Online Social Networks and Media

Polarization

Fairness

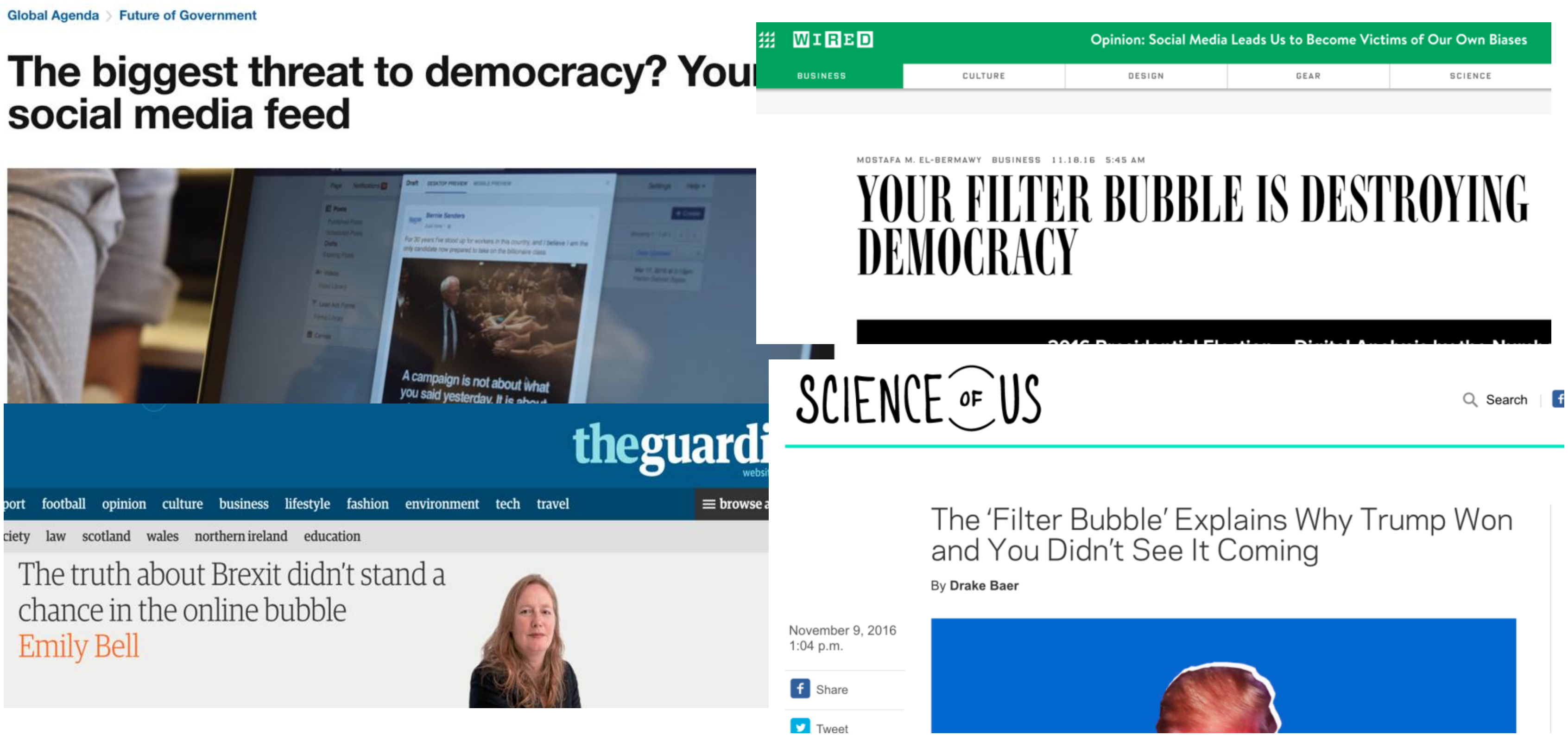
Polarization

Slides taken from EUROCSS 2019 Tutorial:
Polarization on Social Media

Kiran Garimella, Gianmarco De Francisci Morales,
Michael Mathioudakis, Aristides Gionis

Polarization in Social Networks

- There is a growing concern that social media make the public more polarized and extreme



polarization + online bubble → more polarization

What is polarization?

- The term is used in various domains with similar meaning
- Political polarization (Wikipedia) *"the divergence of political attitudes to ideological extremes."*
  https://en.wikipedia.org/wiki/Polarization_(politics)
- Social polarization *"the segregation within a society that may emerge from income inequality, real-estate fluctuations, economic displacements, etc."*
  https://en.wikipedia.org/wiki/Social_polarization
- Oxford Dictionary *"Division into two sharply contrasting groups or sets of opinions or beliefs."*
  Ref: https://en.oxforddictionaries.com/definition/polarization

Why is it important to study?

- How we handle disagreement is essential to democratic process
  - A large part of the discussion has moved to social media
- Because polarization might be linked to adverse effects
  - Social segmentation and stereotypes
  - Echo chambers
    - Decrease in deliberation
    - Hinders deliberative democracy
- Need to be aware of our biases
  - Sometimes we might not hear opposing views
  - Biases around us (e.g., algorithmic personalization)
- However, not necessarily negative in itself

Psychological mechanisms of polarization

- Mechanisms that manifest when humans are confronted with information that challenges their beliefs
- Polarization involves…
  - … arguments and counter-arguments
  - … evidence that is conflicting or interpreted differently
  - … different points of view – that might challenge our own
- How do we react to opposing opinions / arguments / evidence that challenge their opinion?
  - Do we update our beliefs? How?
  - Are we influenced by the beliefs of others?
  - Do we use evidence to update our beliefs?
  - Or use our beliefs to judge evidence?
- Psychologists & cognitive scientists have studied these questions for long

Cognitive dissonance

- People experience discomfort when presented with information that challenges their beliefs or decisions

  Fischer et al. “The theory of cognitive dissonance: State of the science and directions for future research.” 2008.

- Extensively studied behavior, theory first formulated in the 1950’s

  Festinger. “A Theory of Cognitive Dissonance.” 1957.

Cognitive dissonance

- ’Cognition’: broadly defined
  - Element of knowledge, belief, value
- ’Dissonance’ – i.e., ’lack of harmony or agreement’
  - Subjective perception of incompatibility / discrepancy between cognitions
  - Psychological discomfort
  - Motivation to reduce discomfort
- Reduce discomfort by…
  - Adding or highlighting consonant cognitions
  - Removing or downplaying dissonant cognitions

Examples of Cognitive Dissonance

- Selective exposure
  Klapper. “The effects of mass communication.” 1960
  - Subjects choose to examine items that agree with their decision
- Biased assimilation
  Lord et al. “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.” 1979
  - Subjects find consonant evidence more convincing
- Free-choice
  Brehm. “Postdecision changes in the desirability of alternatives.” 1956
  - Spreading-apart-of-alternatives
- Induced compliance
  Festinger and Carlsmith. “Cognitive consequences of forced compliance.” 1959
  - Subjects justify their decisions a-posteriori, even if they originally disagreed

Cognitive Bias

- Cognitive dissonance may lead to cognitive bias:
  - *a systematic thought process caused by the tendency of the human brain to simplify information processing through a filter of personal experience and preferences. The filtering process is a coping mechanism that enables the brain to prioritize and process large amounts of information quickly.*

  https://www.techtarget.com/searchenterpriseai/definition/cognitive-bias
  - *a systematic pattern of deviation from norm or rationality in judgment.*

  https://en.wikipedia.org/wiki/Cognitive_bias
- Examples of Cognitive Bias:
  - Confirmation bias
  - Priming
  - Framing
  - Anchoring

Group biases

- Earlier discussion: bias mechanisms at individual level
- Biases can also manifest at group level
- Social identity complexity
  - Individuals associate themselves with social identities
    - race, religion, gender, class

Roccas, S. and Brewer, M.B., 2002. Social identity complexity. Personality and Social Psychology Review.

- Group polarization
  - The tendency for a group to make decisions that are more extreme than the initial inclination of its members

Sunstein, C.R., 2002. The law of group polarization. Journal of political philosophy.

Summary

- Cognitive dissonance prompts people to expose themselves to confirming information
  - What is consonant or dissonant might also depend on group participation

- What could go wrong?

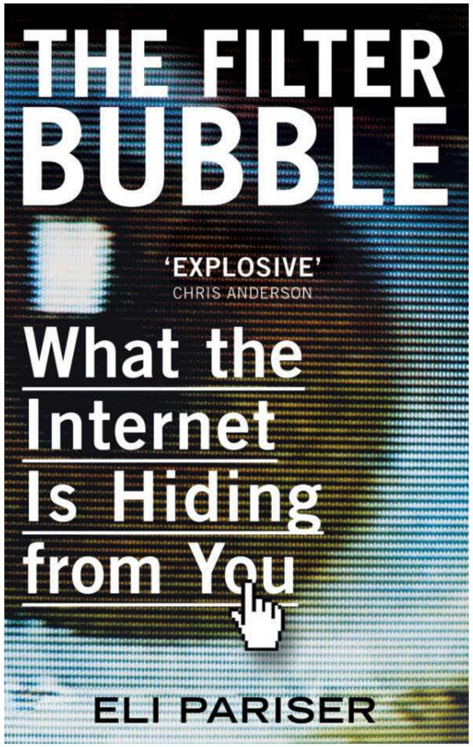
- People share their views on the same platforms they use to consume information
  - Eg: Facebook, Twitter
- If platforms are aware of user views and aim to maximize user satisfaction, what content will they show to users?
  - Why show dissonant content?

Media bias

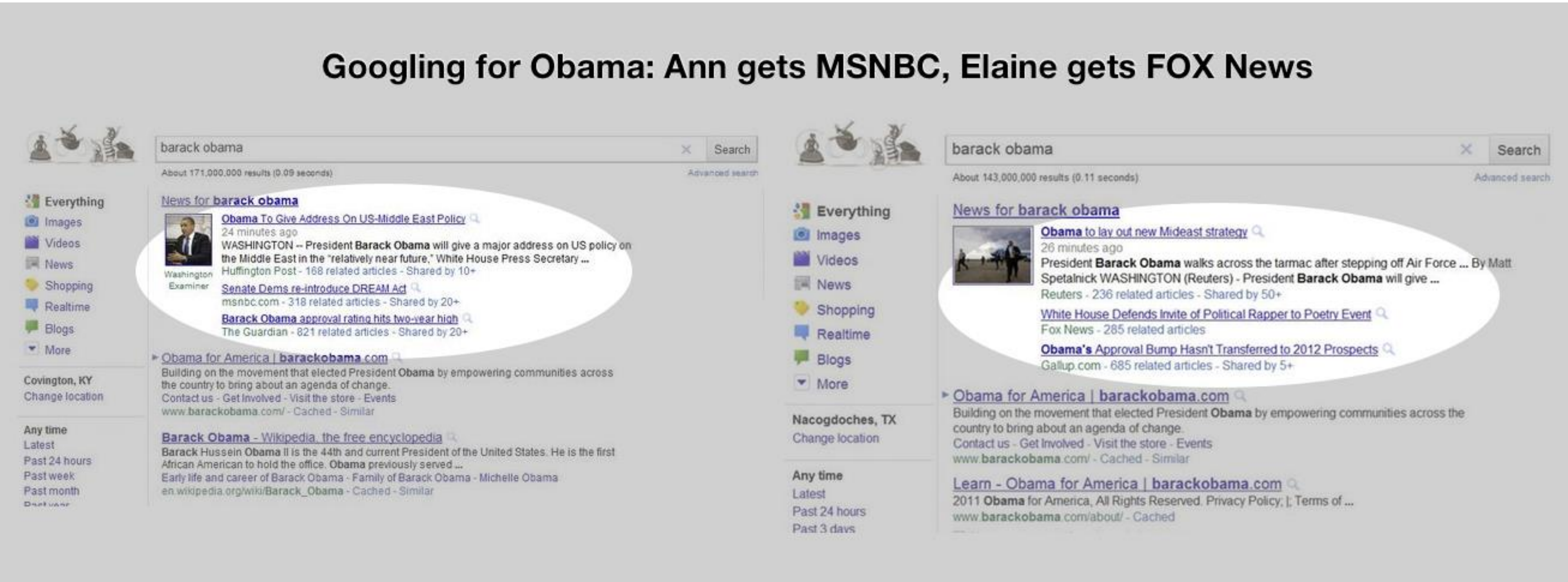
- Media present information differently based on their audience

Algorithmic bias

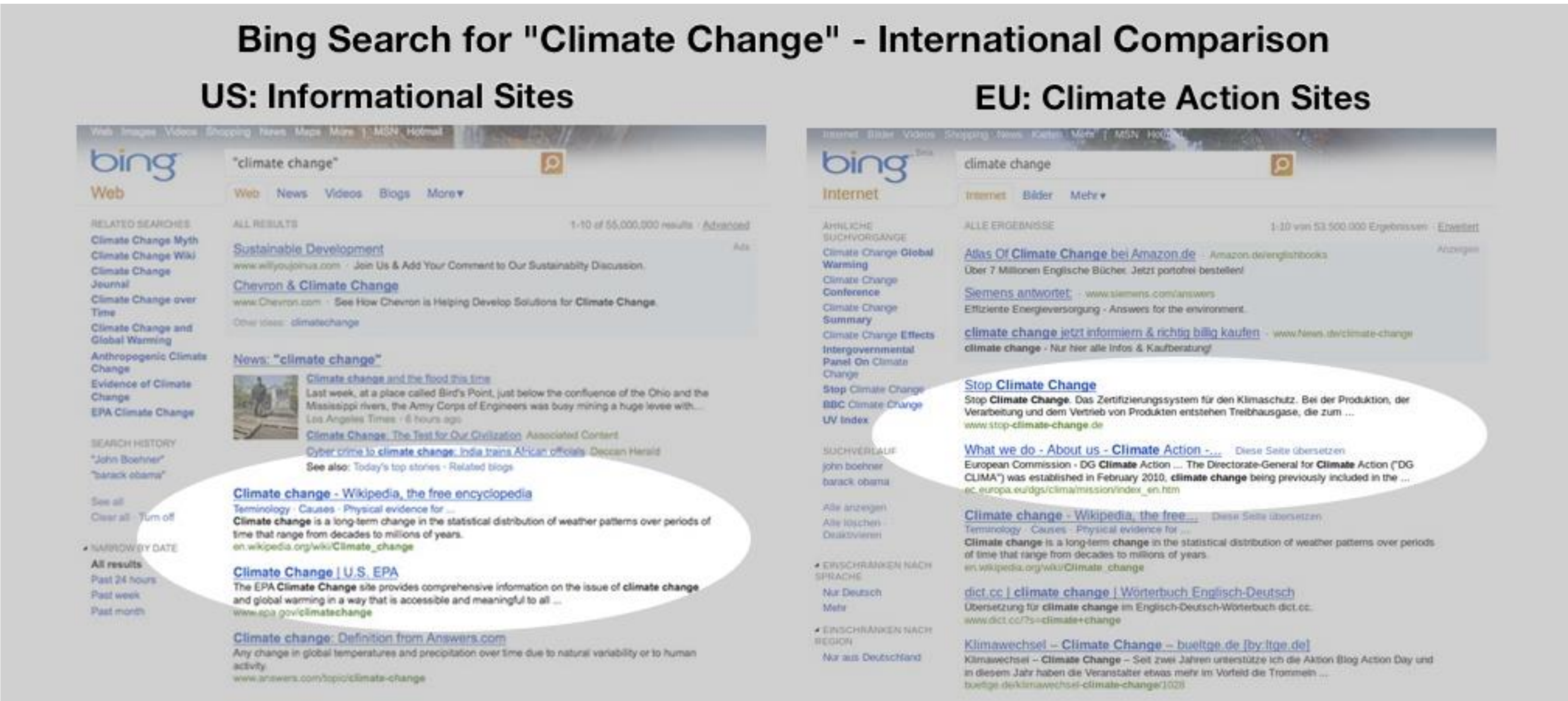
- Online content platforms present information to match individual users
- Algorithmic personalization
  - News
  - Search engines
  - Social media
- Filter bubble
  - We do not see the same content



Filter bubble



Filter bubble

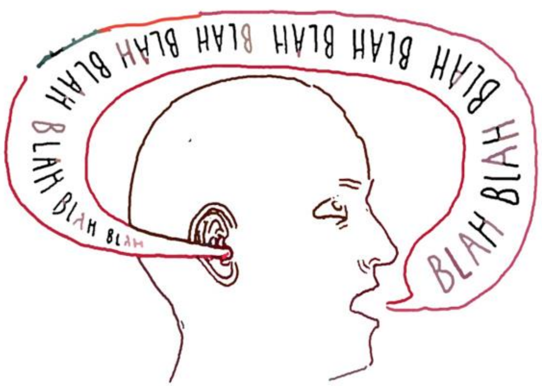


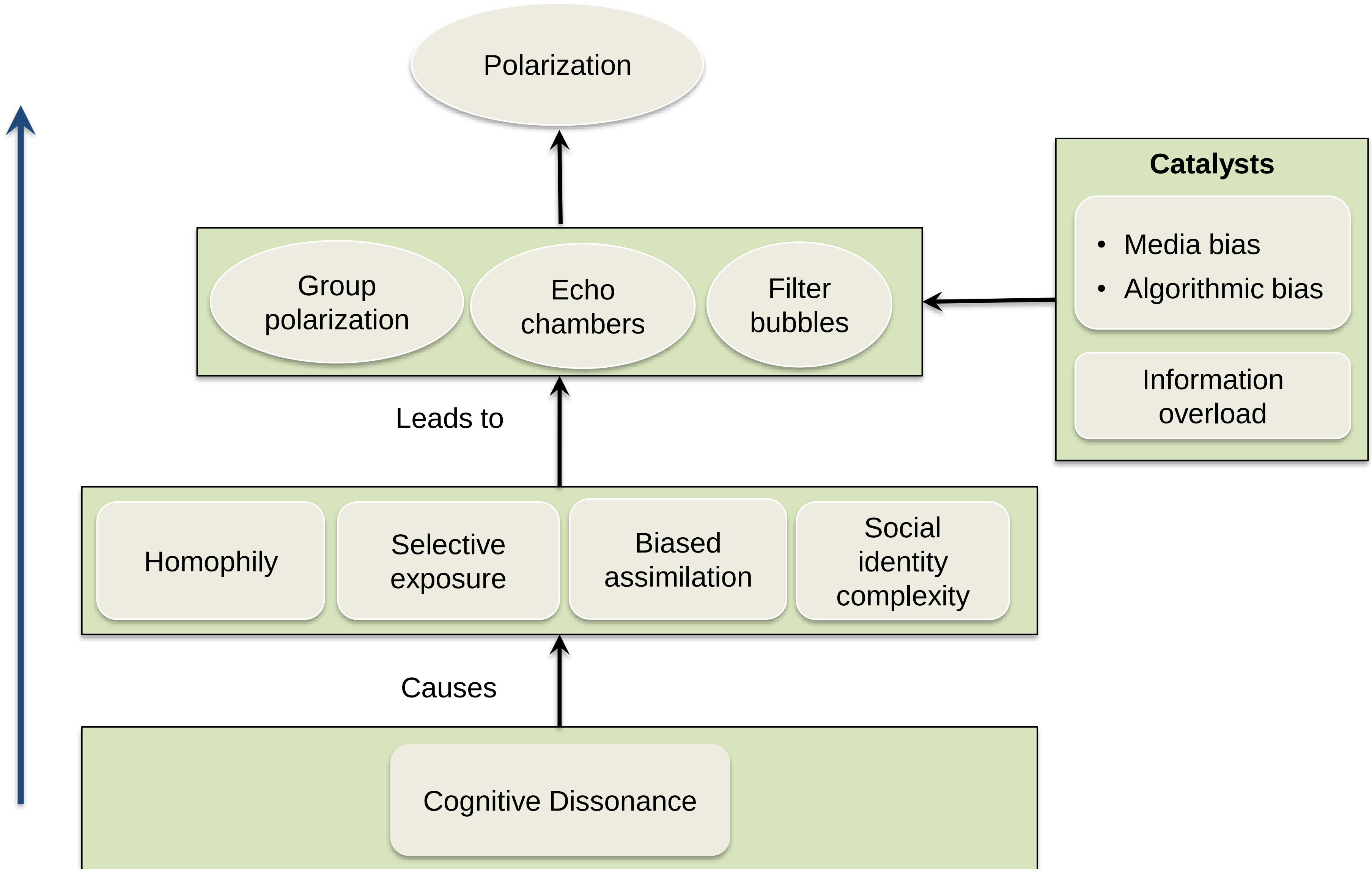
Why the Web might increase polarization

- Increase in available information
- Increase in filtering power
  - People tend to avoid reading conflicting information
- Increase in social feedback (with social media)
  - Homogeneity and group-think reinforced

