TECHNISCHE UNIVERSITÄT WIEN

Opinion Formation and Polarization in Social Networks

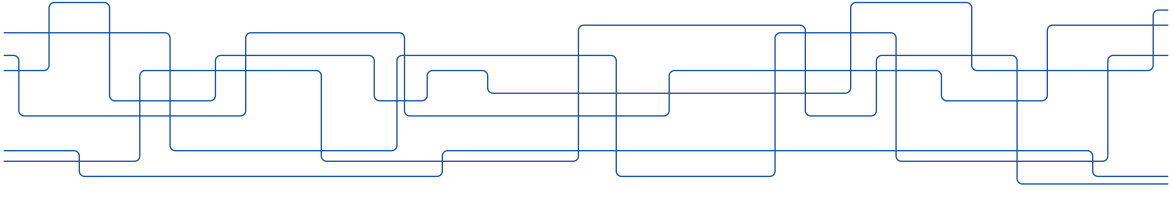
Tutorial at the ESSAI summer school, July 6, 2026

Aristides Gionis, KTH

Stefan Neumann, TU Wien stefan.neumann@tuwien.ac.at

Bruno Ordozgoiti

WWTF



the early days of social media

- in the early days of social media, focus was on new possibilities
  - connecting people across the world in real time
  - "democratization of information"
  - people get exposed to a variety of viewpoints
- that did not last long ...

social-media algorithms under scrutiny

The New York Times

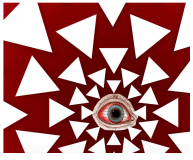
Opinion

YouTube, the Great Radicalizer

By Zeynep Tufekci

March 10, 2018

Share full article 101



COMMITTEE SENSITIVE – RUSSIA INVESTIGATION ONLY

116TH CONGRESS 1st Session | SENATE | REPORT 116-XX

(U) REPORT

OF THE

SELECT COMMITTEE ON INTELLIGENCE

UNITED STATES SENATE

Algorithmic amplification of politics on Twitter

Ferenc Huszár[a,b,c,1,2], Sofia Ira Ktena[a,1,2], Conor O'Brien[a,1], Luca Belli[a,1], Andrew Schlaikjer[a], and Moritz Hardt[d]

[a]Machine Learning Ethics, Transparency, and Accountability Team, Twitter, San Francisco, CA 94103; [b]Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, United Kingdom; [c]Gatsby Computational Neuroscience Unit, University College London, London, W1T 4JG, United Kingdom; and [d]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

Edited by David Laitin, Department of Political Science, Stanford University, Stanford, CA; received December 11, 2020; accepted October 5, 2021

Content on Twitter's home timeline is selected and ordered by personalization algorithms. By consistently ranking certain content higher, these algorithms may amplify some messages while reducing the visibility of others. There's been intense public and scholarly debate about the possibility that some political groups right enjoys higher algorithmic amplification than the mainstream political left. Consistent with this overall trend, our second set of findings studying the US media landscape revealed that algorithmic amplification favors right-leaning news sources. We further looked at whether algorithms amplify far-left and far-right political groups more than moderate ones; contrary to prevailing public belief, we did not find evidence to support this hypothesis. We hope our findings will contribute to an evidence-based debate on the role personalization algorithms play in shaping political content consumption.

social media | algorithmic personalization | media amplification | political bias

When Twitter introduced machine learning to personalize the Home timeline in 2016, it excluded a randomly chosen control group of 1% of all global Twitter users from the new personalized Home timeline. Individuals in this control group have never experienced personalized ranked timelines. Instead, their

"In six out of seven countries studied, the mainstream political right enjoys higher algorithmic amplification than the mainstream political left."

be isolated from indirect effects of personalization, as individuals in the control group encounter content shared by users in the treatment group. Therefore, although a randomized controlled experiment, our experiment does not satisfy the well-known Stable Unit Treatment Value Assumption from causal inference (23). As a consequence, it cannot provide unbiased estimates of causal quantities of interest, such as the average treatment

Significance

The role of social media in political discourse has been the

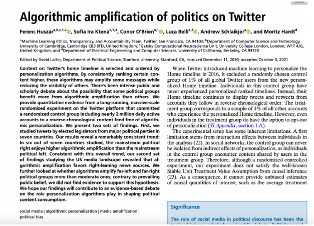
POLITICAL SCIENCE

COMPUTER SCIENCES

- many studies about online social networks are empirical
- some data is collected, relevant phenomena are analyzed

  do filter bubbles exist? does the Twitter/X algorithm promote left- or right-wing content?
- the algorithms are unknown and only user actions can be observed

  $\Rightarrow$ when Twitter/X changes its algorithm, unclear how findings generalize



Algorithmic amplification of politics on Twitter

Ferenc Huszár[a,b,c,1,2], Sofia Ira Ktena[a,1,3], Conor O'Brien[a,1], Luca Belli[a,2], Andrew Schlaikjer[a], and Moritz Hardt[d]

[a]Machine Learning Ethics, Transparency, and Accountability Team, Twitter, San Francisco, CA 94103; [b]Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, United Kingdom; [c]Gatsby Computational Neuroscience Unit, University College London, London, W1T 4JG, United Kingdom; and [d]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

Edited by David Laitin, Department of Political Science, Stanford University, Stanford, CA; received December 11, 2020; accepted October 5, 2021

Content on Twitter's home timeline is selected and ordered by personalization algorithms. By consistently ranking certain content higher, these algorithms may amplify some messages while reducing the visibility of others. There's been intense public and scholarly debate about the possibility that some political groups benefit more from algorithmic amplification than others. We provide quantitative evidence from a long-running, massive-scale randomized experiment on the Twitter platform that committed a randomized control group including nearly 2 million daily active accounts to a reverse-chronological content feed free of algorithmic personalization. We present two sets of findings. First, we studied tweets by elected legislators from major political parties in seven countries. Our results reveal a remarkably consistent trend: In six out of seven countries studied, the mainstream political right enjoys higher algorithmic amplification than the mainstream political left. Consistent with this overall trend, our second set of findings studying the US media landscape revealed that algorithmic amplification favors right-leaning news sources. We further looked at whether algorithms amplify far-left and far-right political groups more than moderate ones; contrary to prevailing public belief, we did not find evidence to support this hypothesis. We hope our findings will contribute to an evidence-based debate on the role personalization algorithms play in shaping political content consumption.

