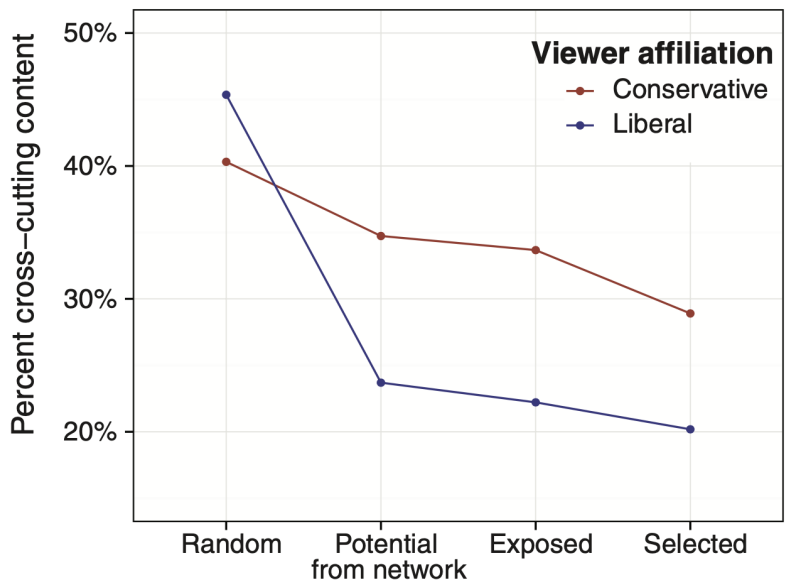
- Hard to measure opinions from decades ago with same fidelity as now.
- Hard to separate effect of on-line tools from broader environment.

Bakshy-Messing-Adamic 2015

- Study performed at Facebook, published in *Science*.
- Analyzed the effect of News Feed's algorithmic ranking of content.
- Cross-cutting content: sources different from user's ideological leaning.

Findings: A cascade of three filters

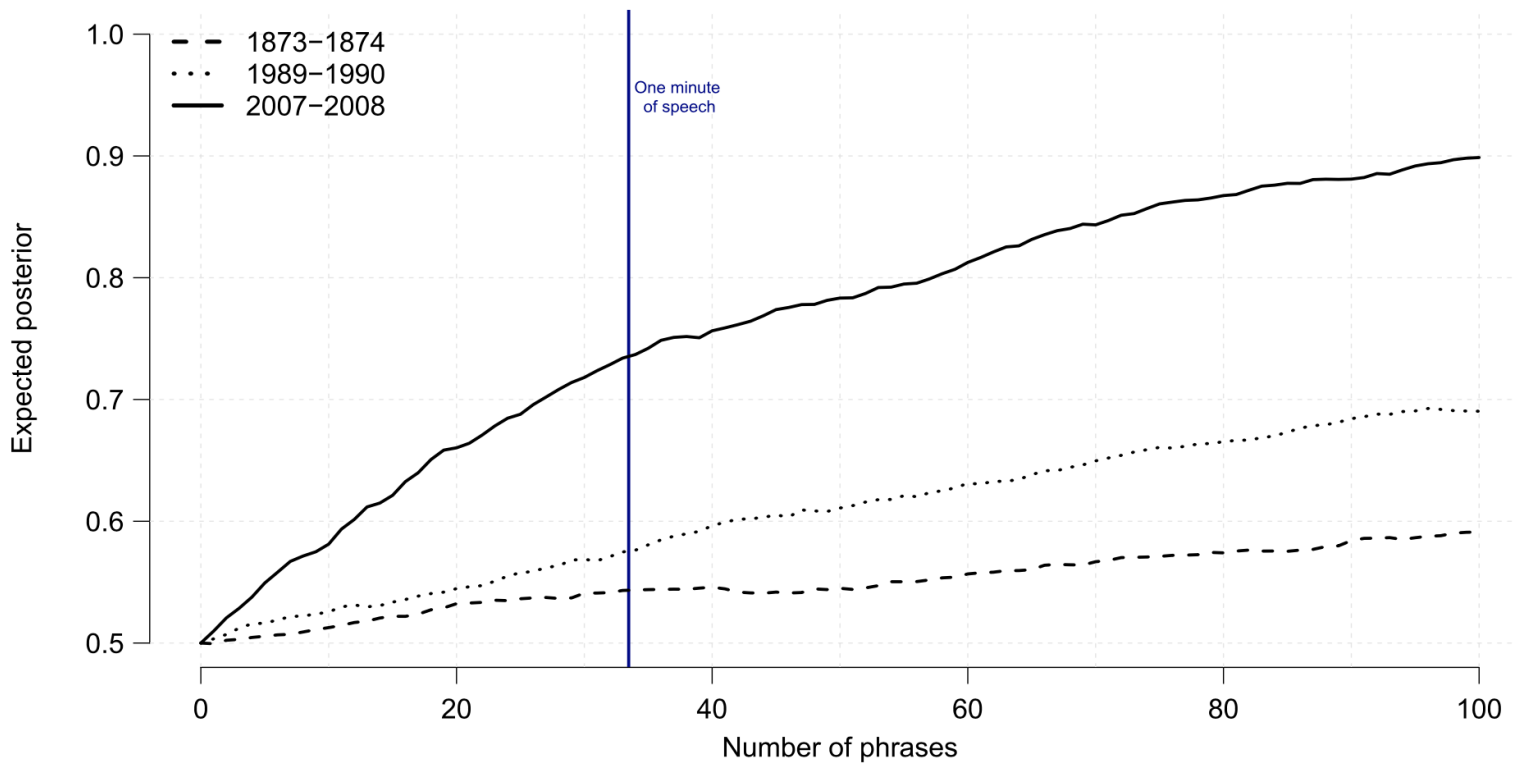
- Similarity to friends: principle of *homophily*.
- Exposure by News Feed algorithm.
- Selection by the user.