Echo chambers

‘Tribal enclaves’ in which people hear and reinforce their own opinions





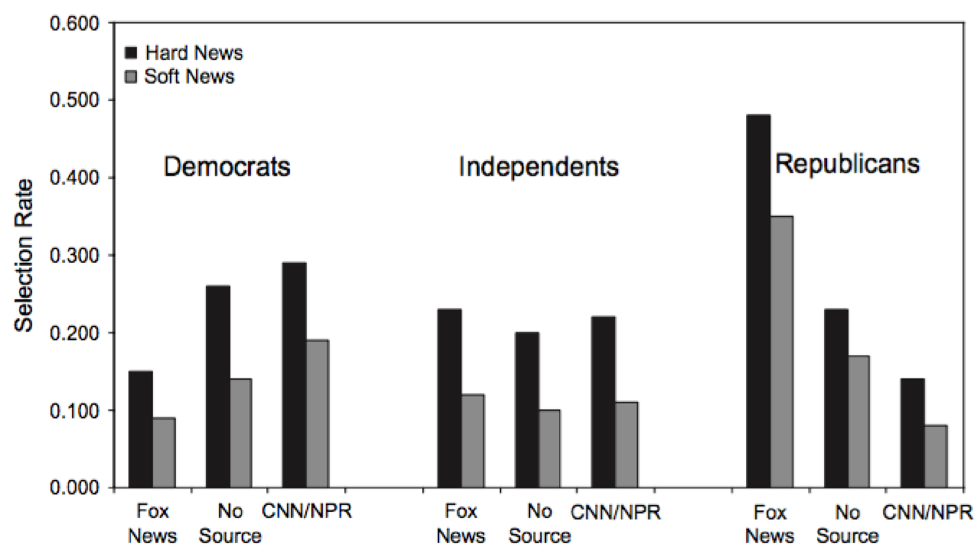
Polarization
Catalysts
Media bias
Algorithmic bias
Information overload
Group polarization
Echo chambers
Filter bubbles
Leads to
Homophily
Selective exposure
Biased assimilation
Social identity complexity
Causes
Cognitive Dissonance

Ideological Selectivity in News

Iyengar, S., & Hahn, K. S. "Red media, blue media: Evidence of ideological selectivity in media use." (2009)

- People prefer to read news from sources close to their leaning
- Finding consistent with selective exposure

- Online user study with randomized experiments in US
- Headlines for 4 articles, labeled randomly as coming from 4 different sources:
  - Fox News, CNN, NPR, BBC
  - Control group sees same stories with no media logo
- 380 stories, 1020 users
- Tendency to select news based on anticipated agreement as predicted by cognitive dissonance theory
- Effect stronger for hard news

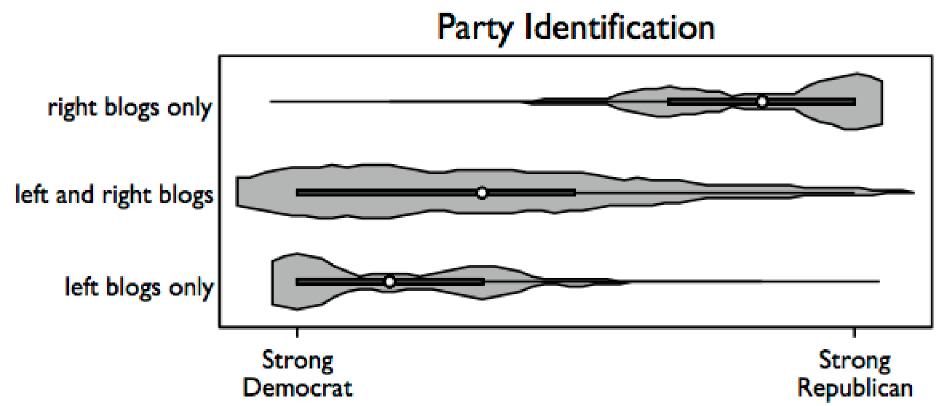


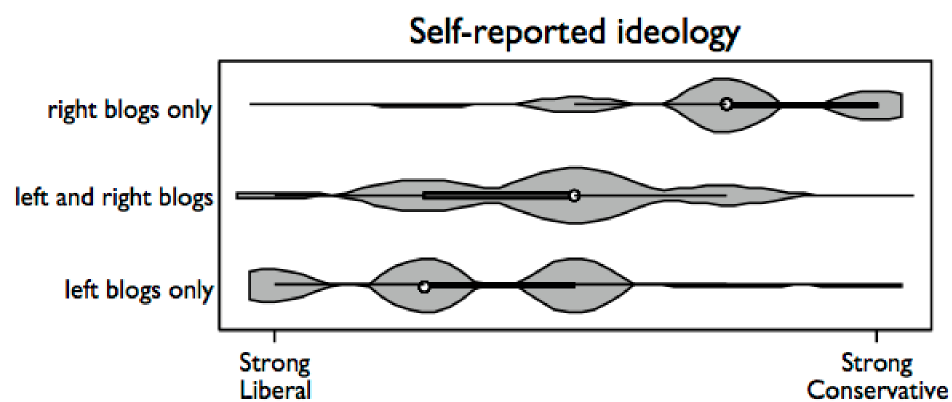
Echo Chambers in Blog Readership

Lawrence, E., Sides, J., & Farrell, H. “Self-segregation or deliberation? Blog readership, participation, and polarization in American politics.” (2010)

- Data from large survey (N=36,000)
- Blog readers are attracted to blogs aligned with their political views (94%)
- Polarization both by party identification and self-reported ideology

- Finding consistent with selective exposure

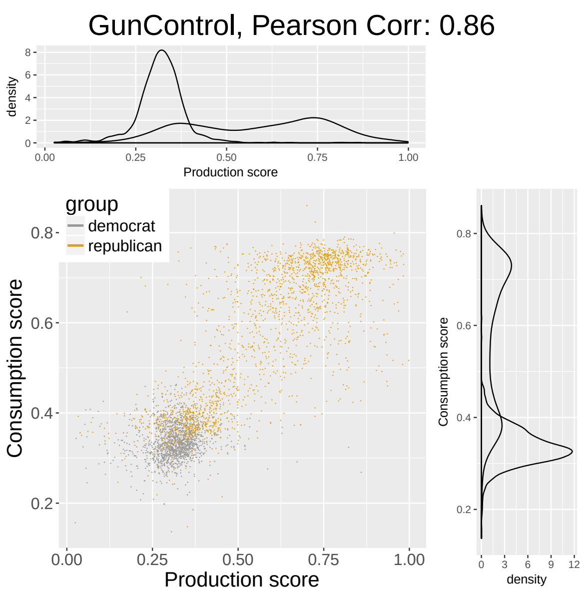


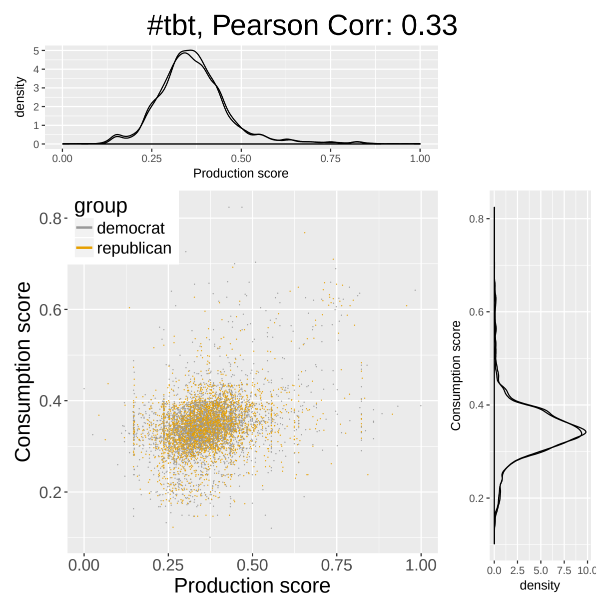


Echo Chambers on Twitter

Garimella et.al., "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship." WWW2018.

- Fixed set of politically active users
- Set of tweets that mention #topic
- Production vs consumption score
- Main finding: correlation of production and consumption scores
- Finding consistent with selective exposure

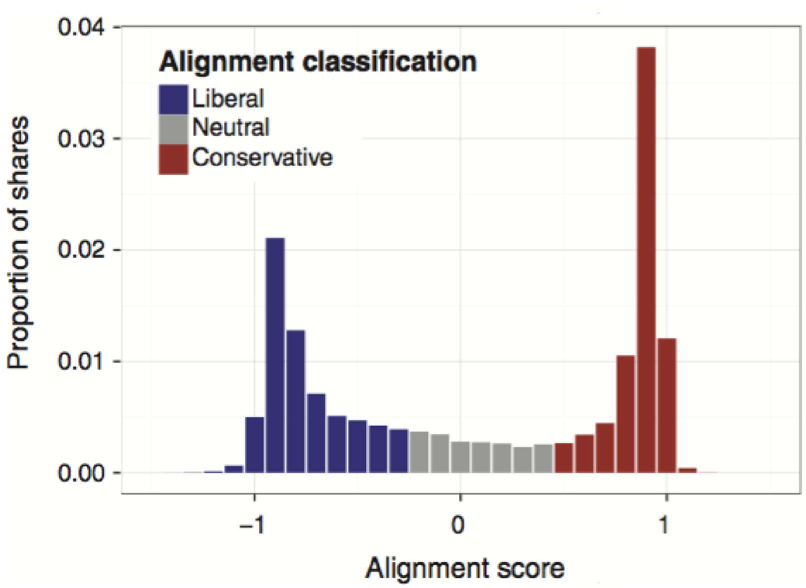


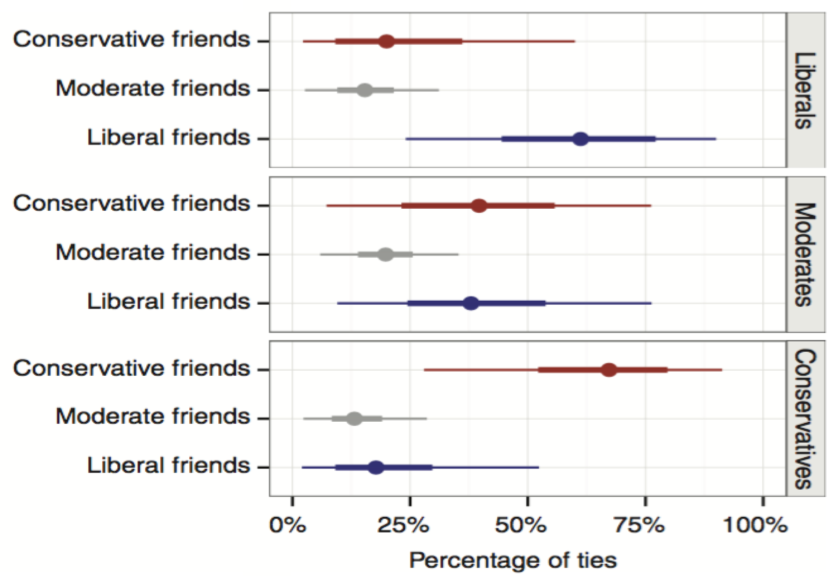


Partisan Exposure on Facebook

Bakshy, E., Messing, S., & Adamic, L. A. "Exposure to ideologically diverse news and opinion on Facebook." (2015)

- US Facebook users with self-reported ideological affiliation
- Analysis on hard news (national news, politics, world affairs)
- Each news associated with a political alignment
  - Average of the affiliation of users who shared the story
- Cross-cutting news if the alignment of the news and the user differ

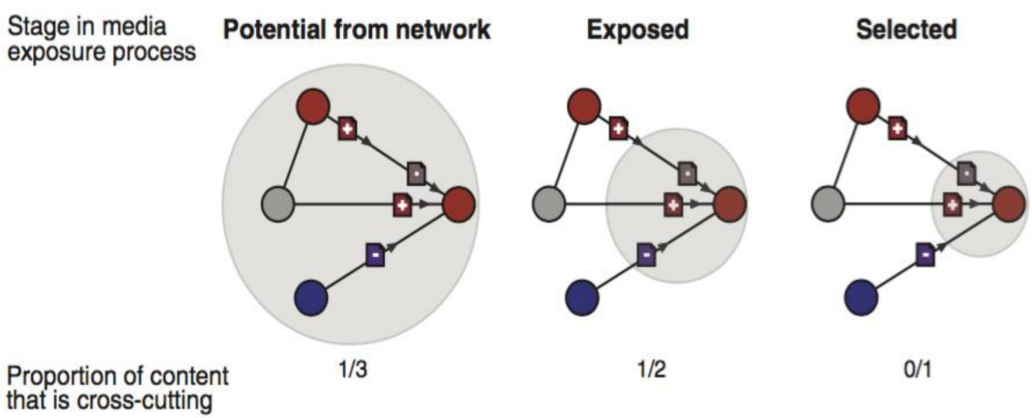


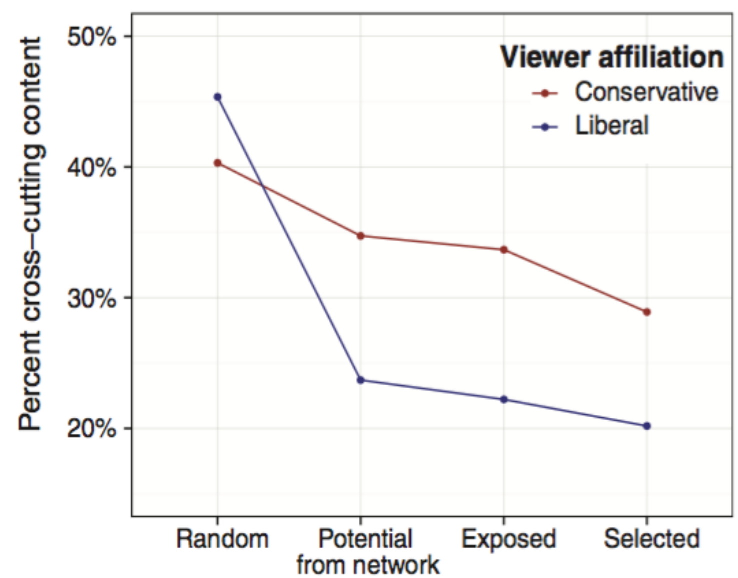


Partisan Exposure

Bakshy, E., Messing, S., & Adamic, L. A. "Exposure to ideologically diverse news and opinion on Facebook." (2015)

- Measure the fraction of cross-cutting news among:
  - ones posted in a user’s network (potential)
  - ones shown in the user’s timeline (exposed)
  - one the user clicked on (selected)
- Compared to random from the whole set, each step reduces the exposure and creates a narrower echo chamber
- Largest reduction from network (social), rather than algorithmic (filtering), selective exposure still plays a role





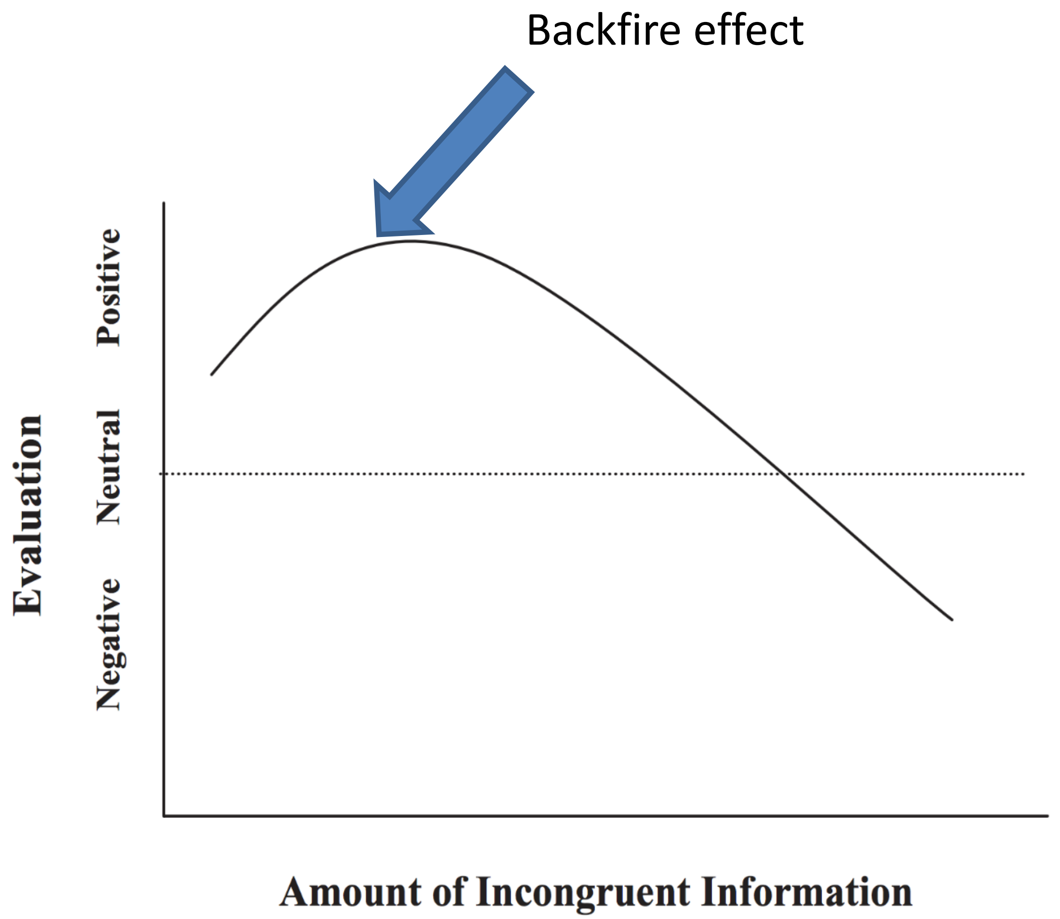
Why the Web might <u>not</u> increase polarization

- Homophily is not observed only for one type of issues (political)
  - The tendency of individuals to associate and bond with similar others
  - Could be based on various facets
    - Gender, age, race, status, religion, geography, beliefs

- Reality kicks in
  - Evidence accumulates at some point

Is there a tipping point?

Redlawsk D, The Affective Tipping Point: Do Motivated Reasoners Ever "Get It"? (2010)



Backfire effect

- Recent study (Bail et al, 2018)
- Surveyed a large sample (N=1652) of politically active twitter users, Democrats and Republicans
- Paid them to follow a Twitter bot for one month that exposed them to content of opposing political ideologies.
- Resurveyed after 1 month
- Finding:
  - Republicans who followed a liberal Twitter bot became substantially more conservative post-treatment
  - Democrats who followed a conservative Twitter bot became slightly more liberal post-treatment

Bail, et al "Exposure to Opposing Views can Increase Political Polarization: Evidence from a Large-Scale Field Experiment on Social Media." (2018)

MEASURING POLARIZATION

Polarization in content

- Sentiment variance in news
- Controversial topic - a concept that invokes conflicting sentiments
- Subtopic - factor that gives a particular sentiment (+ve or -ve)
- Assumption - a controversial topic receives contrasting sentiment (of different kind)
  - positive vs. negative feelings, pros vs. cons, rightness vs. wrongness in their judgments
- Similar results observed by
  - Garimella et al. WSDM 2016
  - Klenner et al. KONVENS 2014

Choi, Jung , and Myaeng. "Identifying controversial issues and their sub-topics in news articles." PAW-ISI 2010.
Garimella, De Francisci Morales, Gionis, and Mathioudakis. "Quantifying Controversy in Social Media." WSDM 2016.
Klenner, Amsler, Hollenstein, and Faaß. "Verb Polarity Frames: a New Resource and its Application in Target-specific Polarity Classification." KONVENS 2014.