social media | algorithmic personalization | media amplification | political bias

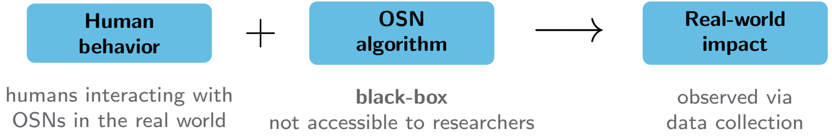
When Twitter introduced machine learning to personalize the Home timeline in 2016, it excluded a randomly chosen control group of 1% of all global Twitter users from the new personalized Home timeline. Individuals in this control group have never experienced personalized ranked timelines. Instead, their Home timeline continues to display tweets and retweets from accounts they follow in reverse chronological order. The treatment group corresponds to a sample of 4% of all other accounts who experience the personalized Home timeline. However, even individuals in the treatment group do have the option to opt-out of personalization (*SI Appendix*, section 1.A).

The experimental setup has some inherent limitations. A first limitation stems from interaction effects between individuals in the analysis (22). In social networks, the control group can never be isolated from indirect effects of personalization, as individuals in the control group encounter content shared by users in the treatment group. Therefore, although a randomized controlled experiment, our experiment does not satisfy the well-known Stable Unit Treatment Value Assumption from causal inference (23). As a consequence, it cannot provide unbiased estimates of causal quantities of interest, such as the average treatment

Significance

The role of social media in political discourse has been the

the empirical approach

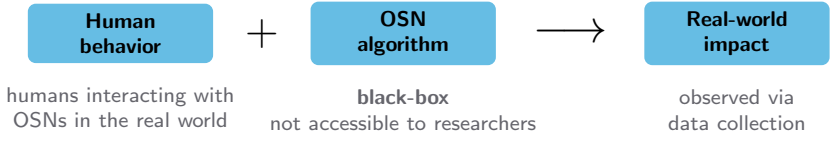


- does not allow to try out different OSN algorithms
- **causality testing difficult**

the agent-based approach

- allows trying out different OSN algorithms
- **causal model:** simulate the impact of interventions
- **vision:** develop technical conditions to regulate OSN algorithms

the empirical approach



- cannot try out different OSN algorithms
- **causality testing difficult**

Ok, but...

What are opinion formation models and how do you model interventions?

overview of this tutorial

- opinion formation models
  - the DeGroot model (consensus)
  - the Friedkin–Johnsen (FJ) model (disagreement & polarization)
  - properties of the FJ model
- algorithmic aspects and interventions for moderating opinions
  - polarization and disagreement indices
  - efficiently estimating user opinions and indices
  - maximizing opinions / minimizing polarization and disagreement
  - emergence of echo chambers

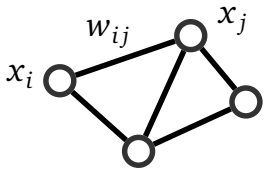
the DeGroot and Friedkin–Johnsen (FJ) models

models of opinion formation

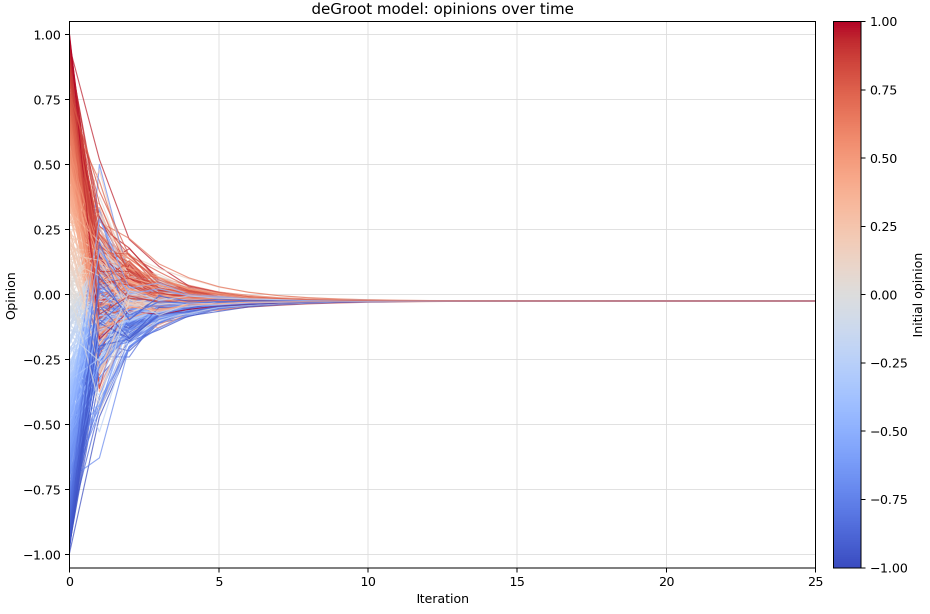
a basic model [DeGroot, 1974]

- we consider a weighted graph modeling a social network
- weight w_{ij} represents influence of node j on i (i trusts j)
- node i has opinion $x_i \in [0,1]$
- node i updates its opinion by

$$x_i^{(t+1)} = \frac{\sum_{j\,|\,(i,j)\in E} w_{ij} x_j^{(t)}}{\sum_{j\,|\,(i,j)\in E} w_{ij}}$$

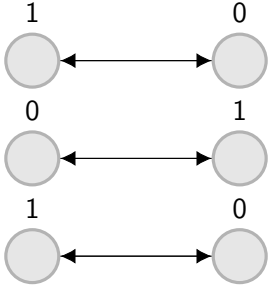


an example of the DeGroot model



properties of the DeGroot model

convergence is not guaranteed



properties of the DeGroot model

Lemma

let G be strongly connected; the DeGroot process converges if and only if G is aperiodic

(a graph is aperiodic if the greatest common divisor of the length of its cycles is 1)

properties of the DeGroot model

What is the consensus value?