Empirical Studies

Challenging to evaluate the question of whether personalization in on-line platforms is leading to increased divergence.

- Hard to measure opinions from decades ago with same fidelity as now.
- Hard to separate effect of on-line tools from broader environment.

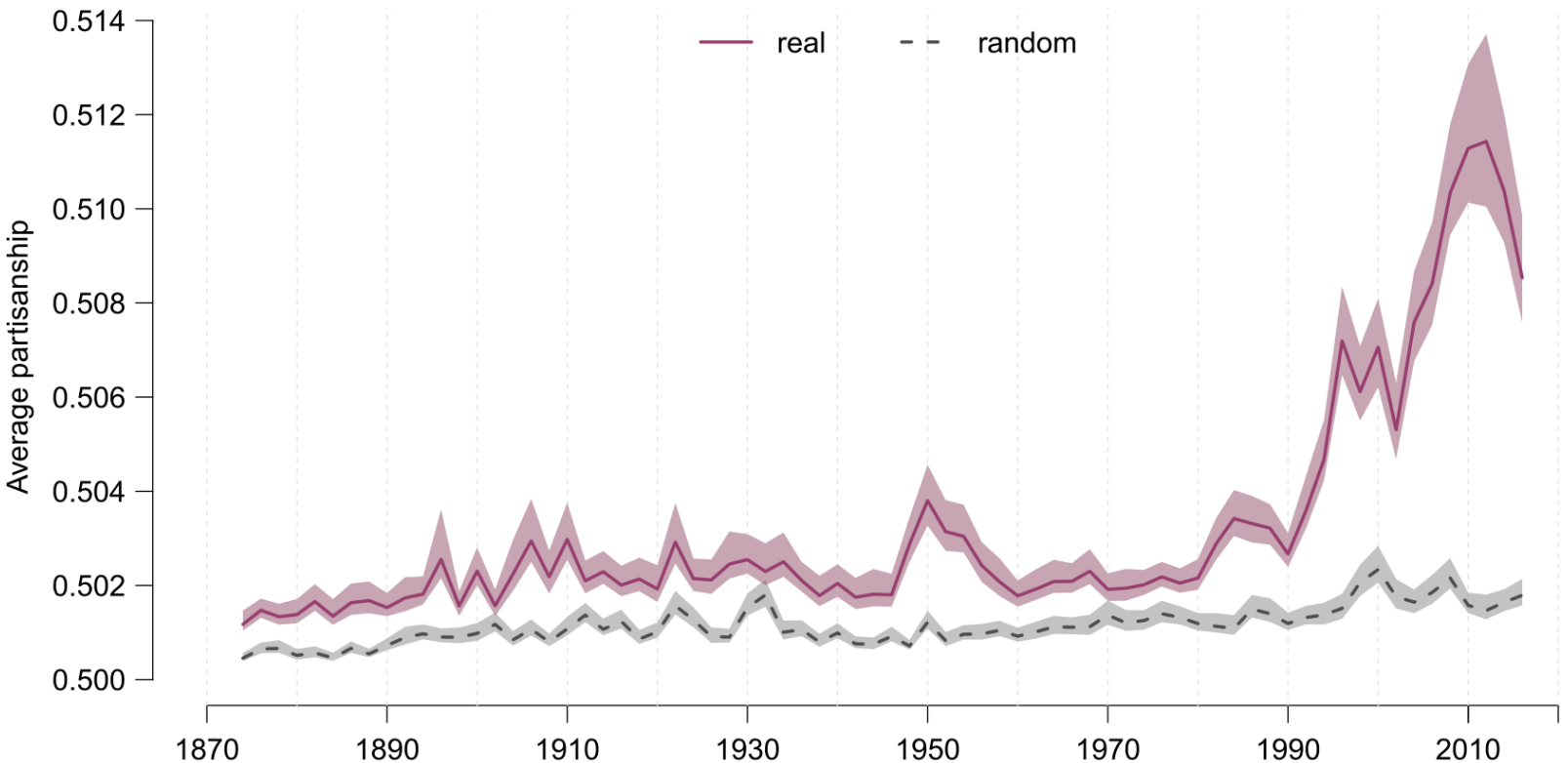


Gentzkow, Shapiro, and Taddy 2019

Empirical Studies

Challenging to evaluate the question of whether personalization in on-line platforms is leading to increased divergence.