Sentiment variance

- Method:
  - Identify candidate entities (noun phrases)
  - Compute sentiment in sentences involving these entities
  - Controversial if positive_sentiment + negative_sentiment > δ and |positive - negative| > Υ

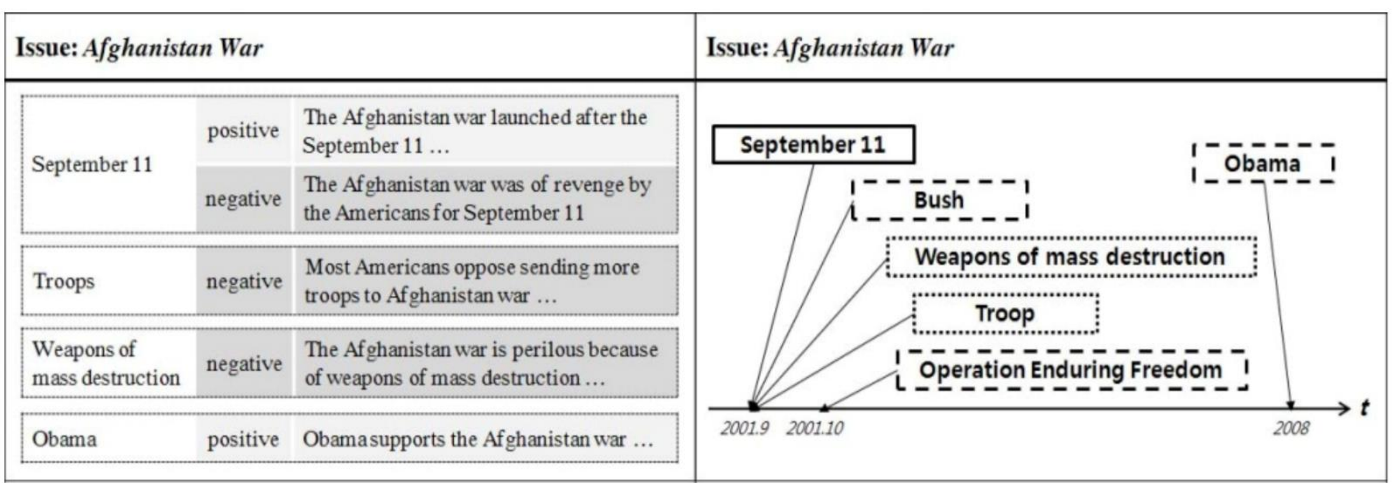
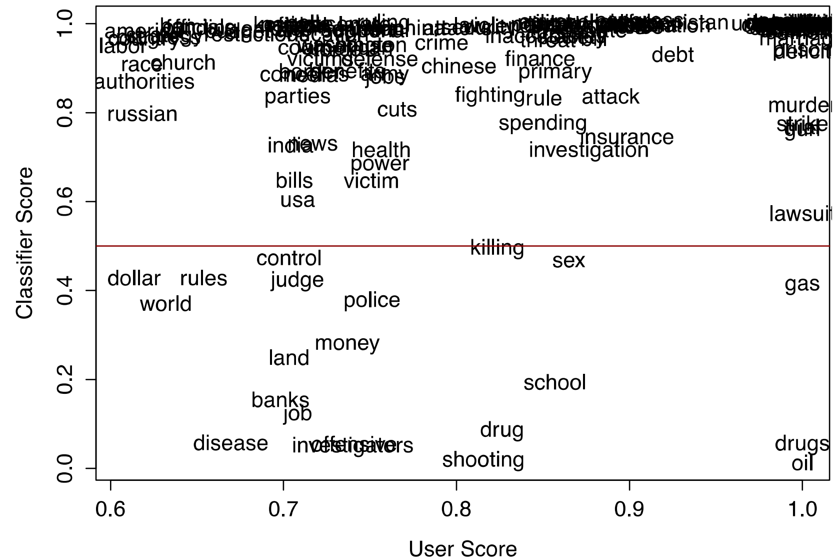


Fig. 1. A summary of the sentiment-generating subtopics for an issue "*Afghanistan War*"

Controversy language in news

- Controversy lexicon
- Controversial topics have:
  - strongly biased terms
  - more negative terms
  - fewer strongly emotional terms
- “we show that we can indicate to what extent an issue is controversial, by comparing it with other issues in terms of how they are portrayed across different media.”



(b) Controversial words; correctly classified words appear above the horizontal line.

Figure 2: Scores of controversial and non-controversial words including classification errors. “User score” is the confidence with which the manual labeling was done (with at least 7 annotators per element), while “classifier score” is the output of the classifier on the training data.

Mejova, Zhang, Diakopoulos, and Castillo. “Controversy and Sentiment in Online News.” C+J Symposium 2014.

Detecting controversy on the Web

- Find out if a Web page discusses a (known) controversial topic
- Map topics (named entities) in a Web page to Wikipedia articles
  - A Web page is controversial if it is similar to a controversial Wikipedia article
  - E.g., If a news article mentions Abortion it is controversial
- Related:
  - There is a lot of work on identifying controversial topics on Wikipedia
  - Edit wars, hyperlink structure, etc.
- Related:
  - Jang et al. show that in addition to this, language models can be built to directly detect controversy

Dori-Hacohen and Allan. "Detecting Controversy on the Web." CIKM 2013.
Jang, Foley, and Allan. "Probabilistic Approaches to Controversy Detection." CIKM 2016.

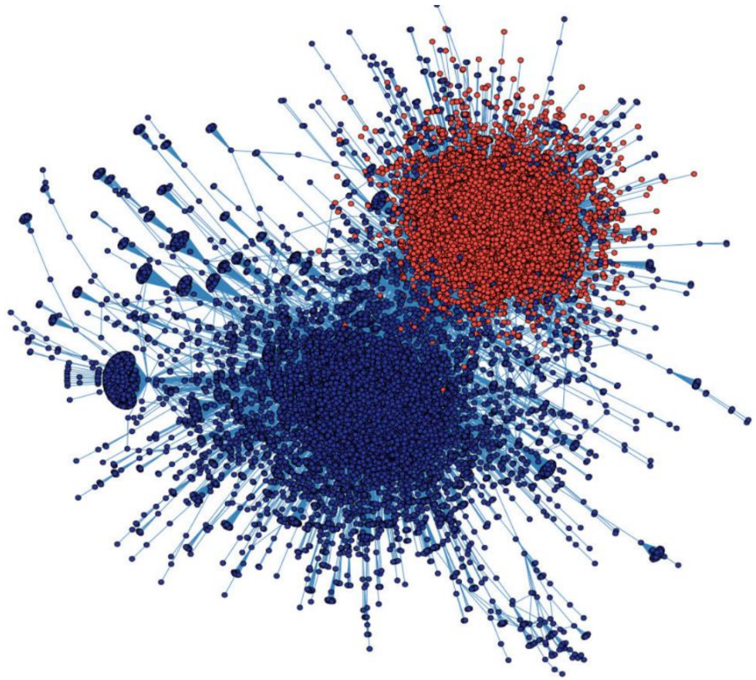
Identifying polarization - Network

- Methods based on network structure
  - Social media, hyperlinks
  - Twitter: Retweet,ReplymSocial (follow)

- Idea: Controversial topics have a clustered structure in their discussions

Quantifying polarization via Modularity

- Modularity:
  - the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random
  - Compares the number of edges inside a cluster with the expected on a random graph
  - Captures the strength of division of a network into modules

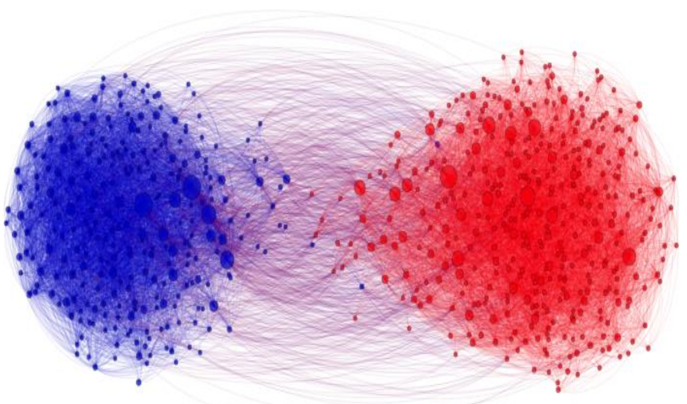


Modularity: 0.48

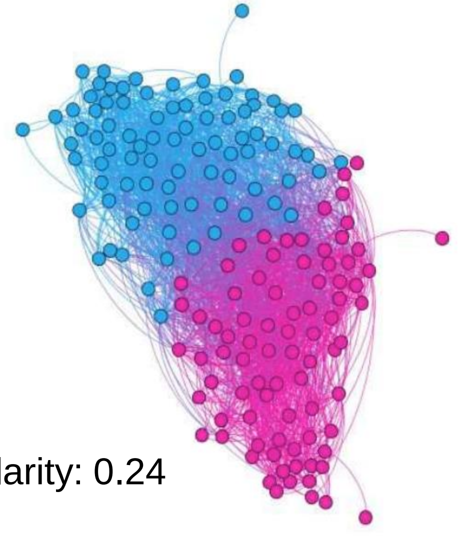
Conover, Ratkiewicz, Francisco, Gonçalves, Menczer, and Flammini. “Political Polarization on Twitter.” ICWSM 2011.

Modularity is not a direct measure of polarization

- We want to capture the in group vs out group interaction preference
- Sensitive to the size of the graph and partitions
- Not “monotone”
  - Strengthening of internal ties can decrease modularity
- How much modularity indicates polarization?



Modularity: 0.42

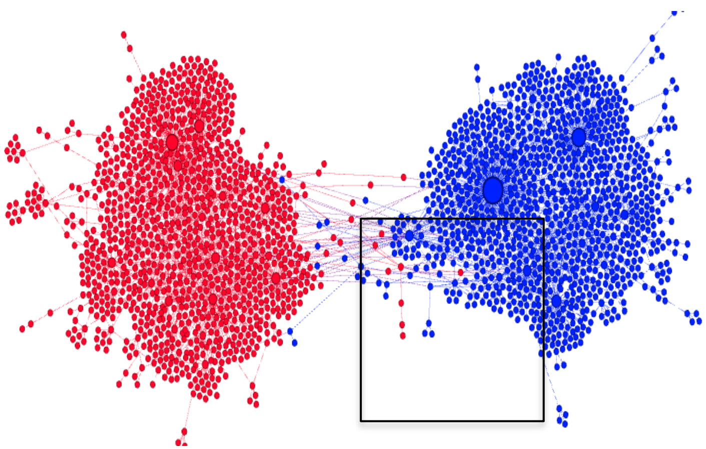


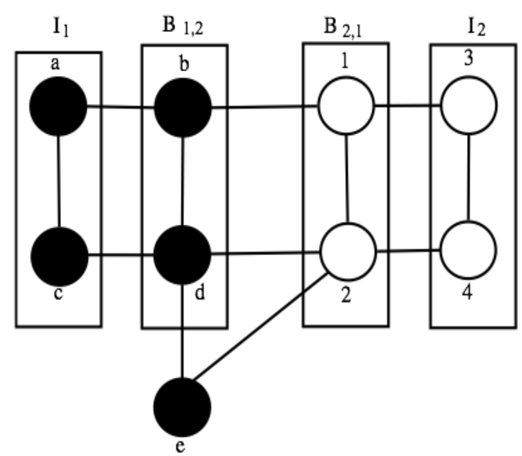
Modularity: 0.24

Community boundary

Guerra, Meira, Cardie, and Kleinberg. "A Measure of Polarization on Social Media Networks Based on Community Boundaries." ICWSM 2013.

- Boundary node:
  - have at least one edge that connecting to the other community
  - have at least one edge connecting to a member of its community which does not link to the other community
- $P(v) = d_{internal}(v)/(d_{external}(v) + d_{internal}(v)) - 0.5$
- $P(v) > 0 \rightarrow$ v prefers internal connections (antagonism?)
- $P(v) < 0 \rightarrow$ v prefers connections with members of the other group

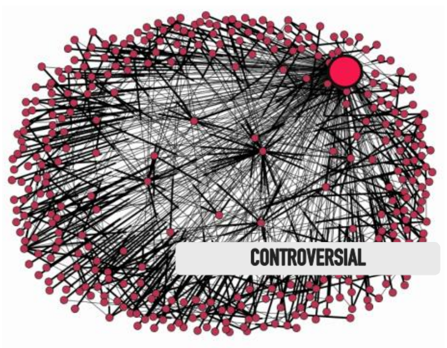


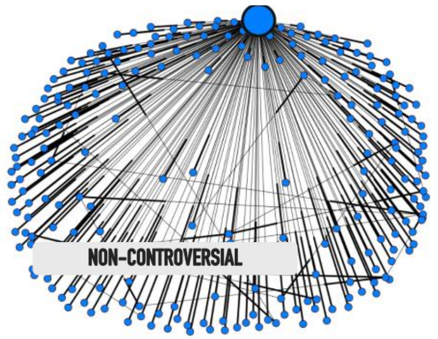


Motif-based approach

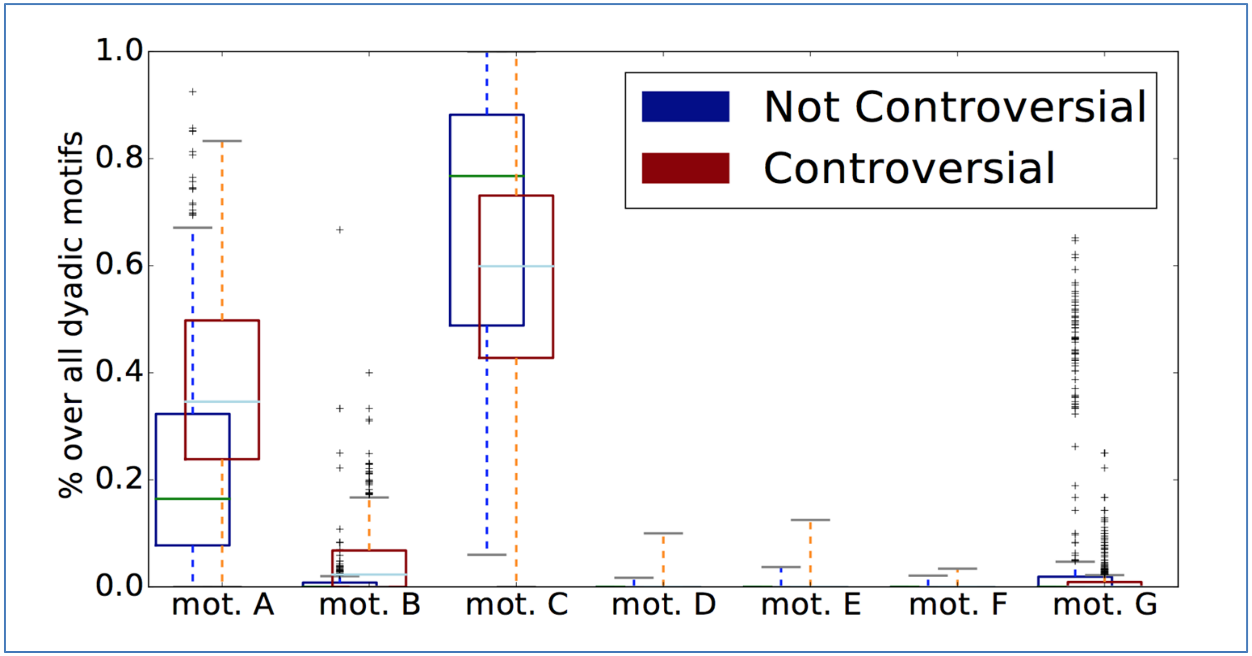
Coletto, Garimella, Luchesse, and Gionis. "A Motif-based Approach for Identifying Controversy." OSNEM. 2017.

- Define reply trees
- Identify frequency of motifs
- Take into account also socia
  - follower information

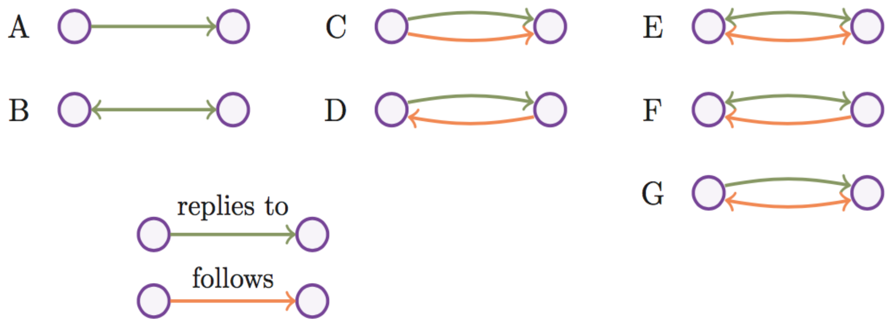




Motifs



1.0
0.8
0.6
0.4
0.2
0.0
% over all dyadic motifs
Not Controversial
Controversial
mot. A
mot. B
mot. C
mot. D
mot. E
mot. F
mot. G



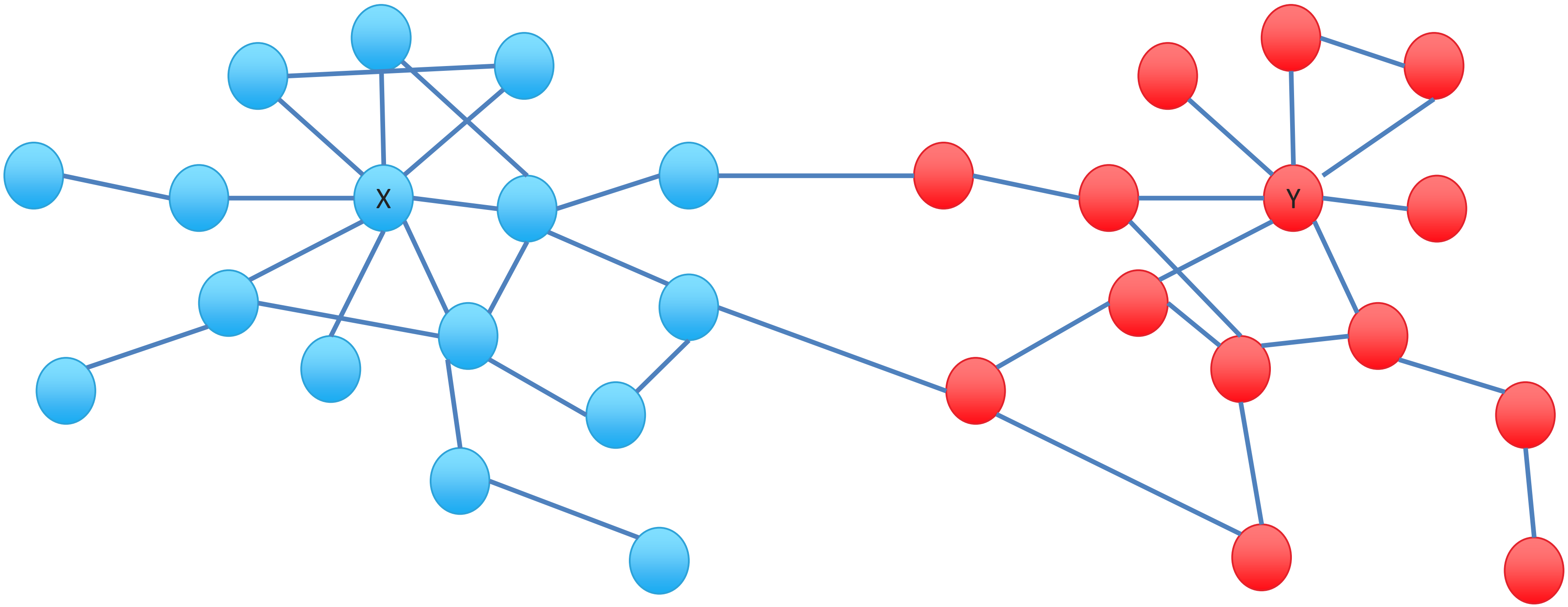
A
B
C
D
E
F
G
replies to
follows

Based on information flow

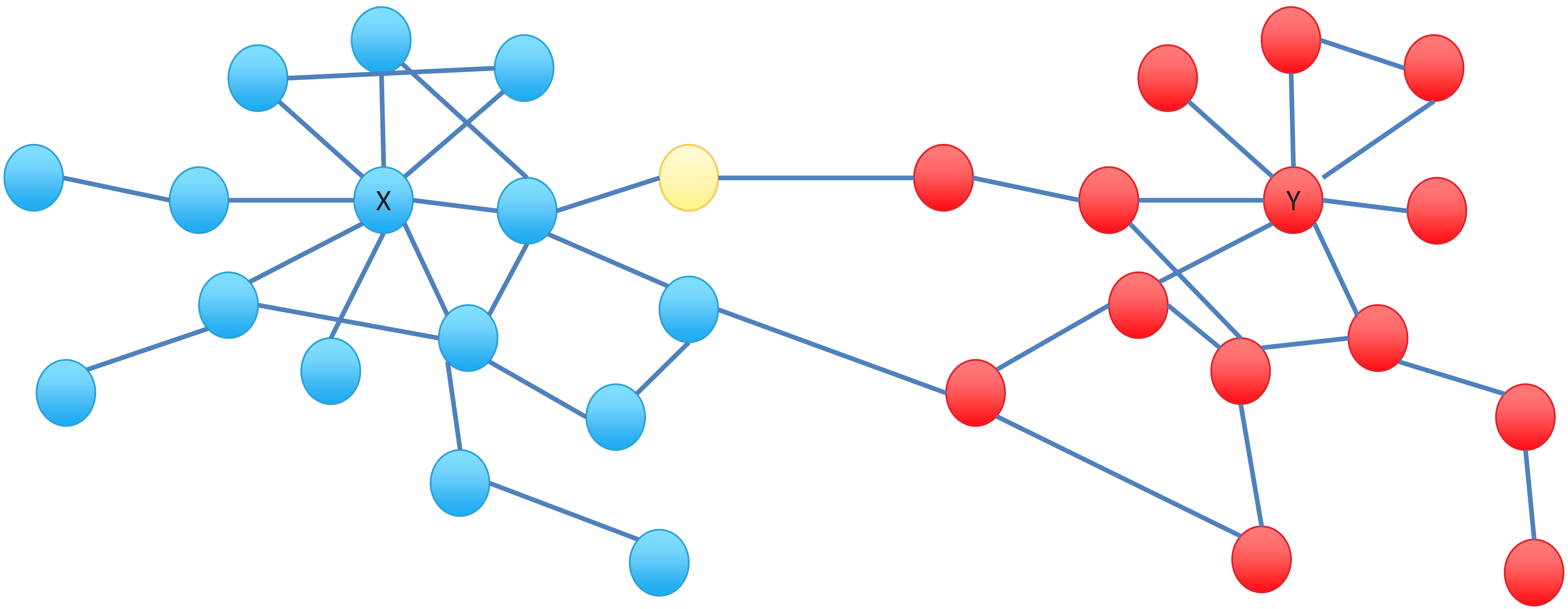
Garimella, De Francisci Morales, Gionis, and Mathioudakis. “Quantifying Controversy in Social Media.” WSDM 2016.