[Golub and Jackson, 2010]

Lemma

let G be an undirected connected graph. If the Laplacian DeGroot process converges, then

$$x^{(t)} \to \bar{x}1 \quad \textit{where} \quad \bar{x} = \frac{1}{n}\sum_i x_i^{(0)}$$

more generally: write $L = D - A$ for the graph Laplacian

if a Laplacian DeGroot process converges to consensus, then

$$\bar{x} = v^\mathsf{T} x^{(0)} \quad \text{where} \quad v^\mathsf{T} L = 0^\mathsf{T},\ v^\mathsf{T} 1 = 1$$

takeaway: the DeGroot model

- the DeGroot model is very fundamental:
  easy to explain, analytical solution, many useful convergence results
- under standard assumptions it converges to one consensus opinion
- this makes it a poor model for studying polarization
- to study polarization and algorithmic interventions, we need more expressive opinion dynamics models

the Friedkin-Johnsen model

[Friedkin and Johnsen, 1990]

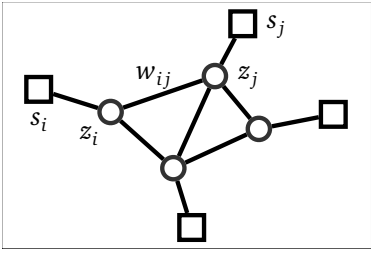
node i has an innate opinion s_i and an expressed opinion $z_i^{(t)}$.

FJ update rule:

$$z_i^{(t+1)} = \frac{s_i + \sum_{j \mid (i,j) \in E} w_{ij} z_j^{(t)}}{1 + \sum_{j \mid (i,j) \in E} w_{ij}}$$

innate opinions are fixed and do not change
$\Rightarrow$ models internal beliefs due to upbringing, etc.

expressed opinion are updated in each round
$\Rightarrow$ models peer pressure



the Friedkin-Johnsen model

[Friedkin and Johnsen, 1990]

node i has an innate opinion s_i and an expressed opinion $z_i^{(t)}$.

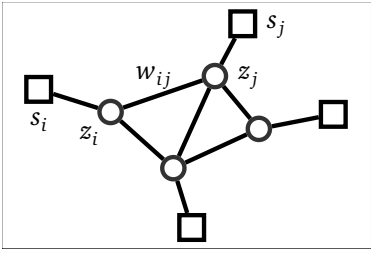
FJ update rule:

$$z_i^{(t+1)} = \frac{s_i + \sum_{j\,|\,(i,j)\in E} w_{ij} z_j^{(t)}}{1 + \sum_{j\,|\,(i,j)\in E} w_{ij}}$$

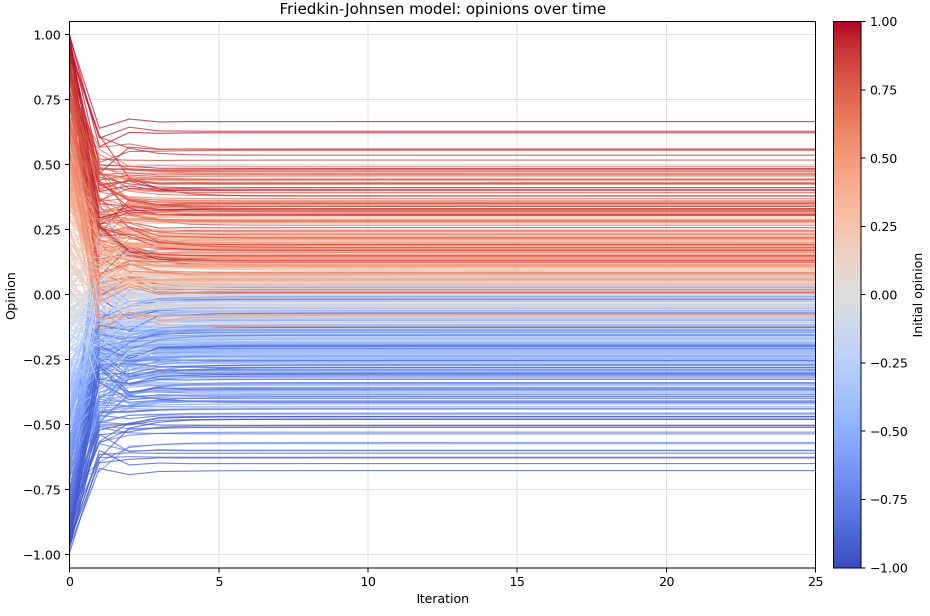
contrast with DeGroot:

$$x_i^{(t+1)} = \frac{\sum_{j\,|\,(i,j)\in E} w_{ij} x_j^{(t)}}{\sum_{j\,|\,(i,j)\in E} w_{ij}}$$

FJ adds a fixed anchor s_i: agents listen to neighbors, but remain pulled toward their innate opinion.



the Friedkin-Johnsen model



notation for the Friedkin–Johnsen model

for an undirected weighted graph $G = (V, E, w)$ with n vertices and m edges:

- adjacency matrix A: $A_{ij} = w_{ij}$ if $(i, j) \in E$, and $A_{ij} = 0$ otherwise
- (weighted) degree $d_i = \sum_j A_{ij}$
- degree matrix $D = \mathrm{diag}(d_1, \ldots, d_n)$
- graph Laplacian $L = D - A$
- identity matrix I of size $|V| \times |V|$

convergence of the Friedkin-Johnsen model

Lemma

for an undirected graph, the FJ process converges to a unique equilibrium

$$z^* = (I + L)^{-1} s$$

proof idea: at convergence, $z_i^{(t+1)} = z_i^{(t)} = z_i^*$, so

$$z_i^* = \frac{s_i + \sum_j A_{ij} z_j^*}{1 + d_i}$$

in matrix–vector notation:

$$z^* = (I + D)^{-1}(s + Az^*)$$

$$\implies (I + D - A) z_i^* = s$$

hence, we have that

$$(I + L) z^* = s \quad \Rightarrow \quad z^* = (I + L)^{-1} s$$

takeaway: the Friedkin–Johnsen model

- the FJ model keeps the simplicity of averaging, but adds innate opinions
- unlike the deGroot model, the equilibrium opinions typically do not reach a consensus
- once opinions can differ, we can define and study polarization, disagreement, and controversy
- equilibrium opinions still have nice analytical form:

$$z^* = (I + L)^{-1} s$$

- it connects to familiar network tools, including random walks, electrical networks, and PageRank-style diffusion

next up

- some properties of the FJ model
- how to compute opinions and indices efficiently
- understanding the impact of algorithmic interventions
  — by studying optimization problems

we will assume that graphs are undirected for the remainder of the tutorial

property of the expressed opinions

- other justifications for the update rule of expressed opinions?