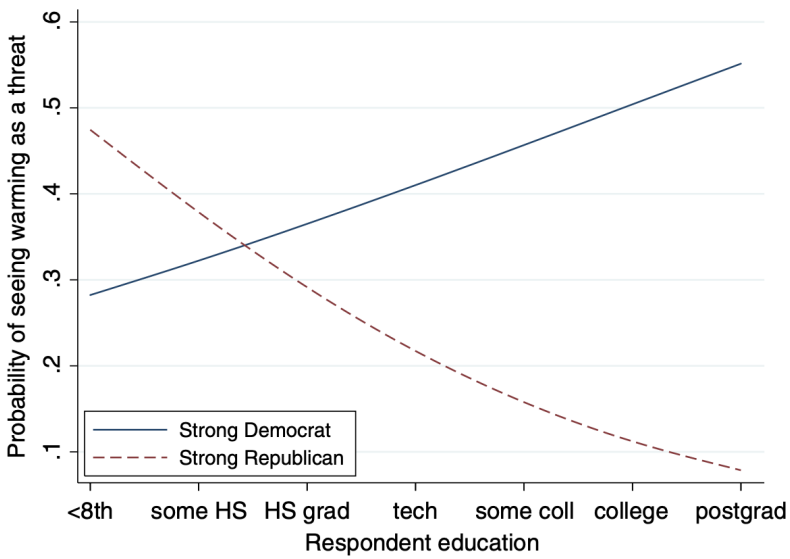
- Hard to measure opinions from decades ago with same fidelity as now.
- Hard to separate effect of on-line tools from broader environment.

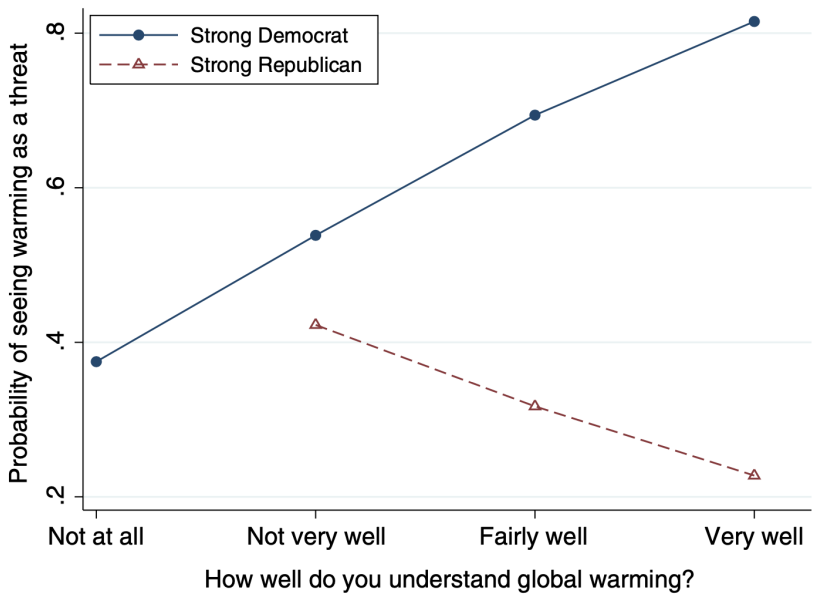


Gentzkow, Shapiro, and Taddy 2019

Empirical Studies

Divergence is not only based on access to relevant information. (Hamilton, 2011)



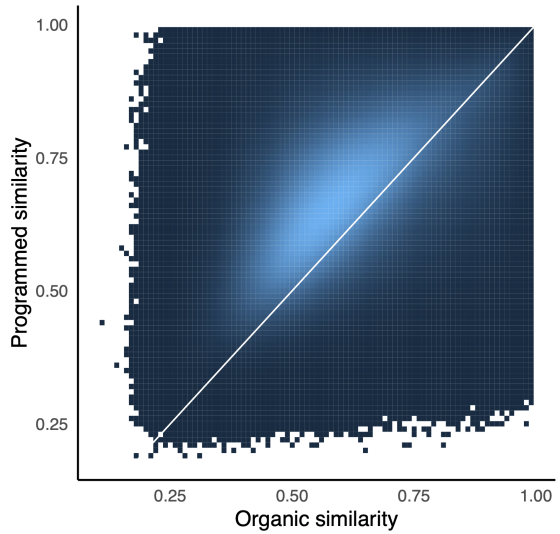


Probability of view climate change as a global threat, as a function of education or self-reported understanding.

Reflections

Personalization creates a natural risk of filter bubbles.

- Challenges for serendipity, personal variety, democratic process.
- Empirically a complex question.



- A cascade of multiple filters: social, algorithmic, individual.
- The process of ranking can distort attention toward the category or genre with the largest probability.
- An interesting tension: Personalization increases heterogeneity between different people's consumption, but may reduce it in one person's consumption.