- Random walk controversy measure (RWC)
  - Authoritative users exist on both sides of the controversy
  - How likely a random user on either side is to be exposed to authoritative content from the opposing side
- Works on both the retweet graph and the social graph
- Requires a partition of the graph

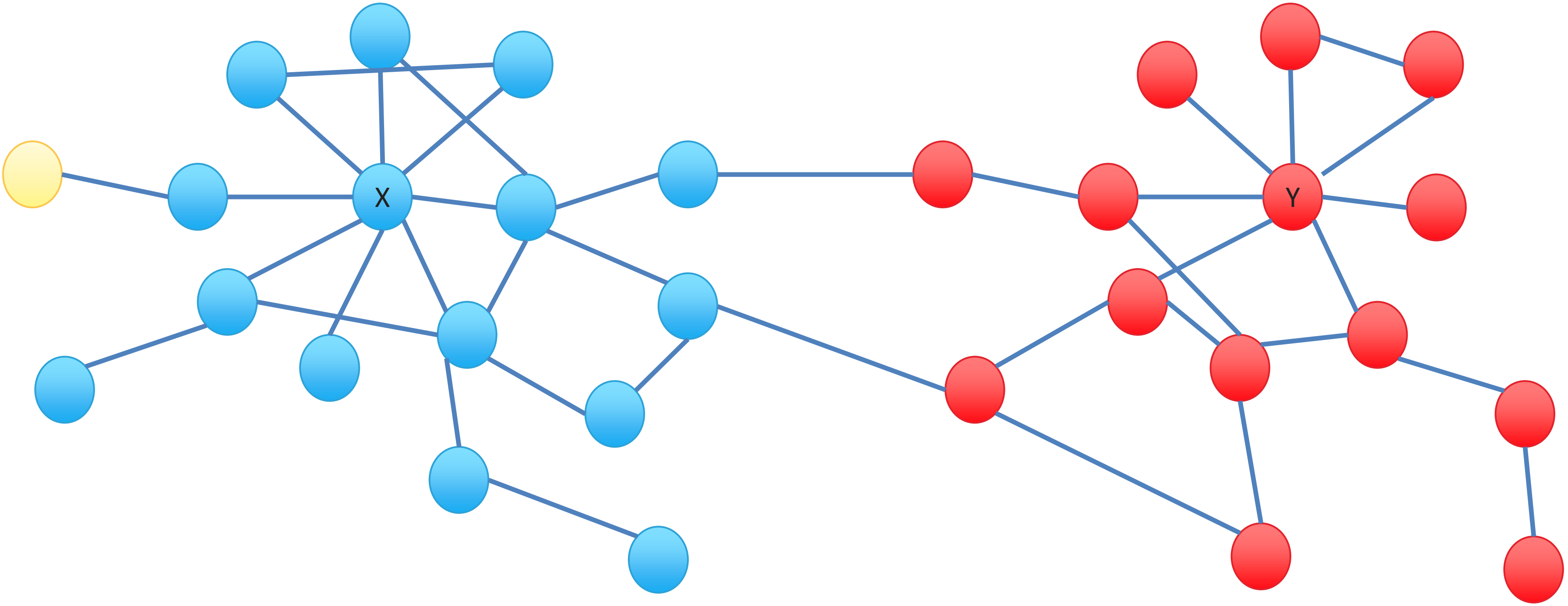
Random walk controversy score



Random walk controversy score

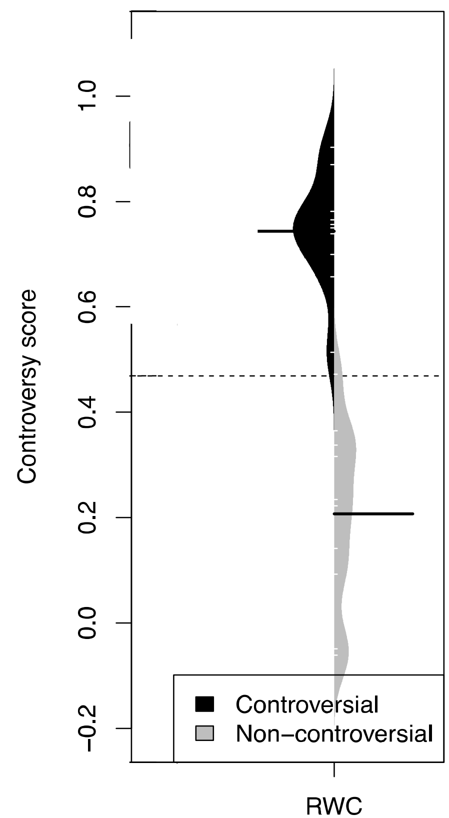


Random walk controversy score



Random walk controversy score (RWC)

$$\boxed{RWC = P_{XX}P_{YY} - P_{YX}P_{XY}}$$



Polarization via Opinion Formation model

- We assume internal opinions in the interval [-1,+1].
  - E.g., Democrats and Republicans.
- Run the FJ model and compute the expressed opinions $\boldsymbol{z}$
- $|z_u|$ measures the degree of polarization of u
  - Expressed opinion values z_u close to 0 signify lack of polarization (neutrality).
  - Expressed opinion values z_u close to -1 or 1 signify polarization.
- For the whole network, we define the polarization index as

$$\pi(\boldsymbol{z}) = \frac{||z||^2}{n}$$

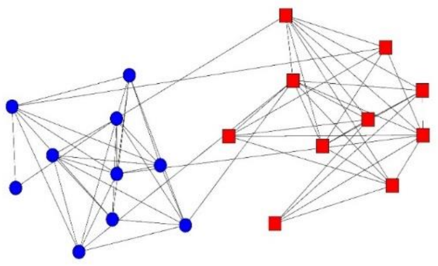
Distance from state of neutrality

A. Matakos et al. *"Measuring and moderating opinion polarization in social networks"*. DMKD 2017

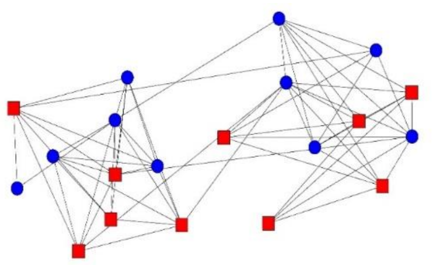
Polarization index interpretation

- Random walk interpretation: the z_u value is the expected intrinsic opinion at the endpoint of a random walk in the graph that starts from node u.
  - Low value of $|z_u|$ implies that u has equal probability of reaching positive and negative viewpoints: network of u is moderate and diverse
  - High value of $|z_u|$ implies that user is surrounded by single-minded users with extreme views.

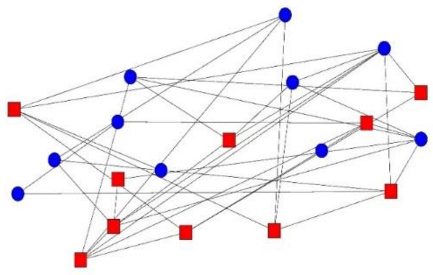
Examples



Echo chamber graph
Polarization Index: 0.30



Community structure with
random opinions
Polarization Index: 0.03



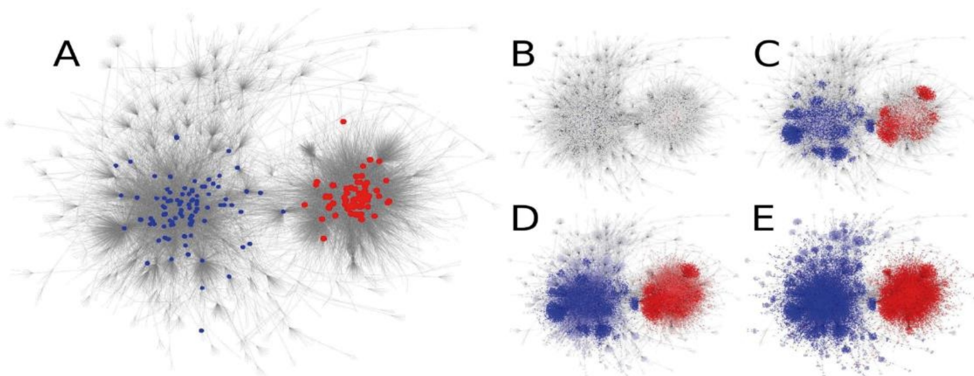
Random graph
Polarization Index: 0.03

- The polarization index captures echo-chambers in the network.

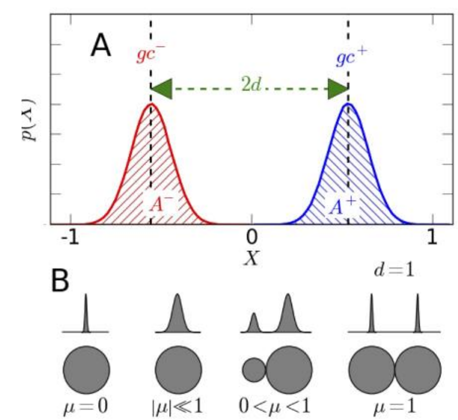
Label propagation

Morales, Borondo, Losada, and Benito. “Measuring political polarization: Twitter shows the two sides of Venezuela.” Chaos. 2015.

- Opinion formation:
  - Identify a set of ‘seed’ users and propagate until convergence



- Measure: distance between distributions
  - “Dipole moment”
  - Accounts for the mass of the population



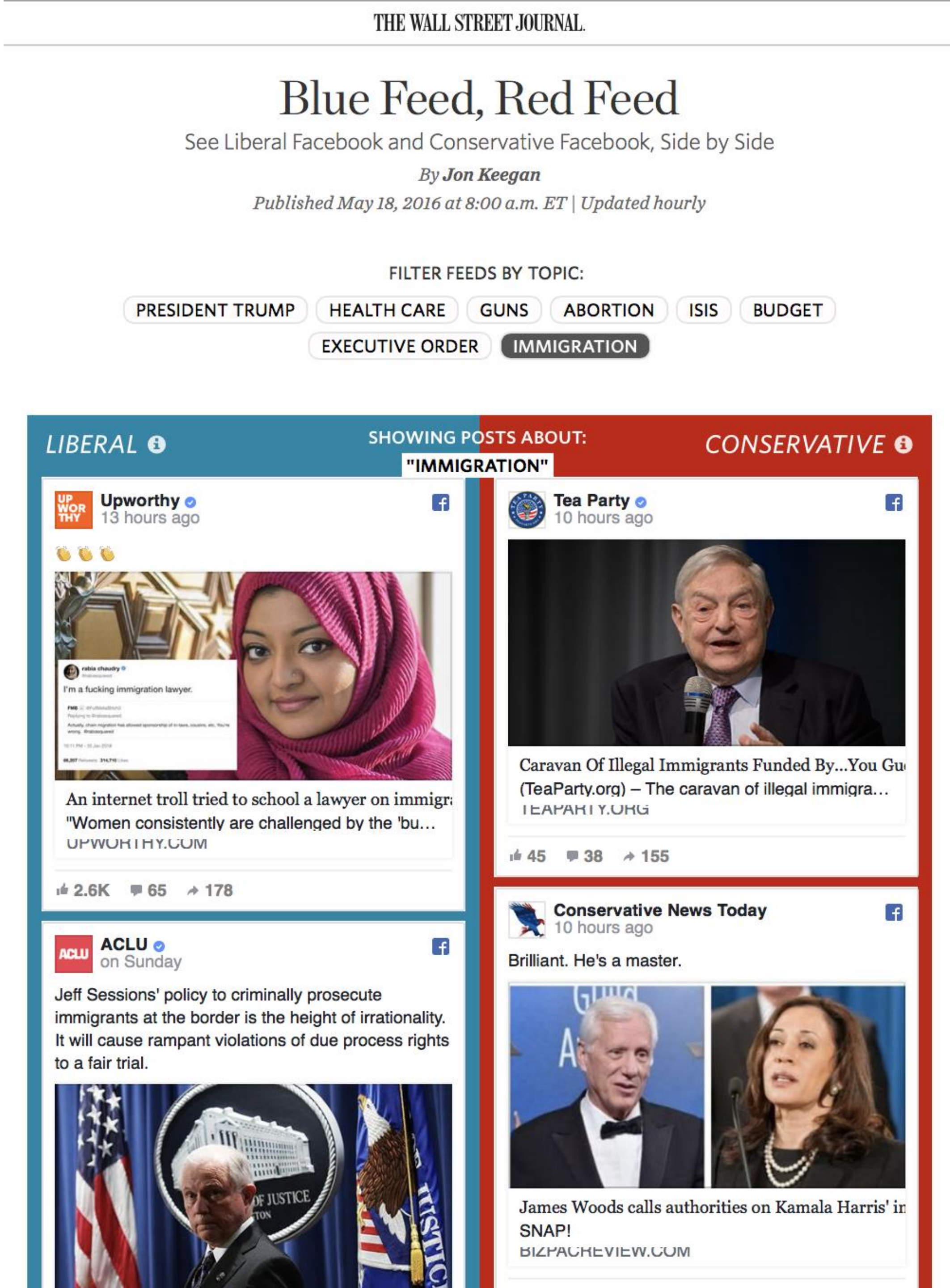
MITIGATING POLARIZATION

Wall Street Journal

Blue Feed, Red Feed

Curated by the newspaper

Aims to show how different the facebook feed can be for different users



Support The Guardian | Subscribe | Find a job | Sign in | Search | International edition

The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

World ▶ Europe | US | Americas | Asia | Australia | Middle East | Africa | Inequality | Cities | Global development

Burst your bubble

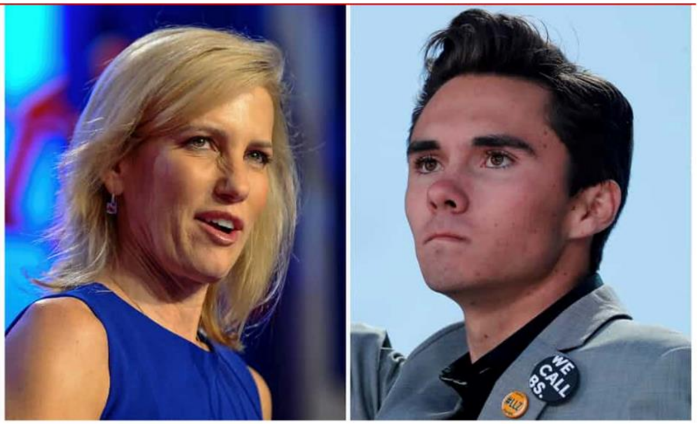
The Guardian's weekly guide to conservative articles worth reading to expand your thinking

2 April 2018

Laura Ingraham is a victim of a totalitarian campaign from the left, apparently

The American right have revealed a vision of free speech that is very expansive for conservatives but far less accommodating for those who disagree with them