$$z_i^{(t+1)} = \frac{s_i + \sum_{j \,|\, (i,j) \in E} w_{ij} z_j^{(t)}}{1 + \sum_{j \,|\, (i,j) \in E} w_{ij}}$$

- for node i, consider the cost function

$$(z_i^{(t)} - s_i)^2 + \sum_{j \,|\, (i,j) \in E} w_{ij} (z_i^{(t)} - z_j^{(t)})^2$$

  - first term corresponds to conflict between internal and expressed opinion
  - second term corresponds to i's conflict with its neighbors
- if node i sets $z_i^{(t+1)}$ to minimize this cost function, the choice of $z_i^{(t+1)}$ is the same as in the update rule above

the price of anarchy in opinion formation

[Bindel et al., 2015]

- how bad is forming your own opinion?
- in the FJ model, each node is independently minimizing its own cost

$$c_i(z_i) = (z_i - s_i)^2 + \sum_{j \mid (i,j) \in E} w_{ij}(z_i - z_j)^2$$

this results to a Nash equilibrium

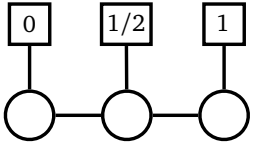
- what instead if we ask to optimize the social cost

$$c(\mathbf{y}) = \sum_{i \in V} c_i(y_i)$$

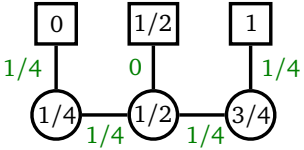
- theorem ([Bindel et al., 2015])
  price of anarchy (ratio of costs) is at most $9/8$ for any undirected graph G

→ this result is for undirected networks;
for directed networks the price of anarchy can be much higher

the price of anarchy in opinion formation — example [Bindel et al., 2015]

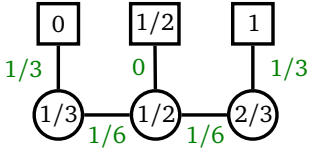


Nash equilibrium



Nash cost $= 3 * 2 * (1/4)^2 = 3/8$

Social optimal



Opt cost $= 2 * ((1/3)^2 + (1/6)^2) + 2 * (1/6)^2 = 1/3$

$$\text{Price of anarchy} = \frac{\text{Nash cost}}{\text{Opt cost}} = \frac{3/8}{1/3} = \frac{9}{8}$$

quantities of interest in the Friedkin-Johnsen model

- given the equilibrium expressed opinions z^* and innate opinions s, we can study more complex phenomena in the network
- we can quantify polarization, disagreement, ...

sum of opinions	$\mathcal{S} = \sum_{i \in V} z_i^*$
polarization index	$\mathcal{P} = \sum_{i \in V} (z_i^* - \bar{z})^2$
controversy index	$\mathcal{C} = \sum_{i \in V} (z_i^*)^2$
internal-conflict index	$\mathcal{I} = \sum_{i \in V} (z_i^* - s_i)^2$
disagreement index	$\mathcal{D} = \sum_{(i,j) \in E} w_{ij} (z_i^* - z_j^*)^2$
polarization-disagreement index	$\mathcal{I}_{pd} = \mathcal{P} + \mathcal{D}$

sum of opinions: sums all node opinions — relevant for marketing campaigns polarization: the variance of the opinions,

quantities of interest in the Friedkin-Johnsen model

sum of opinions	$\mathcal{S} = \sum_{i \in V} z_i^*$	$= 1^\intercal z^* = 1^\intercal (I + L)^{-1} s$
polarization index	$\mathcal{P} = \sum_{i \in V} (z_i^* - \bar{z})^2$	$= s^\intercal (I + L)^{-1} (I - \frac{11^\intercal}{n})(I + L)^{-1} s$
controversy index	$\mathcal{C} = \sum_{i \in V} (z_i^*)^2$	$= s^\intercal (I + L)^{-2} s$
internal-conflict index	$\mathcal{I} = \sum_{i \in V} (z_i^* - s_i)^2$	$= s^\intercal (I + L)^{-1} L^2 (I + L)^{-1} s$
disagreement index	$\mathcal{D} = \sum_{(i,j) \in E} (z_i^* - z_j^*)^2$	$= s^\intercal (I + L)^{-1} L (I + L)^{-1} s$

using that $z^* = (I + L)^{-1} s$, we can express these measures as quadratic forms where 1 is the all-ones vectors, I is the identity matrix, L is the graph Laplacian, s is the vector of innate opinions and $\bar{z} = \frac{1}{|V|} \sum_{i \in V} z_i^*$. **all of the matrices are positive semidefinite!**

sums of innate and expressed opinions

lemma. for undirected graphs it holds that $\sum_i z_i^* = \sum_i s_i$

proof:

- note that $\sum_i z_i^* = 1^\intercal z^*$ and $\sum_i s_i = 1^\intercal s$
- recall that $z^* = (I + L)^{-1}$ and $1^\intercal L = 0^\intercal$ since L is the Laplacian with $L = D - A$
- since $s = (I + L)z^*$ we have that

$$1^\intercal s = 1^\intercal (I + L)z^* = (1^\intercal + 0^\intercal)z^* = 1^\intercal z^*$$

list of useful properties (selection)

- the equilibrium opinions z^* do not depend on the initial opinions $z^{(0)}$
- we assumed opinions in $[-1, 1]$. when moving to intervals in $[0, 1]$ or to $[-a, b]$, all indices can be rescaled accordingly — except controversy
- for controversy $= \sum_i (z_i^*)^2$ the interpretation is very different depending on whether the opinions are in $[-1, 1]$ or in $[0, 1]$
- for the matrices $\mathcal{M}$ in the quadratic forms of polarization, disagreement and internal-conflict, it holds that $\mathcal{M}1 = 0$

use that $L1 = 0$, $(I + L)^{-1}1 = 1$ and $(I - \frac{11^\mathsf{T}}{n})1 = 0$

- if z^* (or s) is mean-centered, i.e., $\sum_i z_i^* = 0$,
  then controversy $\mathcal{C}$ and polarization $\mathcal{P}$ are identical
- it holds that $\sum_i z_i^* = \sum_i s_i$
- the matrices in the quadratic forms of the indices are positive semidefinite
- conservation law of conflict: $\mathcal{I} + 2\mathcal{D} + \mathcal{C} = s^\mathsf{T} s$ [Chen et al., 2018]

algorithmic aspects and interventions for moderating opinions

from definitions to computation

- previously, we have seen a lot of definition of expressed opinions, polarization, ...
- for example, polarization $= \sum_{i \in V} (z_i^* - \bar{z})^2$

$\longrightarrow$ we want to compute the expressed opinions z^*

- reminder: $z^* = (I + L)^{-1} s$
- problems:
  - computing the matrix inverse $(I + L)^{-1}$ takes cubic time in practice
  - for connected graphs, the inverse $(I + L)^{-1}$ has $\Omega(n^2)$ non-zero entries
    — even storing it is costly

$\longrightarrow$ How to compute z^* and the indices efficiently?

efficiently computing expressed opinions and indices

[Xu et al., 2021]

- good news: z^* is the solution to the linear system $(I + L)x = s$
- we can compute ϵ-approximation of z^* in time $\tilde{O}(m)$, where m is the number of edges in the graph [Cohen et al., 2014]
- based on this approximation of z^*, we can compute ϵ-approximations of polarization, disagreement, ... in near-linear time

$\longrightarrow$ orders of magnitude faster than computing matrix inverse

$\longrightarrow$ in practice often enough to iterate opinion update equation a few times

interventions

- how sensitive are polarization, disagreement, ... to interventions?
- examples for interventions: a timeline algorithm changes the network structure, an adversary makes people change their innate opinions, ...

- *formal way to study this:* define an optimization problem, where:
  - the objective function encodes the desired goal
  - the constraints encode the "power" of the intervention

- example: a social network provider wants to minimize polarization and disagreement by changing the network structure [Musco et al., 2018, Zhu et al., 2021]
- example: an adversary wants to maximize the disagreement and has the power to change k node opinions [Chen and Racz, 2021, Gaitonde et al., 2020]

interventions: realistic? goal?

- "tampering with timeline algorithms seems unethical?"
  - studies conducted using simulations
  - goal of this area is to understand the "usefulness" and practicality of certain interventions
  - to decide what is ethical in practice, we need an interdisciplinary discussion, including philosophers, social scientists and policymakers

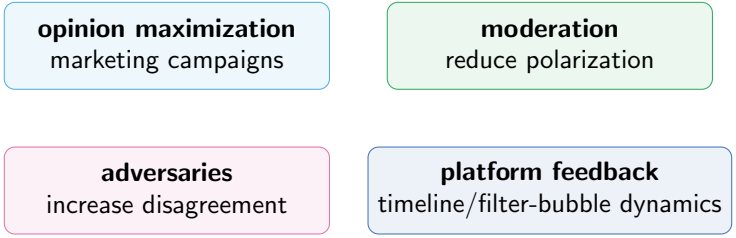
- "but it seems quite unrealistic that an adversary can change k node opinions arbitrarily"
  - perhaps. but we need to understand "simple" adversaries before we can consider more complicated ones
  - if it turned out that even "too powerful" adversaries have little impact on polarization and disagreement, the same will hold for weaker ones

interventions: literature overview

- what to optimize
  - minimize price of anarchy [Bindel et al., 2015]
  - reduce polarization and disagreement [Matakos et al., 2017, Musco et al., 2018]
  - maximize sum of opinions [Gionis et al., 2013, Tu and Neumann, 2022]
  - increase disagreement [Chen and Racz, 2021, Gaitonde et al., 2020]

- what properties to modify
  - innate or expressed opinions [Gionis et al., 2013, Matakos et al., 2017]
  - graph weights [Abebe et al., 2018]
  - graph structure [Bindel et al., 2015, Musco et al., 2018] [Zhu et al., 2021, Rácz and Rigobon, 2022]

roadmap for interventions



- same template throughout: choose an objective, specify what can be changed, then optimize

opinion maximization

- goal: make the overall opinion in the network as positive as possible
- intervention: choose a small number of nodes to influence directly
- algorithmic theme: monotonicity, submodularity, greedy selection

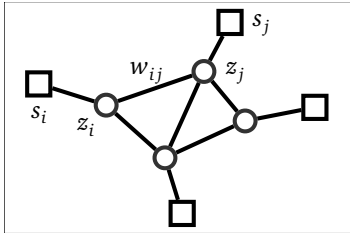
opinion maximization in social networks

[Gionis et al., 2013]

- select k nodes to set their expressed opinion to $z_i^* = 1$ so as to maximize the sum of opinions

$$\mathcal{S} = \sum_{i \in V} z_i^*$$

  - motivation: lobbying for a cause or campaign



- GREEDY gives $(1 - 1/e)$ approximation
- objective function is monotone and submodular
- technical observation: consider an absorbing random walk, with absorbing states the nodes that correspond to the innate opinions. then z_i^* is can be interpreted as the expected value at absorption, when starting a random walk in node i

moderating polarization and disagreement

- goal: reduce controversy, polarization, or disagreement
- interventions: moderate selected opinions or change edges/weights
- algorithmic theme: hardness, approximation, and convex optimization

measuring and moderating opinion polarization in social networks

[Matakos et al., 2017]