7:41 PM | 694



Burst your bubble by the guardian

The Guardian is left-wing

The column shows selected conservative articles from around the web

Escape your bubble

Browser (chrome) extension

Asks you which type of people you would like to be more accepting to

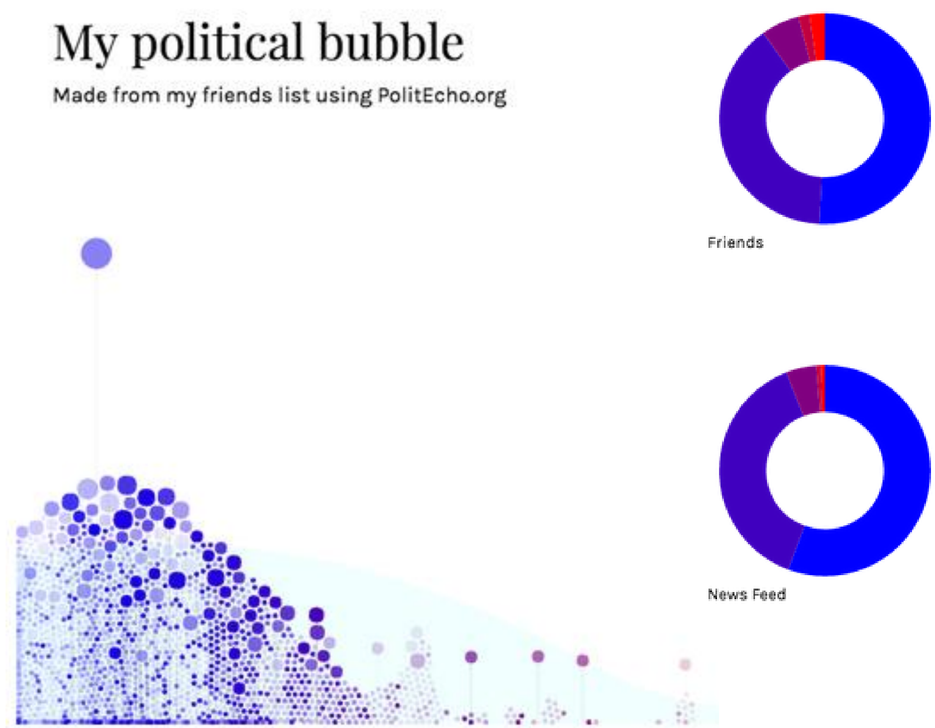
App inserts human-curated, positive articles and images into Facebook News Feed, which paint those you would like to be more accepting of in a positive light

politecho.org

Browser (chrome) extension

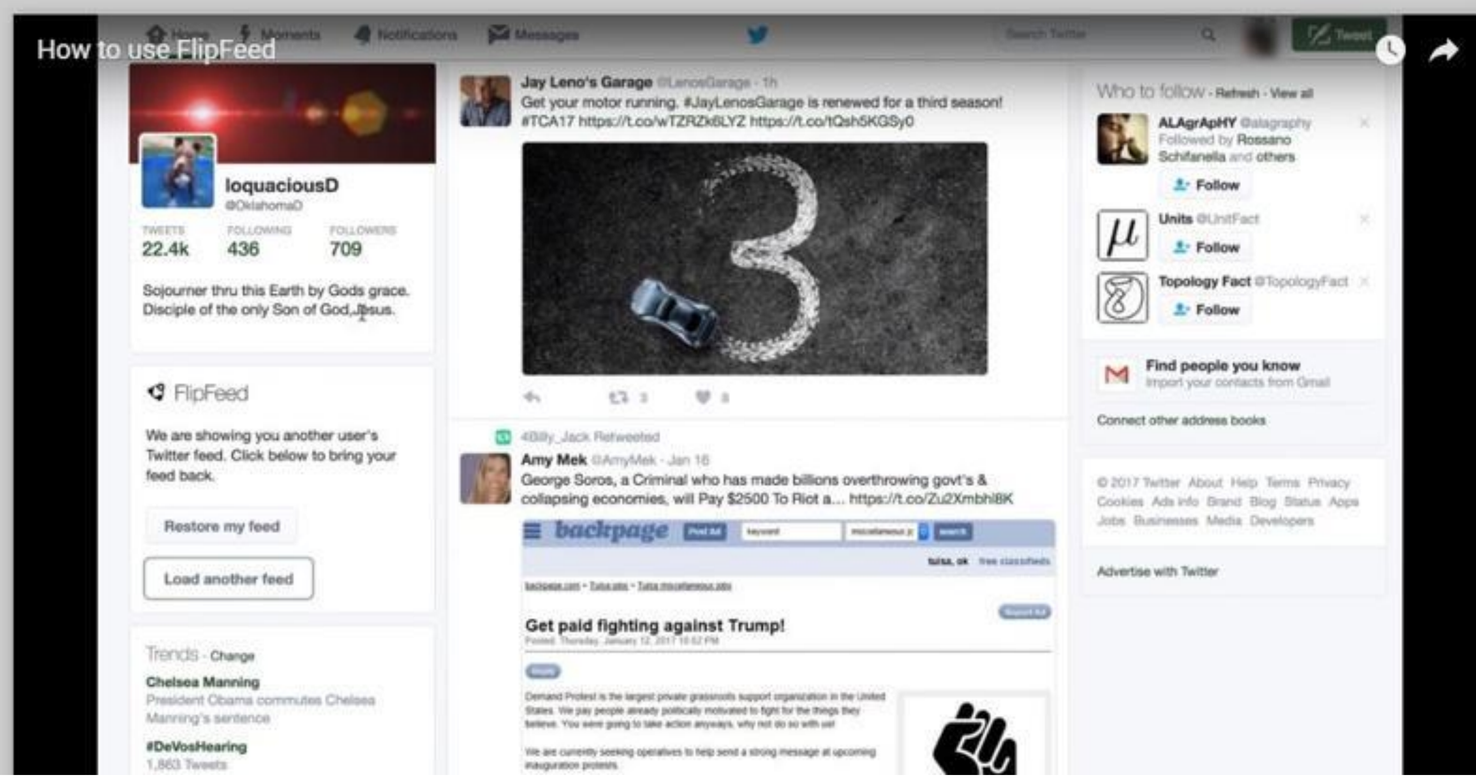
Shows political distribution of own Facebook feed vs. that of friends

Compares liked political pages with a reference set of political pages



FLIPFEED

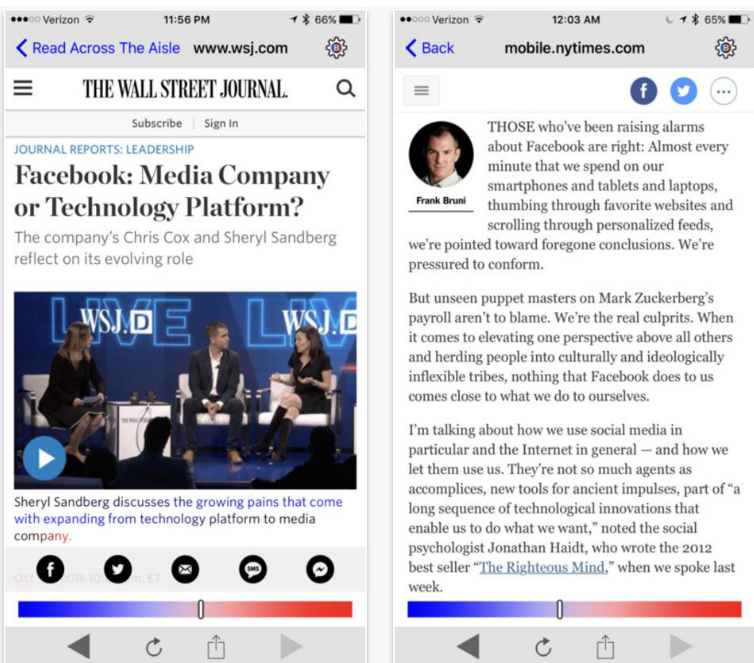
Step into someone else's Twitter feed



Browser (chrome) extension

Allows Twitter users to see a feed that resembles that of another user who has been pre-classified as right- or left-leaning

Laboratory for Social Machines at MIT Media Lab



Mobile (iPhone) app and chrome extension

News reader for select sources

Keeps track of personal reading history

Informs user of news diet bias

Algorithmic mediation/recommendations

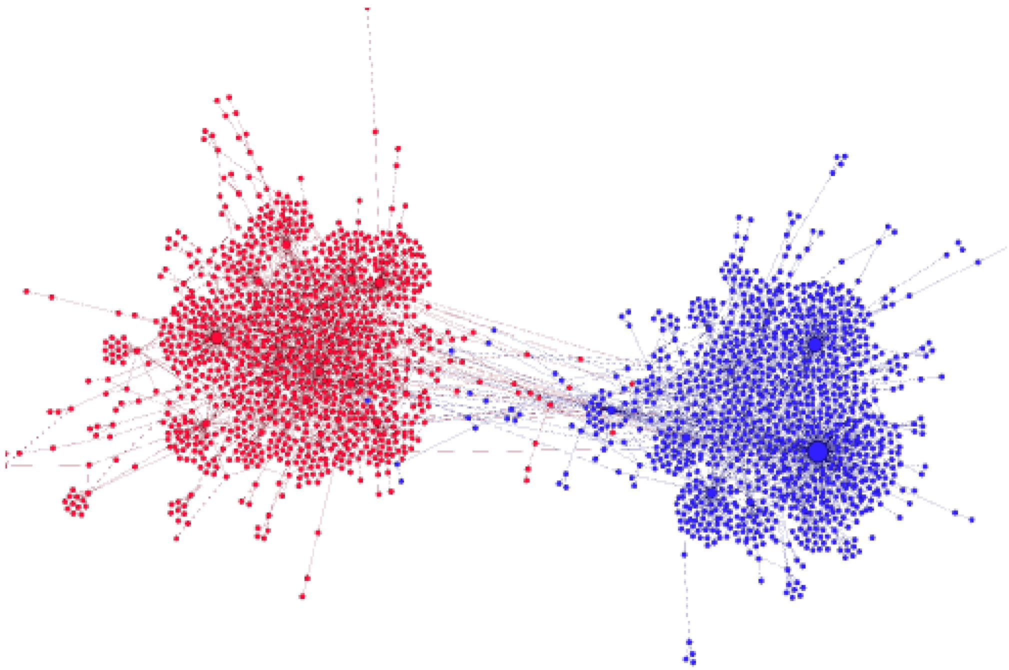
- Task : make a recommendation helping to reduce polarization
- Different approaches driven by polarization metrics
  - Pick a favorite metric : RWC, opinion diversity, influence-based
  - Compute recommendation that reduces polarization according

    to the selected metric
  - Account for recommendation acceptance probability

- Another dimension: What to recommend? User vs. content

1. Recommendations based on RWC

- *Recall* : random-walk controversy score
- Quantifies the degree of polarization of a given topic
- Based on the structure of the retweet graph of the topic



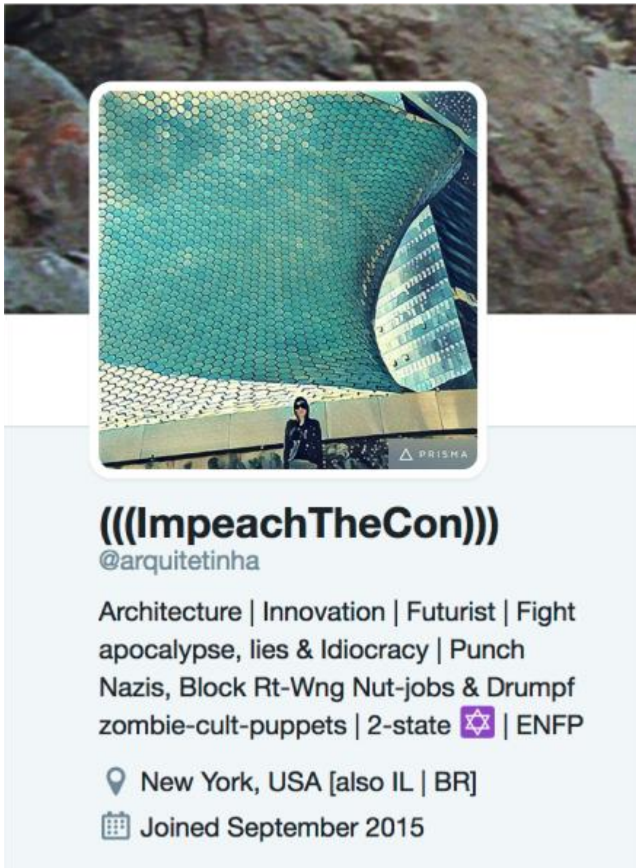
1. Recommendations based on RWC

- Assuming : polarization is measured by RWC
- Problem : add k edges to maximally reduce RWC
- Enhance greedy with efficient incremental computation
- Edge additions are interpreted as recommendations
- Incorporate probability of accepting a recommendation
  - compute user polarity, and
  - acceptance probability as a function of user polarity

K. Garimella et al. *"Reducing controversy by connecting opposing views"*. WSDM 2017

Reducing polarization : real example

Polarity = −0.99



Polarity = 0.95

Reducing polarization : real example

Polarity = −0.99

Polarity = 0.15

Reducing polarization : results

		obamacare		guncontrol	
		node1	node2	node1	node2
Ignoring acceptance probabilities	ROV	mittromney realdonaldtrump barackobama barackobama michelebachmann	barackobama truthteam2012 drudge_report paulryanvp barackobama	ghostpanther mmflint miafarrow realalexjones goldiehawn	barackobama robdelaney chuckwoolery barackobama jedediahbila
With acceptance probabilities	ROV-AP	kksheld lolgop irritatedwoman hcan klsouth	ezraklein romneyresponse motherjones romneyresponse dennisdmz	chuckwoolery liamkfisher csgv jonlovett drmartyfox	csgv miafarrow dloesch spreadbutter huffpostpol

K. Garimella et al. *"Reducing controversy by connecting opposing views"*. WSDM 2017

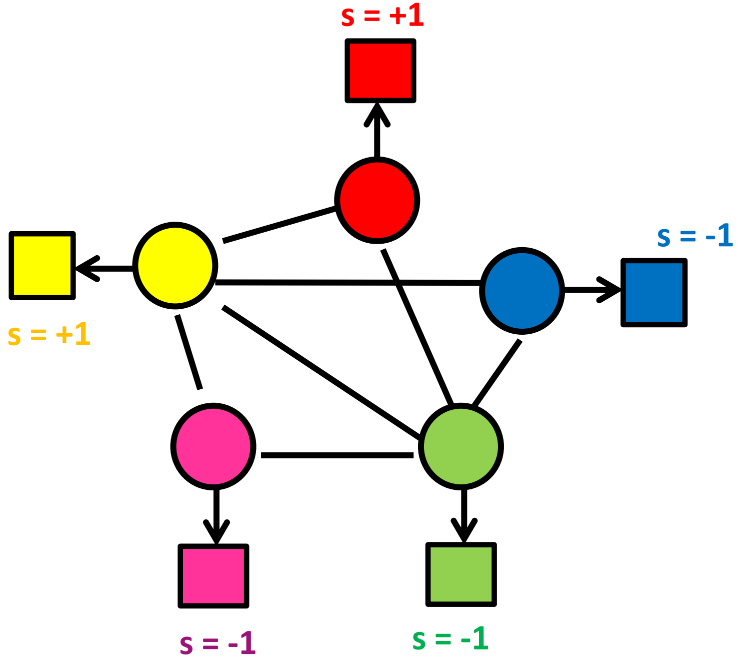
Reducing the polarization index

- Reduce polarization by convincing users to adopt a neutral opinion
  - We assume a budget (k) of such interventions.
  - Find the k interventions that minimize the polarization index π

– *Moderate Internal*: Neutralize k internal opinions ($s_i = 0$)

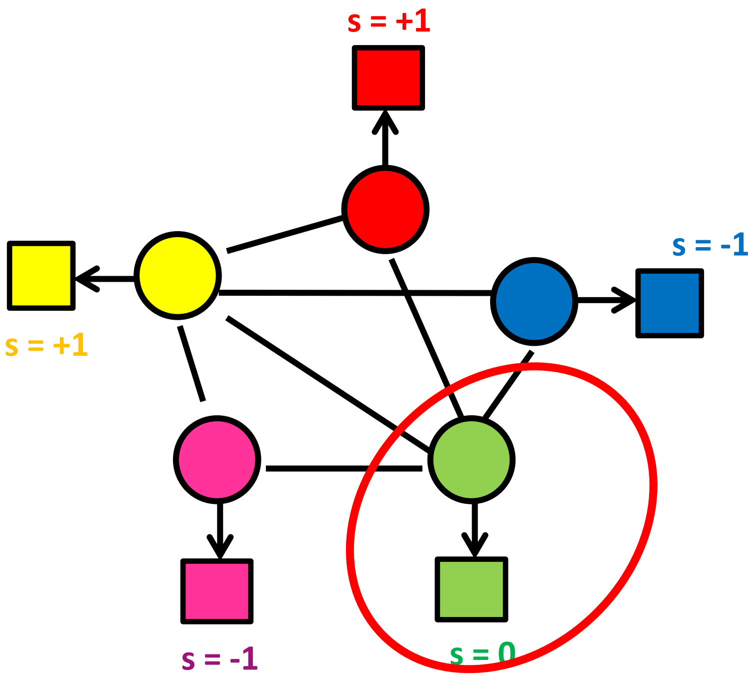
– *Moderate Expressed*: Neutralize k expressed opinions ($z_i = 0$)

Moderating opinions – Example

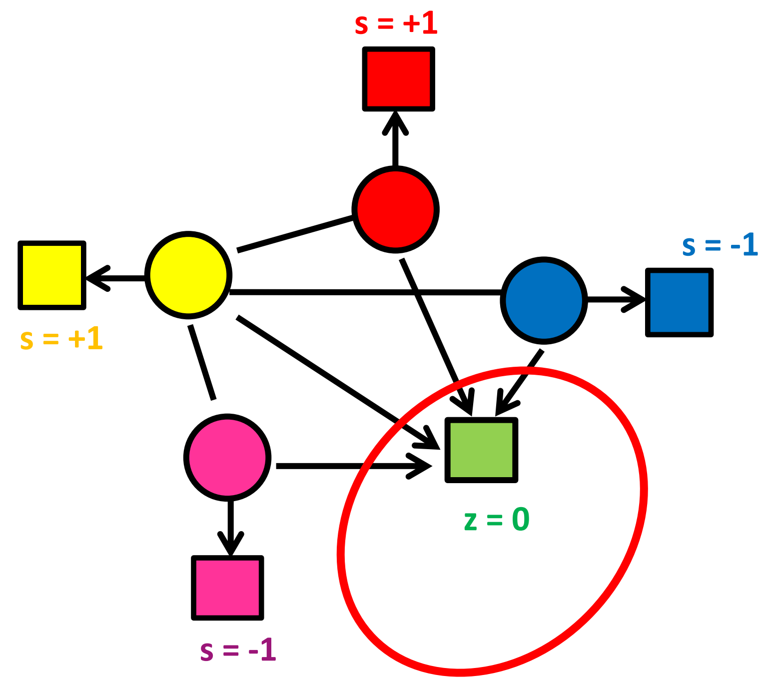


Moderating opinions - Example

Moderate Internal



Moderate Expressed



Algorithms

- Both problems are NP-hard

- Linear algebraic property of model: $\boldsymbol{z} = (L + I)^{-1}\boldsymbol{s}$
  - Use this property to design efficient algorithms

- The ModerateInternal problem has an interesting connection to the Sparse Approximation problem
  - Intuitively: we want a sparse selection $\boldsymbol{s}'$ of $\boldsymbol{s}$ such that $(\mathrm{L} + \mathrm{I})^{-1}\boldsymbol{s}' \approx \boldsymbol{z}$. Subtracting $\boldsymbol{s}'$ from $\boldsymbol{s}$ will minimize the metric
  - BOMP algorithm: a variation of orthogonal matching pursuit for sparse approximation

Greedy algorithms

- Iteratively select k nodes, each time neutralizing the node that causes the maximum decrease in the polarization index.
  - For the *Moderate External* problem, to estimate the decrease in polarization we need to recompute the $(L+I)^{-1}$ for each candidate – too expensive.
  - Efficient implementation using the Sherman-Morrison Formula

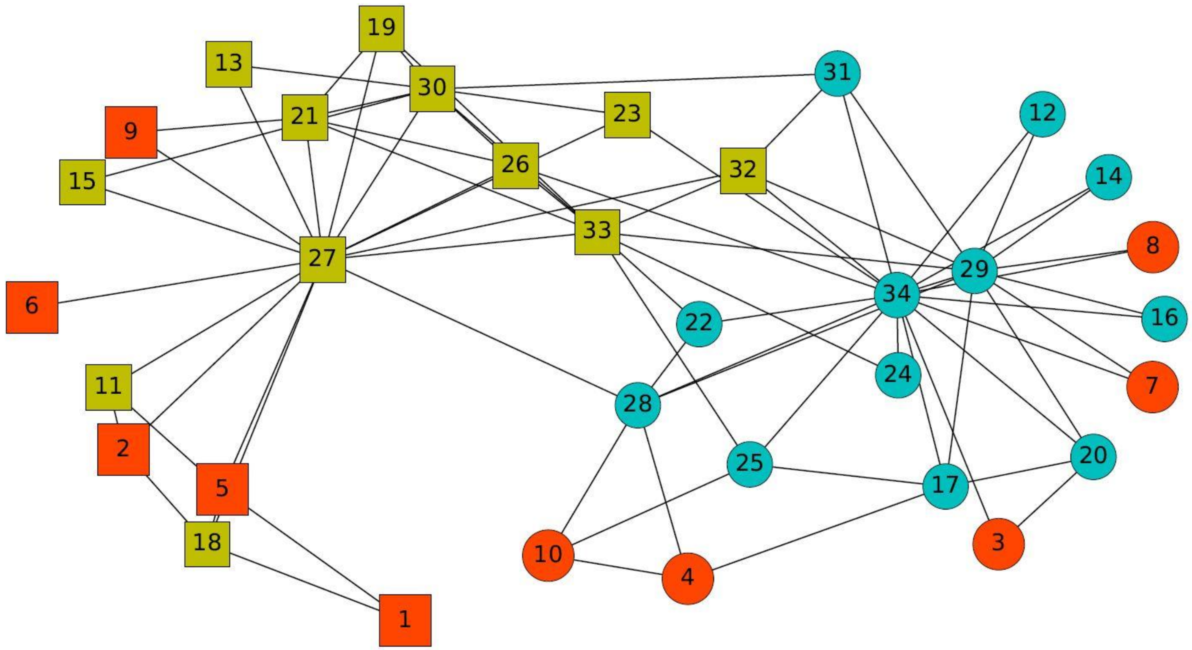
Heuristics for opinion moderation

- ExtremeExpressed: At each step neutralize the node with the highest expressed opinion.

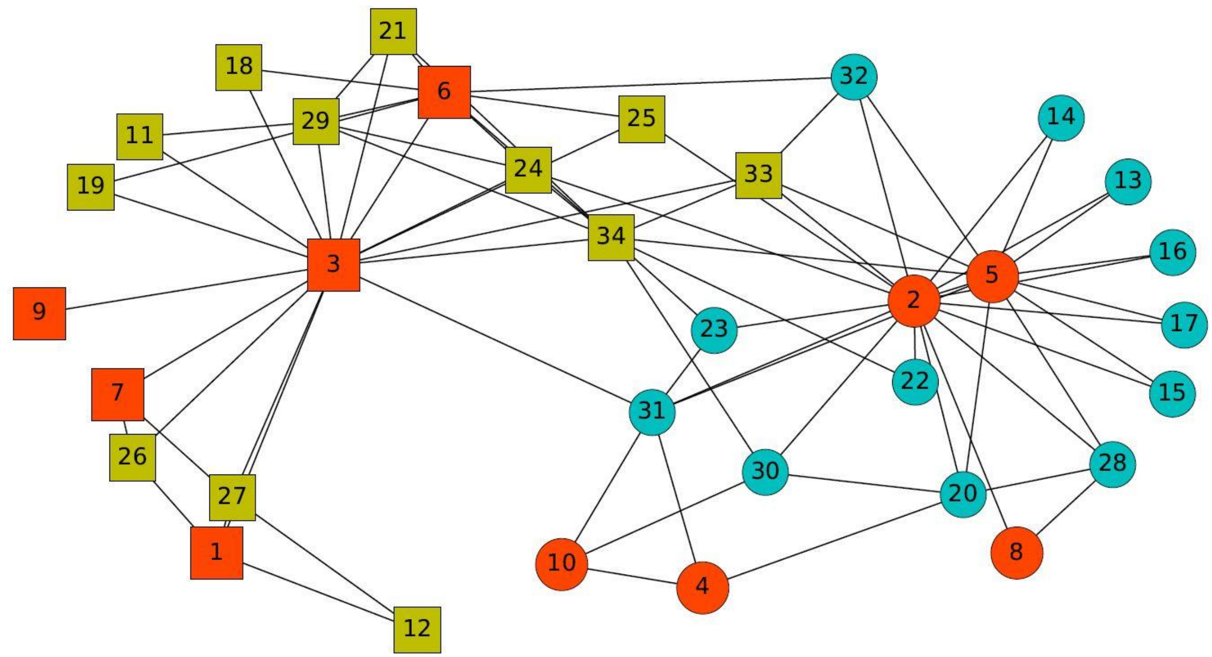
- ExtremeNeighbors: At each step neutralize the node whose neighbors have the highest absolute sum of expressed opinions

- Pagerank: Select the nodes in decreasing order according to their PageRank value

Selected nodes by GreedyInt



Selected nodes by GreedyExt



Recommendations based on information propagation models

- *Recall* the classic viral-marketing setting
  - Given a social network and a propagation model
    e.g., independent-cascade model
  - an action (e.g., meme) propagates in the network
- The influence-maximization problem
  - find k seed nodes to maximize spread
- The standard solution
  - spread is non-decreasing and submodular
  - greedy gives (1−1/e) approximation

Balancing information exposure

- Proposed setting
  - a social network and two campaigns
  - seed nodes I_1 and I_2 for the two campaigns
  - a model of information propagation
- The problem of balancing information exposure
  - find additional seeds S_1 and S_2, with $|S_1| + |S_2| \leq k$
  - s.t. minimize # of users who see only one campaign
  - or maximize # of users who see both or none

K. Garimella et al. *"Balancing information exposure in social networks"*. NIPS 2017

Algorithmic Fairness

Algorithmic Fairness in Machine Learning

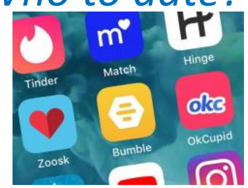
Algorithmic Fairness in Networks

Algorithmic Fairness: Why?

We live in a world where opinions are formed, and decisions are assisted or even taken by AI algorithms: often **black boxes;** driven by **enormous amount of data**

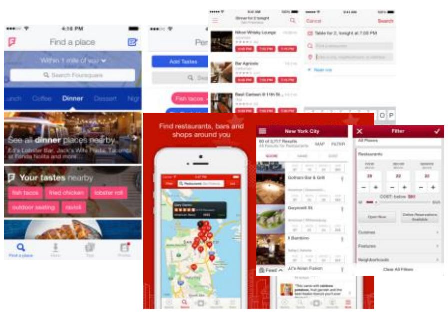
From simple, or not that simple, *personal* ones

Who to date?

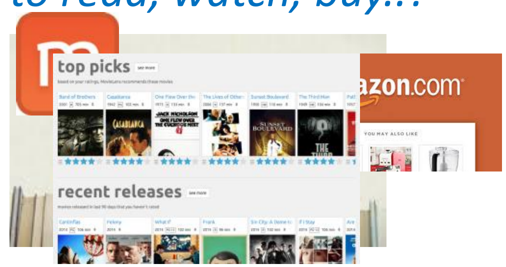


What is happening in the world?

Where to dine?



What to read, watch, buy..?



Get informed, learn?

Which job to take? Which school to attend? Whom to follow? Whom to vote…? ..?

Shape our opinions and our view of the world

Algorithmic Fairness: Why?

And not just by individuals:

- Medicine: prognosis, diagnosis, treatment recommendation
- Insurance, Credit, Benefit (resource) allocation, Housing
- Pricing of goods and services
- Education, e.g., school admission
- Law enforcement, e.g., sentencing decisions
- Job recruitment

Raise several concerns

And this concern has not been without reason:
a steady stream of empirical findings has shown that data-driven methods can unintentionally both encode **existing human biases** and **introduce new ones**.

A Snapshot of the Frontiers of Fairness in Machine Learning
By Alexandra Chouldechova, Aaron Roth
Communications of the ACM, May 2020, Vol. 63 No. 5, Pages 82-89

Case Studies

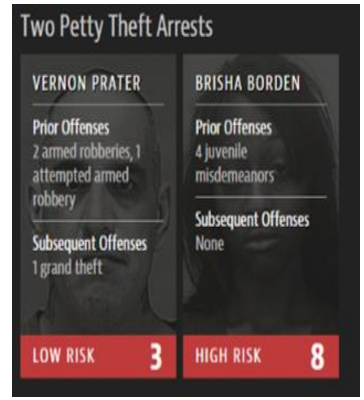
Algorithmic (Un)Fairness: Examples

Example: **The COMPAS recidivism prediction[1]**

Commercial tool that uses a *risk assessment algorithm* to predict some categories of future crime

Used in courts in the US for bail and sentencing decisions

Study by ProPublica



	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Algorithmic (Un)Fairness: Examples

Example: **Word Embeddings**[(2)]

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

representation of words/texts as a vector of numbers

- Banana →(0.3, 5.8, 7.3, 0.1)
- Father →(0.4, 0.7, 1.2, 0.4)
- Baby → (0.3, 0.6, 1.5, 3.0)

Why? Algorithms work with numbers.

Gender stereotype *she*-*he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Trained on a corpus of Google News texts

The trained embedding exhibit female/male gender stereotypes, learning that "doctor" is more similar to man than to woman

Such embeddings as input to downstream ML tasks

[(2)] Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in Neural Information Processing Systems* (2016).

Algorithmic (Un)Fairness: Examples

Example: **Amazon recruitment** [3]
In 2015, Amazon realized that their algorithm used for hiring employees was biased against women

- algorithm was based on the number of resumes submitted over the past ten years
- most of the applicants were men, it was trained to favor men over women.

[3] https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Example: **Dutch Tax authority fraud risk assessment**[4]

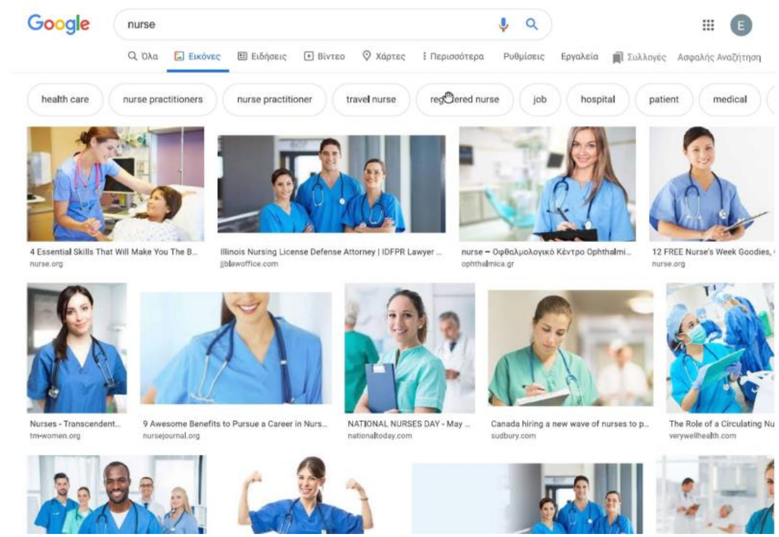
- Dutch tax authorities use a self-learning algorithm to create risk profiles to spot childcare benefits fraud.
- The criteria for the risk profile were developed by the tax authority, having dual nationality was marked as a big risk indicator, as was a low income.
- The Dutch tax authorities faces a €3.7 million fine

[4] https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Algorithmic (Un)Fairness: Examples

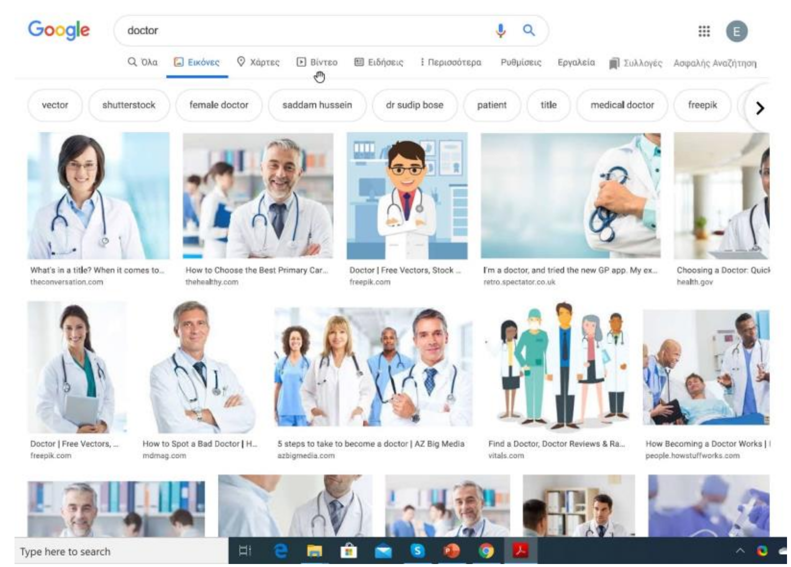
Example: **Image Search**[5]

What images do people choose to represent careers?
E.g., percentage of images portraying women in image search for professions



In search results:

- evidence for *stereotype exaggeration*
- systematic *underrepresentation of women (compared with the actual percentage as estimated by the US bureau of labor and statistics)*



[3] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. CHI 2015

Algorithmic (Un)Fairness: Examples

Example: **Health care risk assessment**

Black patients lose out on critical care when systems equate health needs with costs [6]

Used on more than 200 million people in US

- identify which patients will benefit from "high-risk care management" programs: access to specially trained nursing

Heavily favored white patients over black patients

- Race wasn't a variable, but **healthcare cost history** […]
- Black patients incurred lower health-care costs than white patients with the same conditions on average

Among all patients classified as very high-risk, black individuals turned out to have 26.3 percent more chronic illnesses than white ones

[6] https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

Algorithmic (Un)Fairness: Examples

Jobs ads[7]

Facebook Inc. *disproportionately shows certain types of job ads to men and women*.
Job ads were more likely to present job ads to users if their gender identity reflected the concentration of that gender in a particular position or industry

(study led by University of Southern California researchers)

Example:

Ads for delivery driver job listings that had similar qualification requirements but for different companies.

- The ads did not specify a specific demographic.
- One was an ad for Domino's pizza delivery drivers, the other for Instacart drivers.
- Instacart has more female drivers but Domino's has more male drivers.

Facebook targeted the Instacart delivery job to more women and the Domino's delivery job to more men.

[7] https://www.wsj.com/articles/facebook-shows-men-and-women-different-job-ads-study-finds-11617969600

Algorithmic (Un)Fairness: Examples

Facial recognition technology[8]

Last year, he was accused of reaching into a vehicle, grabbing a cellphone from a man and damaging it.

Officials concluded Oliver **had been misidentified** as the perpetrator and dismissed the case.

Detroit Police used **facial recognition technolog**y in the investigation.

- Facial recognition systems have been used by police forces f**or more than two decades**.
- While the technology works relatively well on white men, **the results are less accurate for other demographics**
  (Recent studies by M.I.T. and the National Institute of Standards and Technology)
- In part because of a **lack of diversity in the images** used to train the algorithm.

[8] https://eu.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/

Bias

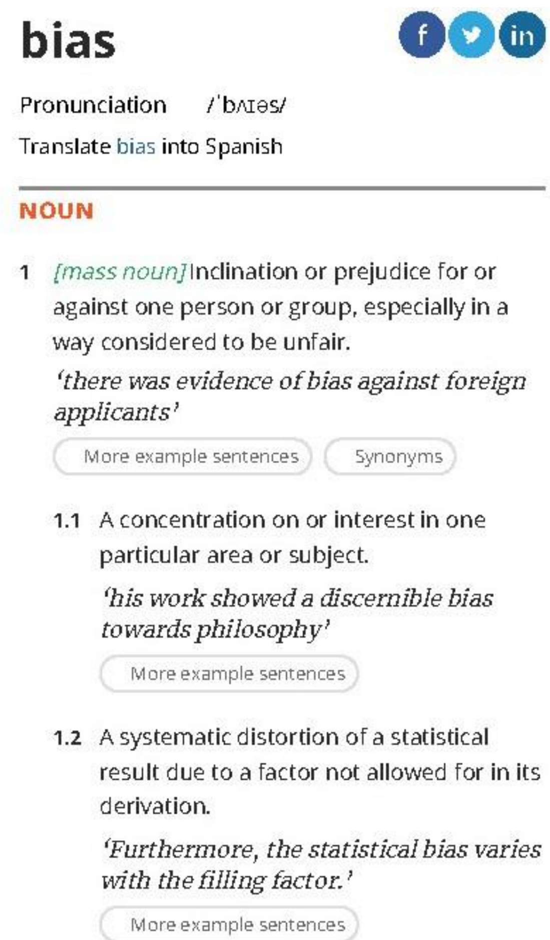
What is the cause of unfairness: bias

Bias

Overloaded term used to capture various forms of *misusing data and information, prejudice behavior, and favoritism.*
Also, various interpretations in ML

According to Oxford English Dictionary
- an inclination, or prejudice for, or against one person, or group, especially in a way considered to be unfair

Two different types
- Statistical
- Societal


bias

Pronunciation /ˈbʌɪəs/

Translate bias into Spanish

NOUN

1 *[mass noun]* Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.

'there was evidence of bias against foreign applicants'

More example sentences Synonyms

1.1 A concentration on or interest in one particular area or subject.

'his work showed a discernible bias towards philosophy'

More example sentences

1.2 A systematic distortion of a statistical result due to a factor not allowed for in its derivation.

'Furthermore, the statistical bias varies with the filling factor.'

More example sentences

What is the cause of unfairness: bias

Statistical bias

Non-representative sampling: mismatch between the sample/data used to train a predictive model, and the world as it currently is.

Sampling bias: the dataset selected as input to an algorithm is not representative of the full population to which the algorithm will be applied.

For example, missing MRI scans of healthy people, cardiovascular diseases for women

Selective labels (selection bias): the observed outcome depends on the choice of input.

For example, evaluating whether a loan will be repaid

Systematic measurement error, particularly when the error is greater for some groups than others

S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. 2020. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Annual Review of Statistics and Its Application 2021 8:1

What is the cause of unfairness: bias

Societal bias

objectionable social structures, human biases, and preferences that are:

- *reflected in data*,
- when *designing, implementing, evaluating and using* algorithms and systems.

Long list of biases: confirmation bias, normative biases, functional biases induced by the platforms, behavioral and temporal biases, cognitive biases

A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. 2019. A critical review of online social data: Biases, methodological pitfalls, and ethical boundaries. Front. Big Data 2, 13 (2019).

What is the cause of unfairness: bias

Bias may come from:

- **the actual data (garbage in, garbage out)**
  - if a survey contains biased questions [societal bias]
  - if some specific population is misrepresented in the input data [statistical bias]
    - Sample size disparity: learn on majority, errors concentrated in the minority class
  - if the data itself is a product of a historical process that operated to the disadvantage of certain groups - data as a social mirror [societal bias]

- **the algorithm**
  - reflecting, for example, commercial or other preferences of its designers [societal bias]
  - data processing [statistical bias]

- feedback loop [bias amplification] an algorithm receives biased data produces more biased output data and when this output data are fed back to this, or some other algorithm, *bias keeps increasing in an endless feedback loop*

FAIRNESS DEFINITIONS

Algorithmic Fairness: What?

Fairness is a general term, an elusive goal

Philosophical, ethical, political, judicial interpretations

Equality
Treat everyone the same

Equity
Treat everyone according to their needs

No barriers

(*) https://www.vocabulary.com/dictionary/fairness

Algorithmic Fairness: What?

Algorithmic fairness: Lack of discrimination
the results of an algorithm should not be influenced by *protected*, or *sensitive* attributes, such as gender, religion, age, sexual orientation, race, etc

Two levels:

- Individual fairness: Similar individuals should be treated in a similar manner

- Group fairness: Individuals are partitioned into groups according to their protected attributes. All groups should be treated fairly/similarly.

Depends on the algorithm: Classification, Recommendation, Ranking, Set Selection, Clustering, etc

Classification

Individual Fairness

Distance-based

Define a distance d between individuals and a distance D between the output

$$D\big(O(x), O(y)\big) \sim d(x, y)$$

How to define distances, especially in the input space

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard S. Zemel: Fairness through awareness. ITCS 2012: 214-226

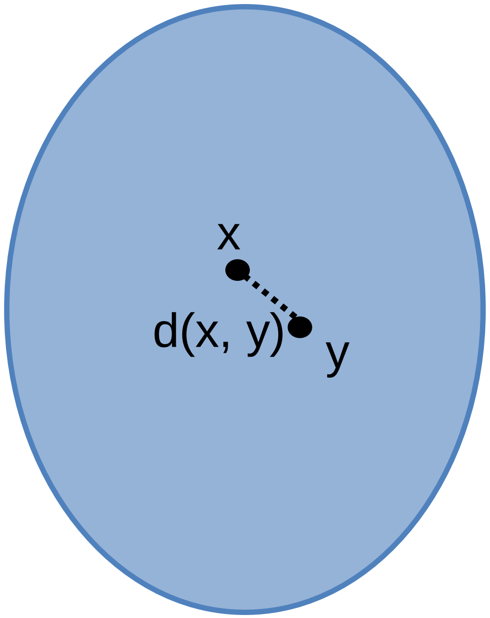
Individual Fairness

Similarity of ***input***

V be a set of individuals.

Distance metric d: $V \times V \rightarrow R$

- *Task-specific*
- Expresses *ground truth* (or, best available approximation)
- Externally imposed, e.g., by a regulatory body, or externally proposed, e.g., by a civil rights organization
- Made public, and open to discussion and refinement.