- opinions are assumed to be in the interval $[-1, 1]$, furthermore assume $\bar{z} = 0$
- goal is to minimize controversy index $\mathcal{C} = \sum_{i \in V} (z_i^*)^2$
  - equivalent to polarization index $\mathcal{P} = \sum_{i \in V} (z_i^* - \bar{z})^2$, since $\bar{z} = 0$
- select k nodes to set $z_i = 0$, or $s_i = 0$ (i.e., become moderate), to minimize $\mathcal{C}$
- authors show that the problem is **NP**-hard
- they propose a binary orthogonal matching pursuit (BOMP) algorithm
- experimentally compare BOMP with GREEDY, PAGERANK, and other baselines
  - BOMP with GREEDY are the best-performing methods

minimizing polarization and disagreement in social networks

[Musco et al., 2018]

- focus on minimizing the following indices
  - polarization: $\mathcal{P} = \sum_{i \in V} (z_i^* - \bar{z})^2$
  - disagreement: $\mathcal{D} = \sum_{(i,j) \in E} w_{ij} (z_i^* - z_j^*)^2$
  - polarization-disagreement: $\mathcal{I}_{pd} = \mathcal{P} + \mathcal{D}$
- constraint: we can decrease the innate opinions within a given budget and ℓ_1-distances, i.e., $||s - s'||_1 \leq B$ and $s' \leq s$
- result: optimizing these indices is convex and can be solved in polynomial time
- what if we can change the graph topology with a fixed number of edges?
  - minimizing $\mathcal{I}_{pd}$ is convex
  - thus, it can be solved with standard-convex optimization methods
  - when one of the terms $\mathcal{P}$ or $\mathcal{C}$ is weighted differently, problem is not convex

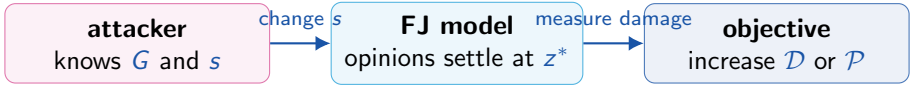
opinion formation with adversaries

- goal: understand how much damage an attacker can cause
- intervention: radicalize or perturb a limited number of opinions
- algorithmic theme: full information vs limited information

maximizing disagreement — understanding the power of adversaries

[Chen and Racz, 2021, Gaitonde et al., 2020]

- suppose an adversary can change k innate opinions arbitrarily
  - full-information setting: the attacker knows G and s
  - russia is hacking bluesky accounts and then spreading propaganda[1]
- how much can the adversary increase polarization and disagreement?
- algorithmic question: which nodes should the attacker target?



[1] https://www.nytimes.com/2026/05/21/business/bluesky-russia-hacking-accounts.html

maximizing disagreement — understanding the power of adversaries

[Chen and Racz, 2021, Gaitonde et al., 2020]

- let $\mathcal{P}$ and $\mathcal{D}$ be the original polarization and disagreement; let $d_{\max}$ be the weighted maximum degree
- if the adversary can change k innate opinions arbitrarily:

$$\mathcal{P}' \leq \mathcal{P} + 3k \qquad \mathcal{D}' \leq \mathcal{D} + 8d_{\max}k$$

- in practice, greedy algorithms can increase polarization and disagreement linearly in k
- if the adversary can choose any innate-opinion vector s with $\|s\|_2 \leq R$:

$$s^{\mathsf{T}}(I + L)^{-1}L(I + L)^{-1}s \leq \frac{R^2}{4}$$

- proof idea: bound the largest eigenvalue of $(I + L)^{-1}L(I + L)^{-1}$

adversaries with limited information

[Tu et al., 2023]

- previous adversary models are often full information:
  - the attacker knows the graph G and all innate opinions s_0
- in practice, opinions are private, noisy, or hard to estimate
  - the network topology is often much easier to observe
- limited information model: the attacker knows only G
  - choose k nodes and radicalize their innate opinions
- question: can topology alone identify nodes whose change increases disagreement or polarization?

full information vs limited information

[Tu et al., 2023]

full information

- input: G and s_0
- can evaluate the real effect of changing a candidate node
- chooses $X \subseteq V$, $|X| = k$, using the actual opinions

limited information

- input: only G
- uses a topology-based proxy for influence on discord
- chooses $X \subseteq V$, $|X| = k$, before seeing the opinions

after choosing X, the attack is the same: set selected innate opinions to an extreme value, e.g. $s_i = 1$ for $i \in X$.

greedy attacks: full vs limited information

[Tu et al., 2023]

GREEDY-F

1. start with $X = \emptyset$
2. for each candidate u, evaluate the true gain from attacking $X \cup \{u\}$
3. add the candidate with largest marginal gain
4. repeat until $|X| = k$

uses: graph and innate opinions

GREEDY-L

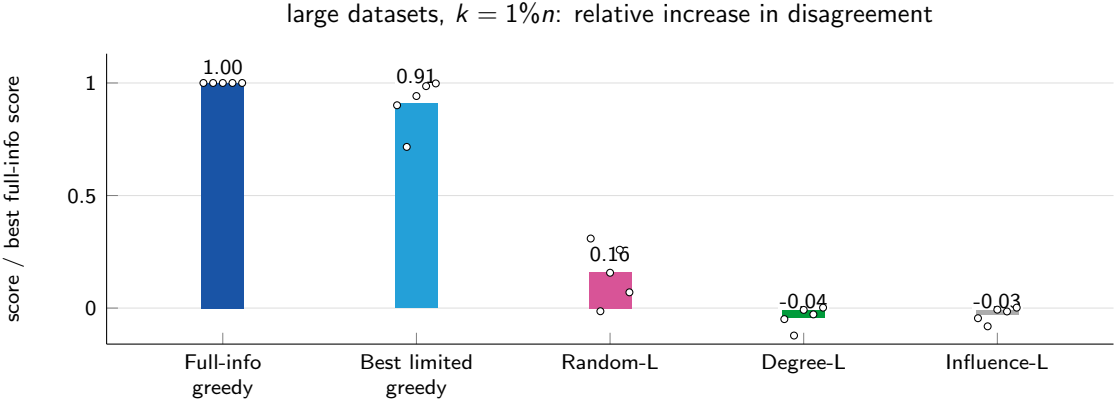
1. start with $X = \emptyset$
2. for each candidate u, score the gain using a topology-only proxy
3. add the candidate with largest proxy gain
4. repeat until $|X| = k$

uses: graph only

same greedy template; the difference is whether marginal gains are scored with real opinions or only with topology.

limited information: representative results

[Tu et al., 2023]



- best limited-information greedy reaches about 91% of the full-information greedy score on average
- topology matters: simple degree, influence-maximization, and random baselines are much weaker

platform feedback and filter bubbles

- goal: model repeated interaction between nodes and a timeline/recommender system
- intervention: the platform changes the network exposure pattern
- algorithmic theme: local optimization can create global polarization

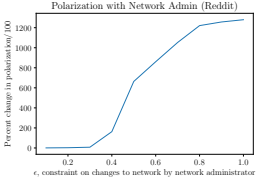
analyzing the impact of filter bubbles on social network polarization

[Chitra and Musco, 2020]

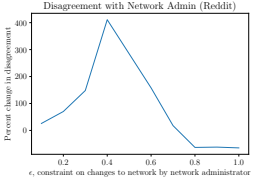
- study the interplay between nodes and a network administrator
- the dynamics proceed in iterations — in each iteration
  - the nodes adjust their expressed opinions according to the FJ model
  - the network administrator slightly adjusts the network to minimize disagreement $\mathcal{D}$

  until convergence
- intuition: network administrators want less disagreement, as this implies "happier" users
- it is shown experimentally that polarization increases
- authors suggest this explains why recommender systems increase polarization and introduce filter bubbles

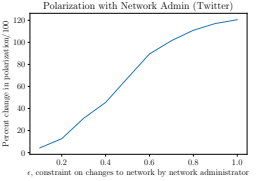
analyzing the impact of filter bubbles on social network polarization



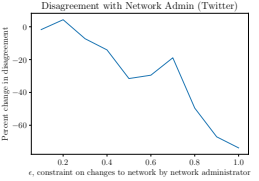
(a) Change in polarization, Reddit network



(b) Change in disagreement, Reddit network



(c) Change in polarization, Twitter network



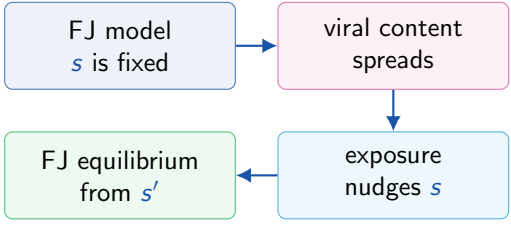
(d) Change in disagreement, Twitter network

[Chitra and Musco, 2020]

incorporating viral information

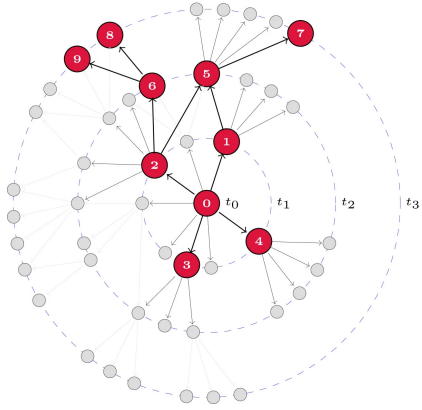
- goal: model how viral information change user opinions and polarization
- intervention: nodes update their innate opinions after seeing viral content
- algorithmic theme: information cascade can increase global polarization

opinion dynamics meets viral content



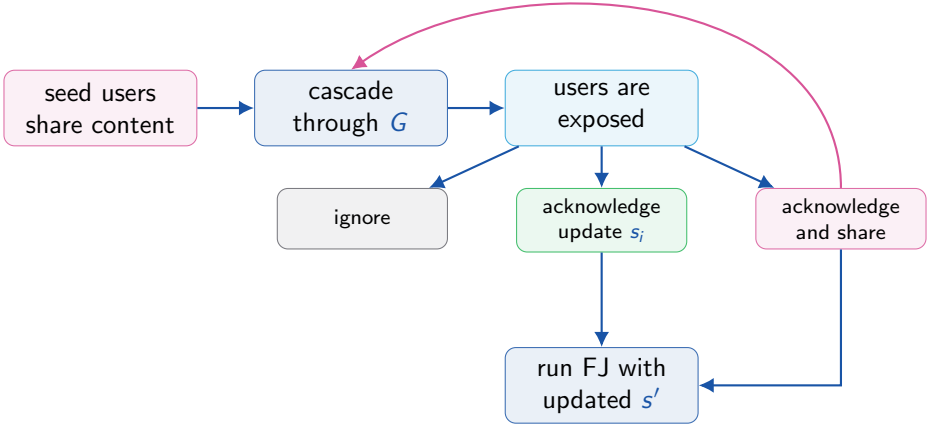
- key idea: exposure can move s before opinions settle
- independent cascade: who sees the content?
- Friedkin–Johnsen: what opinions emerge?

[Tu and Neumann, 2022]



the Spread–Acknowledge model

[Tu and Neumann, 2022]



- when nodes *acknowledge* the content, they update their innate opinion s_i
- a useful equivalent view: first reveal the cascade, then compute $z^* = (I + L)^{-1} s'$

content types matter

[Tu and Neumann, 2022]

marketing content

like it or ignore it

$$s_i' = \min\{1,\, s_i + \epsilon\}$$

Emmanuel Macron
Republic on the Move

Marine Le Pen
National Rally

polarizing content

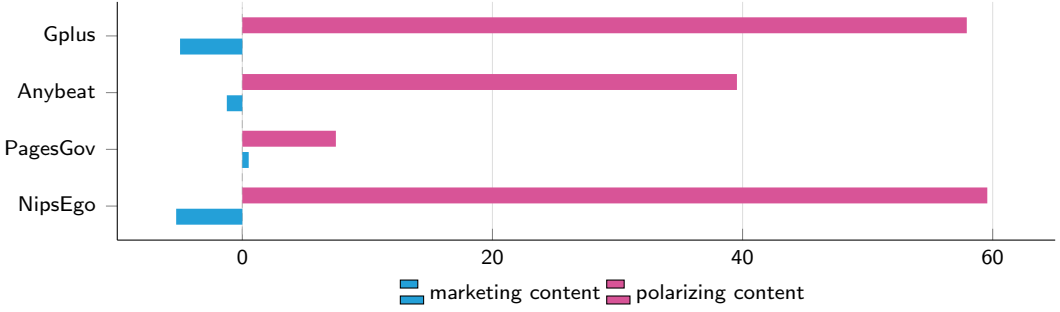
like it or hate it

$$s_i' = \begin{cases} s_i + \epsilon, & s_i \geq \tau, \\ s_i - \epsilon, & s_i < \tau. \end{cases}$$