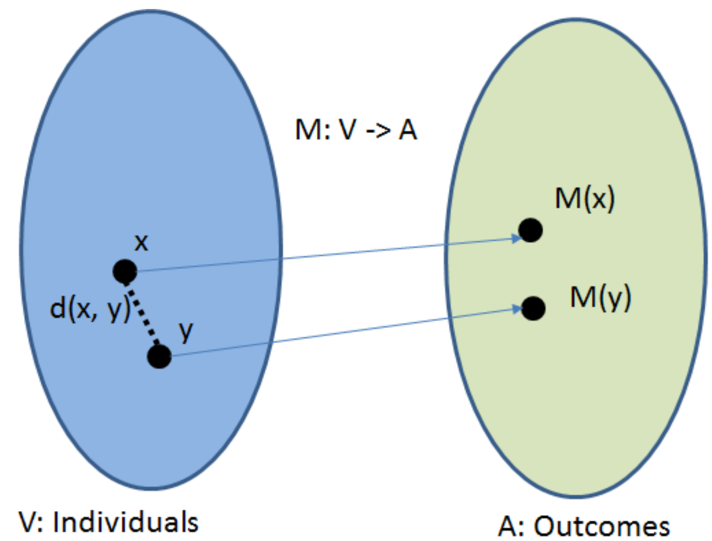


Individual Fairness

Similarity of ***outcome***

Probabilistic classifier M that maps individuals in V to probability distributions over outcomes *A*

- To classify *x* ∈ *V*, we choose an outcome *a* ∈ A according to distribution *M(x)*

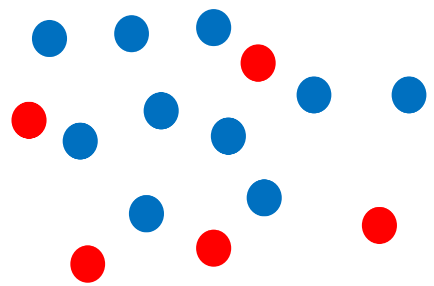


Lipschitz Mapping: a mapping M: V -> Δ(A) satisfies the (*D*, *d*)-Lipschitz property, if for every x, y ∈ V,

$$D\big(M(x) - M(y)\big) \leq L\, d(x, y)$$

where D is a distance measure between probability distributions

Group Fairness



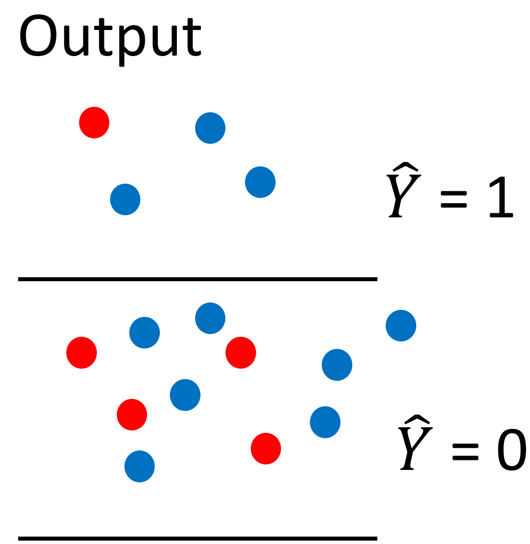
Individuals divided **into *groups*** based on the value of one or more protected attribute

Two groups

- $\boldsymbol{G^{+}}$: Protected (minority) group
- $\boldsymbol{G^{-}}$: Non protected (privileged) group

Group Fairness

Classification



Binary classifier

1 is the positive class (i.e., the class that leads to a favorable decision, e.g., getting a job, or being offer a specific medical treatment)

$\boldsymbol{Y}$ the actual output - ground truth
$\boldsymbol{\hat{Y}}$ the predicted output
$\boldsymbol{S}$ the predicted probability for a specific output

Group Fairness: parity

Compare the probability of a *favorable outcome for the* ***non-protected group*** with the probability of a *favorable outcome for the* ***protected group***

$$\frac{P[\hat{Y} = 1 \mid v \in G^{+}]}{P[\hat{Y} = 1 \mid v \in G^{-}]} = 1$$

demographic parity (statistical parity, independence)

preserves the input ratio: *the demographics of the individuals receiving a favorable outcome the same as demographics of the underlying population*

If there 10% of women among the applicants, 10% of those getting the job are women

Equity, or equality of output: members of each group have the *same chance of getting the favorable output*.

Instead of equal, some other value, e.g., 80% rule, or disparate impact

Group Fairness: error-based

Both the predicted output $\hat{Y}$ **and** the actual output Y

Confusion Matrix

	Actual $Y = 1$	Actual $Y = 0$
Predicted $\hat{Y} = 1$	**TP**	**FP**
Predicted $\hat{Y} = 0$	**FN**	**TN**

For example:

Knows as equal opportunity

$P[\hat{Y} = 1 \mid Y = 1, v \in G^-]$ *TP (true positive) rate* for the non-protected group

$P[\hat{Y} = 1 \mid Y = 1, v \in G^+]$ *TP rate* for the protected group

equal opportunity vs statistical parity: as with statistical parity, the members of the two groups have the same chance of getting the favorable outcome, **but only when** these members qualify

Equal opportunity is closer to an ***equality*** interpretation of fairness

Equalized odds: both true and false positive rates equal for the two groups

Group Fairness: calibration

In probabilistic classifiers where the output is the *probability* that an individual belongs to the positive class), we want the estimates to be *well-calibrated*:

if the algorithm identifies a set of people as having a probability $\boldsymbol{p}$ of belonging to the positive class, then approximately a $\boldsymbol{p}$ fraction is indeed positive instances

We want the classifier to be equally well-calibrated for both groups, for any predicted probability p in [0,1]:

$$P[Y = 1 \,|S = p, v \in G^{-}] = P[Y = 1 \,|S = p, v \in G^{+}]$$

Group Fairness (other names)

- Independence (demographic parity)
- Separation (error rates)
- Sufficiency (calibration)

Counterfactual Fairness

A decision is fair towards an individual, if it is the same in both the actual world and *a counterfactual world* where the individual belonged to a different group

Individual fairness	Two data points
Group fairness	Two demographic groups
Counterfactual fairness	A data point and its counterfactual

Casual inference

ACHIEVING FAIRNESS

Achieving Fairness: How

- In general, there are three approaches to achieving fairness:
  - Pre-processing: Preprocess the data
  - In-processing: Change the algorithm
  - Post-processing: Tweak the output

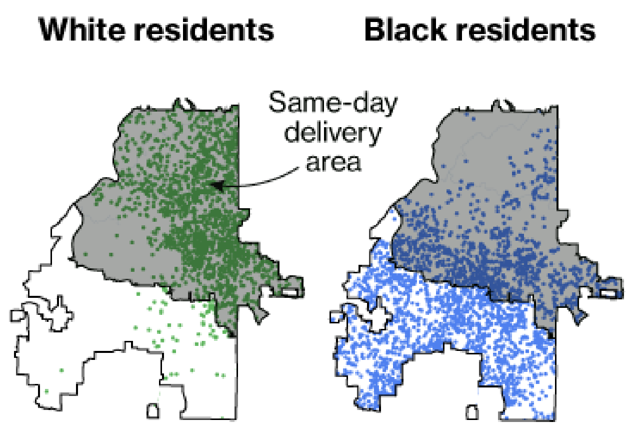
Pre-processing : omit the protected attribute

Blindness/Unawareness: omit/hide the value of the protected attribute

- Other *proxy* attributes correlated with the protected ones (also known as redundant encoding).

Redlining: the practice of arbitrarily denying or limiting financial services to specific *neighborhoods (based on zip codes (*))*

The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.



(*) Amazon doesn't consider the race of its customers. Should It? Ingold, D. and Soper, S., 2016. Bloomberg News.

Pre-processing: overview

	bias in rows	bias in columns	fairness	algorithm
Suppression		✔	group	any
Class Relabeling	✔		group	ranker
Reweighing		✔	individual / group	ranker
Data Transformation		✔	individual / group	ranker
Database Repair	✔	✔	group	ranker
Data Augmentation	✔	✔	individual / group	matrix factorization

remove data

modify data

add data

recommender-oriented

Bias in the rows

when there are not enough representative individuals from minority (sub)groups

Bias in the columns

when features are biased (correlated) with sensitive attributes.

Pre-processing: Class relabeling

Changes the labels of some objects in the dataset to remove the discrimination from the input data.

The method:

- Consider a subset of data from the *minority* group as promotion candidates, and a subset of the *majority* group is chosen as demotion candidates.

- How to select candidates:
  - Learn a classifier; rank the tuples based on their probability of having positive labels
  - Select the *top k of minority* (for promotion) and *the bottom k of majority* (for demotion)

- Flip their labels

Lowering the discrimination will result in lowering the accuracy and vice versa

Intrusive

In-processing: Overview

Depends on the algorithm

Common approaches in learning based:

I. Learning fair representations
II. Adding regularization terms to the objective function

In-processing: learning fair representations

Basic idea:

- Introduce *an intermediate level* **Z** between the input space **X** that represents individuals and the output space **Y** that represents classification outcomes

X ⇒ Z ⇒ Y

Input: individuals — **Fair representation** — Classification output

Z: representation of **X**

- best encodes X and
- obfuscates any information about membership in the protected group

Z is a multinomial random variable of size *k* where each of the *k* values represents *a prototype (cluster)* in the space of **X**.

- As in pre-processing, but now part of the optimization objective

In-processing: learning fair representations

A learning system that minimizes the loss function

Quality of the encoding — Fairness — Accuracy

$$L = \lambda_x L_x + \lambda_z L_z + \lambda_y L_y$$

Distance from points in X to their representation in Z should be small

Statistical parity

Prediction based on the representation should be accurate

$X \Rightarrow Z \Rightarrow Y$

Input: individuals

Fair representation

Classification output

$\lambda_x, \lambda_z, \lambda_y$ hyper-parameters that control the trade-off among the three objectives

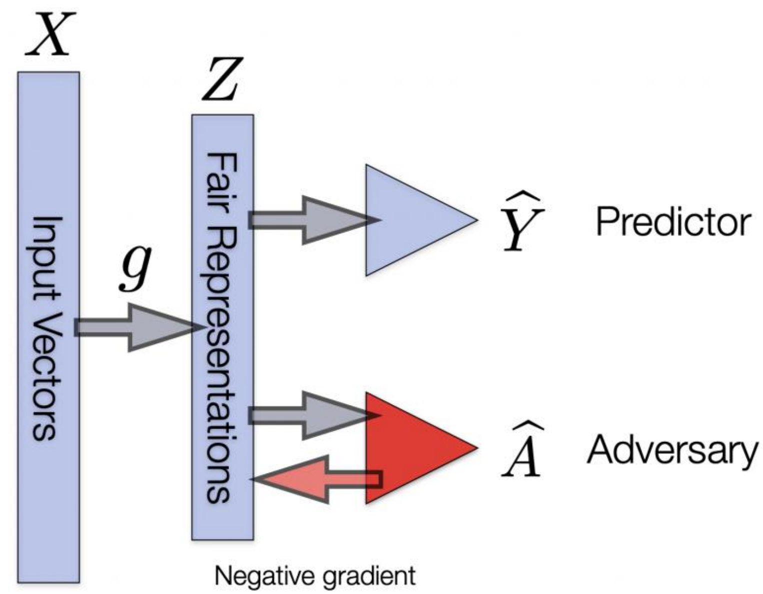
Statistical parity

$$P(z = k | x \in \boldsymbol{G^+}) = P(z = k | x \in \boldsymbol{G^-}) \; \forall \, k$$

The probability that a random element of the protected group maps to a particular prototype of Z is equal to the probability that a random element of the non-protected group maps to the same prototype

In-processing: learning fair representations

Adversarial leaning



Simultaneously two goals:
(1) Predictor Accuracy
(2) Fool the Adversary

The ***adversary*** is trying to predict the relevant sensitive variable from the representation, and so minimizing the performance of the adversary ensures there is little or no information in the representation about the sensitive variable.

In-processing: Regularization

$$L = L_{original} + \lambda\, L_{fairness}$$

The DELTR approach extends the ListNet learning to rank approach

$$L_{DELTR}\big(r(q), \hat{r}(q)\big) = \underbrace{L_{LN}\big(r(q), \hat{r}(q)\big)}_{\text{Accuracy}} + \boxed{\lambda\, \mathrm{F}\big(\hat{r}(q)\big)}\ \text{Unfairness}$$

- λ specifies the desired trade-offs between ranking utility and fairness

$$F\big(\hat{r}(q)\big) = \max(0, (\boldsymbol{Exposure}\big(\boldsymbol{G^{+}}\,\big|\,\boldsymbol{P_{\hat{r}(q)}}\big) - \boldsymbol{Exposure}\big(\boldsymbol{G^{-}}\,\big|\,\boldsymbol{P_{\hat{r}(q)}}\big))^{2})$$

- squared hinge loss: if the protected group already receives as much exposure as the non-protected, just optimize for accuracy
- prefers rankings in which the exposure of the protected group is not less than the exposure of the non protected group but not vice versa

Post-Processing: Constraint optimization

As an optimization problem

F: a fairness measure

U: a measure of the utility (accuracy)

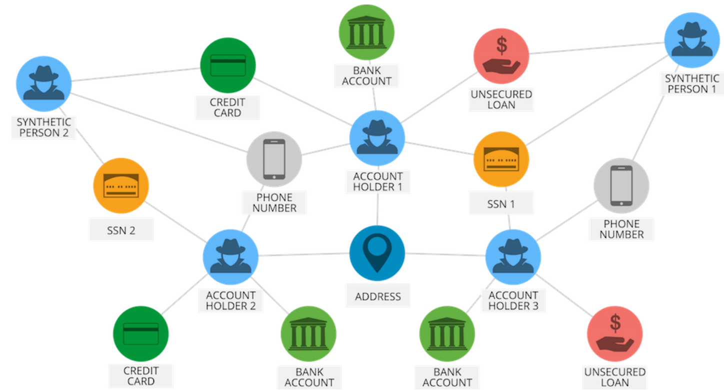
There are two general ways of formulating an optimization problem involving fairness F and utility U namely:

- maximizing fairness subject to a constraint in utility
- maximizing utility subject to a constraint in fairness

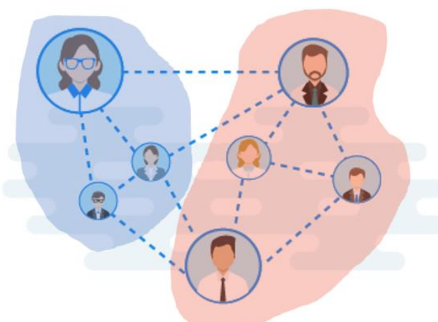
FAIRNESS IN NETWORKS

Graph Mining: Applications

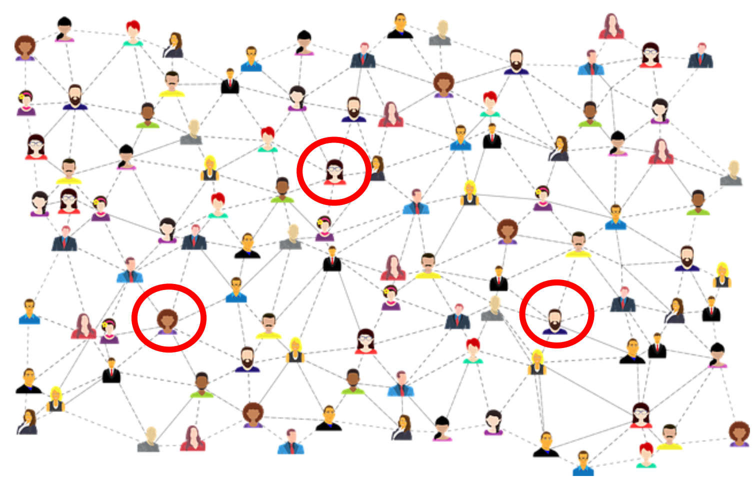
Credit scoring



Financial fraud detection



User community detection

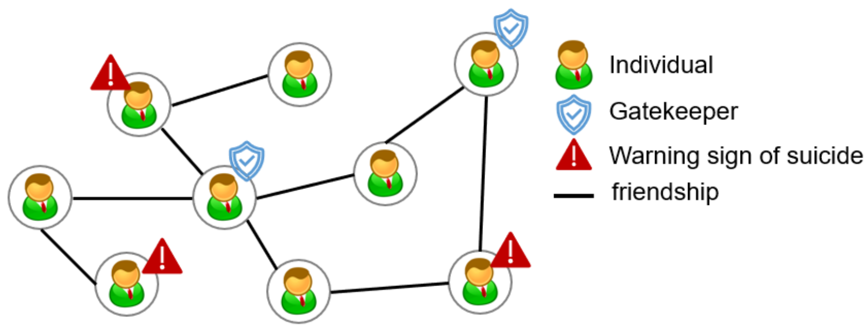


Identifying influencers

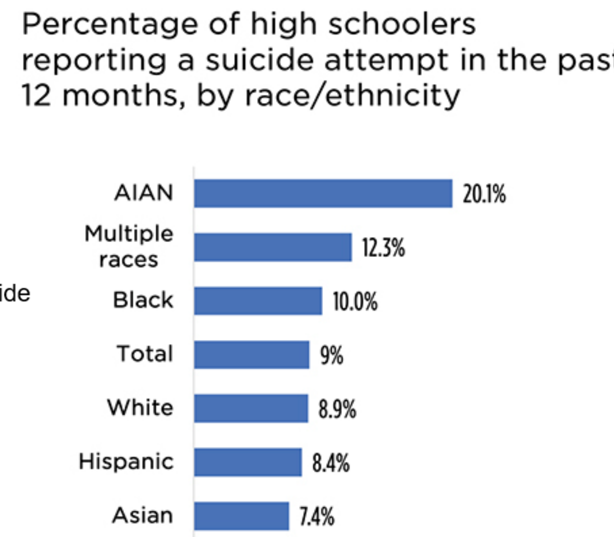
Algorithmic Fairness on Graphs: Suicide Prevention

- Suicide is one of the leading causes of death in US

Gatekeeper training programs



Toy example of a gatekeeper training program



Suicide attempts by race/ethnicity

- **Observation:** existing suicide prevention efforts disproportionately affect individuals of different demographics

[1] https://www.cdc.gov/nchs/data/vsrr/vsrr024.pdf
[2] https://988lifeline.org/
[3] https://www.childtrends.org/publications/addressing-discrimination-supports-youth-suicide-prevention-efforts

Challenge

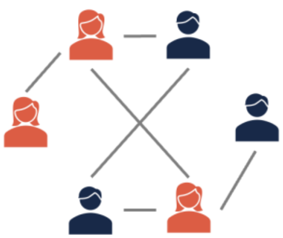
- **Assumption**

	Classic machine learning	**Graph mining**
Data	IID samples	Non-IID graph

  - **IID:** independent and identically distributed
- **Example**

- Individuals are independent
- Cannot affect others

Classic machine learning



- Individuals are connected
- Can affect others through connection(s)

Graph mining

- **Challenges:** implication of non-IID nature on
  - Measuring bias
    - Dyadic fairness, degree-related fairness
  - Mitigating unfairness
    - Enforce fairness by graph structure imputation

Roadmap

- Network Centrality Fairness

- Fair Graph Embeddings

The Pagerank Algorithm

- The best-known algorithm for measuring the centrality/importance of nodes in a graph, introduced by Google
- **Assumption:** important webpage → linked by many others

- Pagerank performs a random walk with restarts:
- At each step of the random walk:
  - With probability c perform a transition according to the transition probability matrix $\boldsymbol{A}$
  - With probability $1 - c$ restart to a randomly selected node according to teleportation (jump) vector $\boldsymbol{e}$
- The Pagerank vector is the stationary distribution $\boldsymbol{r}$ of this random walk

Preliminary: PageRank

- **Formulation**
  - Iterative method for the following linear system

$$\mathbf{r} = c\mathbf{A}^T\mathbf{r} + (1-c)\mathbf{e}$$

    - $\mathbf{A}$: transition matrix
    - $\mathbf{r}$: PageRank vector
    - c: damping factor
    - $\mathbf{e}$: teleportation vector
  - Closed-form solution

$$\mathbf{r} = (1-c)(\mathbf{I} - c\mathbf{A}^T)^{-\mathbf{1}}\mathbf{e}$$

- **Variants**
  - Personalized PageRank (PPR)

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab 1999.
[2] Haveliwala, T. H. (2003). Topic-sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. TKDE 2003.
[3] Tong, H., Faloutsos, C., & Pan, J. Y. (2006). Fast Random Walk with Restart and Its Applications. ICDM 2006.