- same spreading process; different response to exposure

representative empirical result

[Tu and Neumann, 2022]



- with only 0.5% seed users, polarizing content can strongly increase polarization
- marketing content mostly shifts opinions, with much smaller polarization effects

conclusion, limitations, reflections, future directions

summary

- opinion formation in social networks is an active area of research
  - work both in mathematical modeling and computational social science
- in this tutorial, we reviewed common opinion-formation models
  - DeGroot and Friedkin-Johnsen models, other opinion formation models
  - discussed properties of the models and measures of interest
- discussed how polarization may emerge from these models
  - e.g., emergence of echo chambers
- reviewed computational aspects and interventions for moderating opinions
- no discussion on misinformation and disinformation — need a separate tutorial

challenges, limitations

- follow network is losing relevance on social media,
  users are served more out-of-network content
- more and more AI-generated content on platforms
- validation of the mathematical models is very challenging
  - models are often too simplistic, e.g., opinions in [0,1], opinions are updated by a simple weighted-averaging operation
  - models involve parameters that are difficult to estimate in practice

ethical issues on interventions

a common intervention action is to aim to reduce polarization, or increase diversity, by making judicious recommendations

Q: is it ethical to tamper with users' feed?

Q: can such methods facilitate manipulation?

A: UI, user control, and transparency needs to be addressed separately

A: content prioritization and recommendation algorithms are already in place, and they

- are mainly aiming at increasing engangement and monetization
- are not transparent
- are not offering control to the users
- do not have built-in ethical specifications

directions for future work

- validating existing models / developing better models that fit the real world
- incorporate different modalities
  - follow graph / likes / posts / comments
  - natural language processing
- modeling the roles of different users / incorporate personalization
  - but this makes validation only harder
- transparent methods for interventions to reduce polarization / maximize diversity
- which adversary models are realistic in the real world?
- obtain understanding of more complicated adversaries and timeline algorithms
- most works consider adversaries who make a change once — what about feedback loops?
- how can we defend networks against attackers?

A Survey on Algorithmic Interventions in Opinion Dynamics

Miyauchi, Kuroki, Cinus, Neumann, Bonchi (2026)

arxiv.org/abs/2603.10756

recent survey covering objectives, intervention models, algorithms, and future directions

thank you!

questions?

slides

neumannstefan.com/files/
essai-tutorial.pdf

references I


Abebe, R., Kleinberg, J., Parkes, D., and Tsourakakis, C. E. (2018).
Opinion dynamics with varying susceptibility to persuasion.
In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1089–1098.


Bindel, D., Kleinberg, J., and Oren, S. (2015).
How bad is forming your own opinion?
Games and Economic Behavior, 92:248–265.


Chen, M. F. and Racz, M. Z. (2021).
An adversarial model of network disruption: Maximizing disagreement and polarization in social networks.
IEEE Transactions on Network Science and Engineering, 9(2):728 – 739.


Chen, X., Lijffijt, J., and Bie, T. D. (2018).
Quantifying and minimizing risk of conflict in social networks.
In *KDD*, pages 1197–1205.


Chitra, U. and Musco, C. (2020).
Analyzing the impact of filter bubbles on social network polarization.
In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123.

references II


Cohen, M. B., Kyng, R., Miller, G. L., Pachocki, J. W., Peng, R., Rao, A. B., and Xu, S. C. (2014).
Solving SDD linear systems in nearly $m \log^{1/2} n$ time.
In *STOC*, pages 343–352.


DeGroot, M. H. (1974).
Reaching a consensus.
Journal of the American Statistical Association, 69(345):118–121.


Friedkin, N. E. and Johnsen, E. C. (1990).
Social influence and opinions.
Journal of Mathematical Sociology, 15(3-4):193–206.


Gaitonde, J., Kleinberg, J., and Tardos, E. (2020).
Adversarial perturbations of opinion dynamics in networks.
In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 471–472.


Gionis, A., Terzi, E., and Tsaparas, P. (2013).
Opinion maximization in social networks.
In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 387–395. SIAM.

references III


Golub, B. and Jackson, M. O. (2010).
Naive learning in social networks and the wisdom of crowds.
American Economic Journal: Microeconomics, 2(1):112–49.


Matakos, A., Terzi, E., and Tsaparas, P. (2017).
Measuring and moderating opinion polarization in social networks.
Data Mining and Knowledge Discovery, 31(5):1480–1505.


Musco, C., Musco, C., and Tsourakakis, C. E. (2018).
Minimizing polarization and disagreement in social networks.
In *Proceedings of the 2018 World Wide Web Conference*, pages 369–378.


Rácz, M. Z. and Rigobon, D. E. (2022).
Towards consensus: Reducing polarization by perturbing social networks.
CoRR, abs/2206.08996.


Tu, S. and Neumann, S. (2022).
A viral marketing-based model for opinion dynamics.
In *WebConf*.

references IV


Tu, S., Neumann, S., and Gionis, A. (2023).
Adversaries with limited information in the friedkin-johnsen model.
In *ACM International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2201–2210.


Xu, W., Bao, Q., and Zhang, Z. (2021).
Fast Evaluation for Relevant Quantities of Opinion Dynamics.
In *WWW*, pages 2037–2045.


Zhu, L., Bao, Q., and Zhang, Z. (2021).
Minimizing polarization and disagreement in social networks via link recommendation.
In *NeurIPS*.