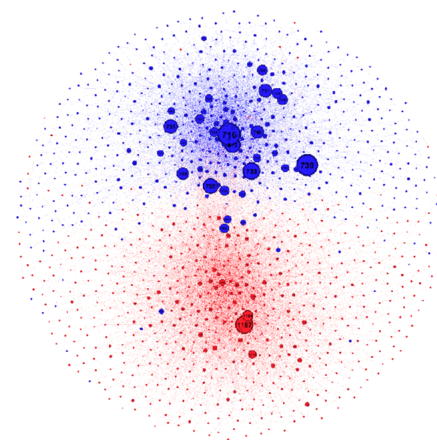
Unfairness in PageRank

- Pagerank distributes the importance values to the nodes in the network
  - But is it fair?
- **Example**
  - **Network:** 1222 nodes of political blogs
  - **Groups:** red (left-leaning) and blue (right-leaning)

Unfair ranking

Similar number of red nodes vs. blue nodes (48% red vs. 52% blue)

Much less PageRank mass of red nodes (33% red vs. 67% blue)

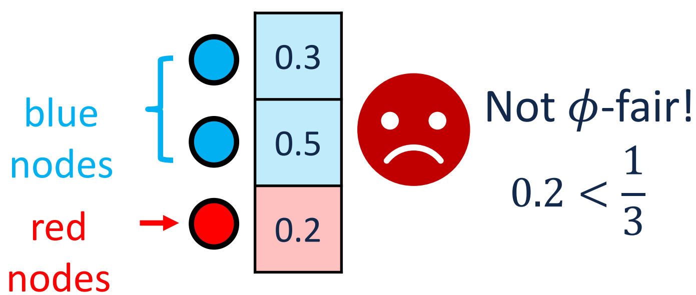


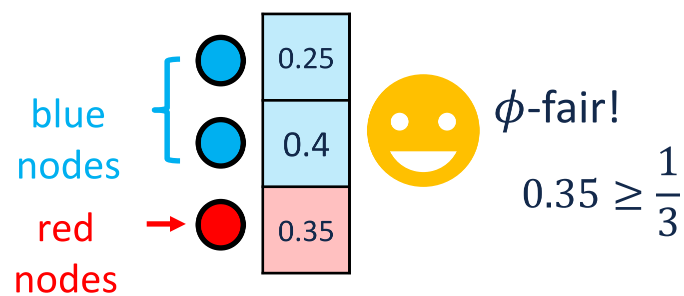
- How can we define Pagerank fairness?
- How do we make Pagerank fair?

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.
[2] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Fairness Measure: ϕ-Fairness

- **Given:** (1) a graph G; (2) a parameter ϕ
- **Definition:** a PageRank vector is ϕ-fair if at least ϕ fraction of total PageRank mass is allocated to the protected group
- **Variants and generalizations**
  - Statistical parity → ϕ = fraction of protected group
  - Affirmative action → ϕ = a desired ratio (e.g., 20%)
- **Example**
  - Protected group = red nodes
  - $\phi = 1/3$





[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.
[2] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Problem Definition: Fair PageRank

- **Given**
  - A graph with transition matrix $\mathbf{A}$
  - Partitions of nodes
    - Red nodes ($\mathcal{R}$): protected group
    - Blue nodes ($\mathcal{B}$): unprotected group
- **Produce:** a fair PageRank vector $\tilde{\mathbf{r}}$ that is
  - ϕ-fair
  - Close to the original PageRank vector $\mathbf{r}$ (minimizes the utility loss)

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.
[2] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Fair PageRank: Solutions

- **Recap:** closed-form solution for PageRank

$$\mathbf{r} = (1 - c)(\mathbf{I} - c\mathbf{A}^T)^{-1}\mathbf{e}$$

- **Parameters in PageRank**
  - **Damping factor** c avoids sinks in the random walk (i.e., nodes without outgoing links)
  - **Teleportation vector e** controls the starting node where a random walker restarts
    - Can we control where the walker teleports to? ← Solution #1: fairness-sensitive PageRank
  - **Transition matrix** $\mathbf{A}$ controls the next step where the walker goes to
    - Can we modify the transition probabilities?
    - Can we modify the graph structure?

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.
[2] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Solution #1: Fairness-sensitive PageRank

- **Intuition**
  - Find a teleportation vector $\mathbf{e}$ to make PageRank vector ϕ-fair

$$\mathbf{r} = \mathbf{Q}^T\mathbf{e}, \quad \mathbf{Q}^T = (1-c)(\mathbf{I} - c\mathbf{A}^T)^{-1}$$

  - Keep transition matrix $\mathbf{A}$ and $\mathbf{Q}^T$ fixed
- **Observation:** mass of PageRank $\mathbf{r}$ w.r.t. red nodes $\mathcal{R}$

$$\mathbf{r}(\mathcal{R}) = \mathbf{Q}^T[\mathcal{R}, :]\mathbf{e}$$

  - $\mathbf{Q}^T[\mathcal{R}, :]$: rows of $\mathbf{Q}^T$ w.r.t. nodes in set $\mathcal{R}$
- **(Convex) optimization problem**

$$\begin{aligned} \min_{\mathbf{e}} \quad & \boxed{\|\mathbf{Q}^T\mathbf{e} - \mathbf{r}\|^2} \\ \text{s.t.} \quad & \boxed{\begin{array}{l} \mathbf{e}[i] \in [0,1], \forall i \\ \|\mathbf{e}\|_1 = 1 \end{array}} \\ & \boxed{\|\mathbf{Q}^T[\mathcal{R}, :]\mathbf{e}\|_1 = \phi} \end{aligned}$$

The fair PageRank $\mathbf{Q}^T\mathbf{e}$ is as close as possible to the original PageRank $\mathbf{r}$

The teleportation vector $\mathbf{e}$ is a probability distribution

The fair PageRank $\mathbf{Q}^T\mathbf{e}$ is ϕ-fair

  - Can be solved by any convex optimization solvers

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

Fairness-sensitive PageRank: Example

- **Settings:** $\phi = 1/3$ and protected node = red node
- **Original PageRank**

rows w.r.t. blue nodes / row w.r.t. red nodes

$$\mathbf{r} = \mathbf{Q}^T\mathbf{e} = \begin{bmatrix} 0.55 & 0.3 & 0.2 \\ 0.4 & 0.6 & 0.2 \\ 0.05 & 0.1 & 0.6 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 0.35 \\ 0.4 \\ 0.25 \end{bmatrix}$$

Not ϕ-fair! $0.25 < \frac{1}{3}$

- **Fairness-sensitive PageRank**

rows w.r.t. blue nodes / row w.r.t. red nodes

$$\tilde{\mathbf{r}} = \mathbf{Q}^T\tilde{\mathbf{e}} = \begin{bmatrix} 0.55 & 0.3 & 0.2 \\ 0.4 & 0.6 & 0.2 \\ 0.05 & 0.1 & 0.6 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 0.31 \\ 0.35 \\ 0.34 \end{bmatrix}$$

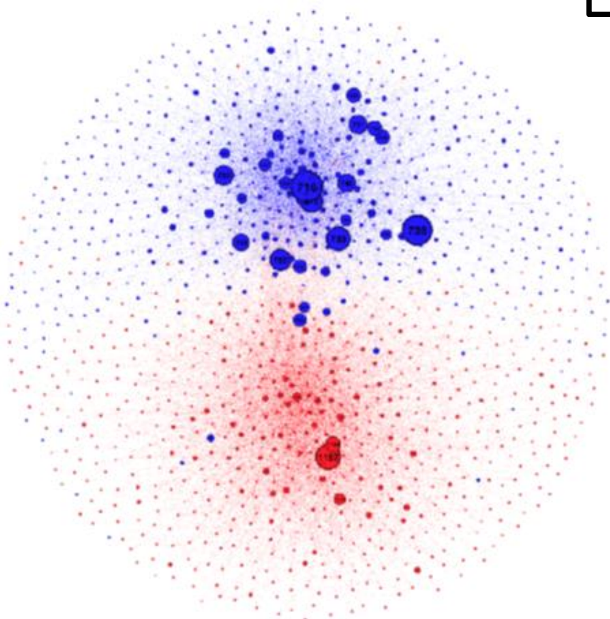
ϕ-fair! $0.34 \geq \frac{1}{3}$

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

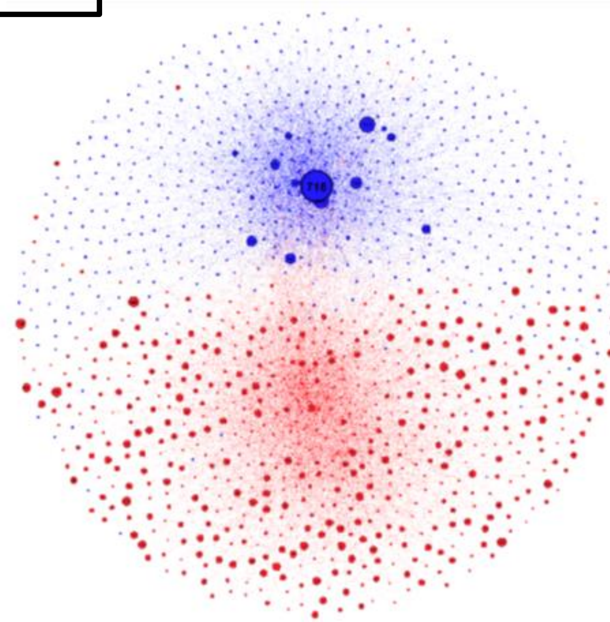
Fairness-sensitive PageRank: Experiment

- **Observation:** the teleportation vector allocates more weight to the red nodes, especially nodes at the periphery of the network
  - More likely to (1) restart at red nodes and (2) walk to other red nodes more often

PageRank



Fairness-sensitive PageRank teleportation vector



NOTE: size is proportional to score in the teleportation vector

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

Fair PageRank: Solutions

- **Recap:** closed-form solution for PageRank

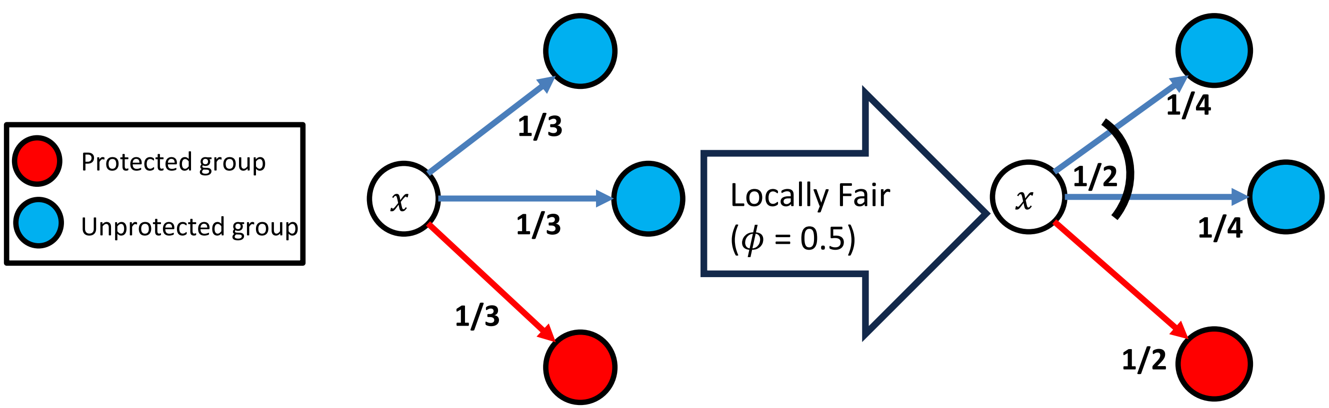
$$\mathbf{r} = (1 - c)(\mathbf{I} - c\mathbf{A}^T)^{-1}\mathbf{e}$$

- **Parameters in PageRank**
  - **Damping factor** c avoids sinks in the random walk (i.e., nodes without outgoing links)
  - **Teleportation vector e** controls the starting node where a random walker restarts
    - Can we control where the walker teleports to?
  - **Transition matrix** $\mathbf{A}$ controls the next step where the walker goes to
    - Can we modify the transition probabilities? Solution #2: locally fair PageRank
    - Can we modify the graph structure?

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.
[2] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Solution #2: Locally Fair PageRank

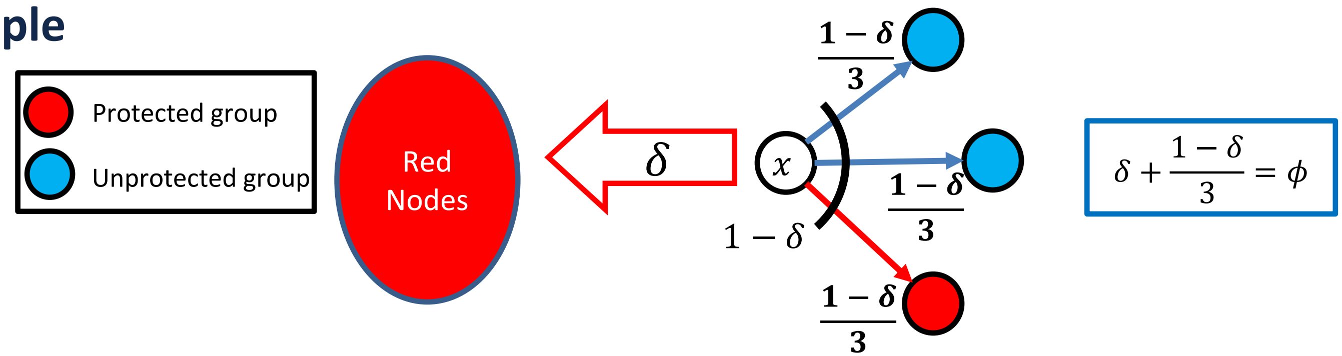
- **Intuition:** adjust the transition matrix A to obtain a fair random walk
- **Neighborhood locally fair PageRank**
  - **Key idea:** jump with probability ϕ to red nodes and (1- ϕ) to blue nodes
  - **Example**



[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

Solution #2: Locally Fair PageRank

- **Residual locally fair PageRank**
  - **Key idea:** jump with
    - Equal probability to 1-hop neighbors
    - A residual probability δ to the other red nodes
  - **Example**



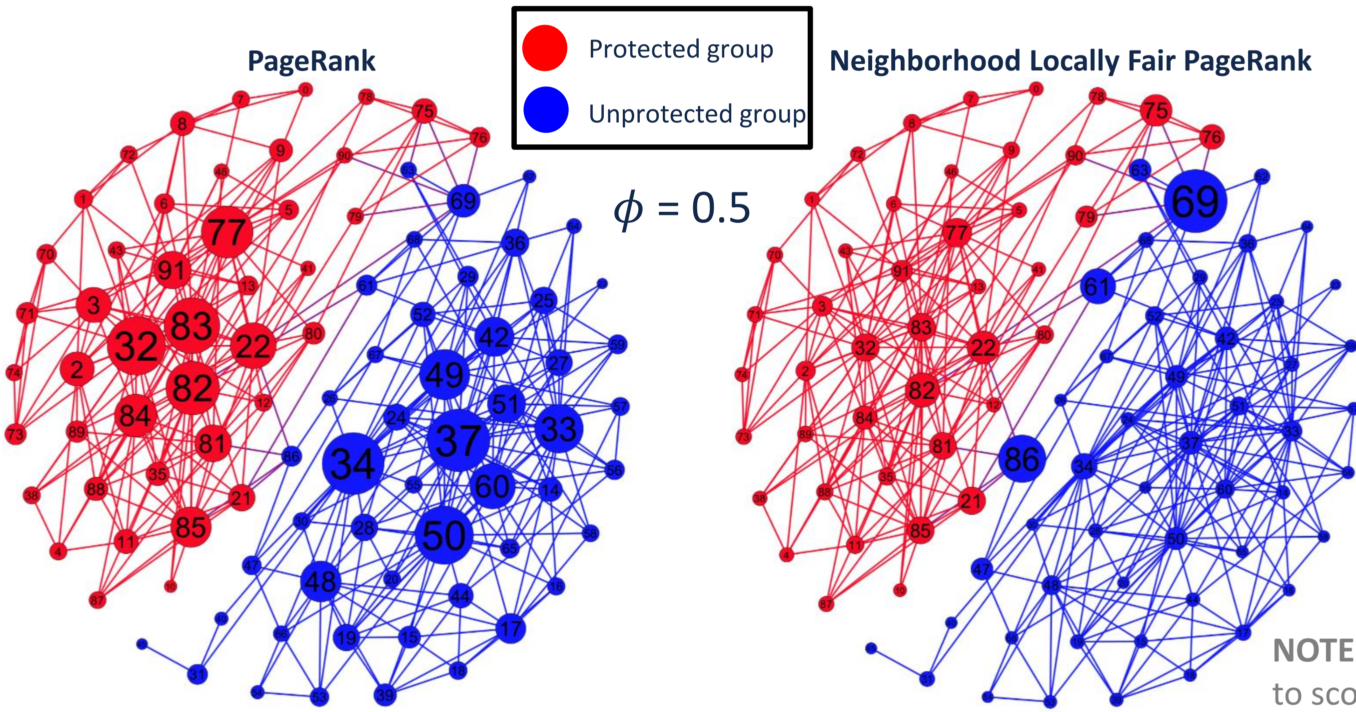
- **Residual allocation policies:** neighborhood allocation, uniform allocation, proportional allocation, optimized allocation

- **Neighborhood allocation:** allocate the residual to protected neighbors, equivalent to neighborhood locally fair PageRank
- **Uniform allocation:** uniformly allocate the residual to all protected nodes
- **Proportional allocation:** allocated the residual to all protected nodes proportionally to their PageRank score
- **Optimized allocation:** allocate the residual to all protected nodes while minimizing the difference with original PageRank score

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

Locally Fair PageRank: Experiment

- **Observation:** PageRank weight is shifted to the blue nodes at boundary



$\phi = 0.5$

NOTE: size is proportional to score in the PageRank vector

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

Fair PageRank: Solutions

- **Recap:** closed-form solution for PageRank

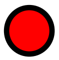
$$\mathbf{r} = (1 - c)(\mathbf{I} - c\mathbf{A}^T)^{-1}\mathbf{e}$$

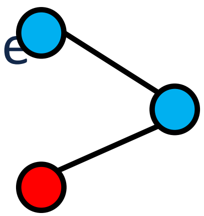
- **Parameters in PageRank**
  - **Damping factor** c avoids sinks in the random walk (i.e., nodes without outgoing links)
  - **Teleportation vector e** controls the starting node where a random walker restarts
    - Can we control where the walker teleports to?
  - **Transition matrix** $\mathbf{A}$ controls the next step where the walker goes to
    - Can we modify the transition probabilities?
    - Can we modify the graph structure?

Solution #3: best fair edge identification

[1] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.
[2] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Solution #3: Best Fair Edge Identification

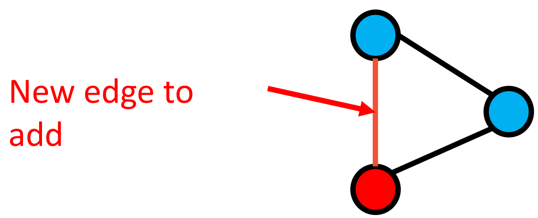
- **Intuition:** add edges that can improve the PageRank fairness to the graph
- **Example**
  -  = protected node
  - $\phi = 1/3$



$$\mathbf{r} = \mathbf{Q}^T \mathbf{e} = \begin{bmatrix} 0.257 \\ 0.486 \\ 0.257 \end{bmatrix}$$

Not ϕ-fair!

$$\frac{0.257}{0.257 + 0.486 + 0.257} < \frac{1}{3}$$



$$\tilde{\mathbf{r}} = \tilde{\mathbf{Q}}^T \mathbf{e} = \begin{bmatrix} 0.333 \\ 0.333 \\ 0.333 \end{bmatrix}$$

ϕ-fair!

$$\frac{0.333}{0.333 + 0.333 + 0.333} = \frac{1}{3}$$

- **Question:** how to find the edges with the highest improvement?

[1] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Best Fair Edge Identification: Problem Definition

- **Given**
  - $G = (\mathcal{V}, \mathcal{E})$
  - $\mathcal{S} \subseteq \mathcal{V}$: protected node set
  - $r_{\mathcal{E}}(\mathcal{S}) = \sum_{i \in \mathcal{V}} r_{\mathcal{E}}(i)$: total PageRank mass of nodes in $\mathcal{S}$ on graph with edge set $\mathcal{E}$
- **Fairness gain of edge addition**

$$\mathrm{gain}(x, y) = \boxed{r_{\mathcal{E} \cup (x,y)}(\mathcal{S}) - r_{\mathcal{E}}(\mathcal{S})}$$

Naive method
Exhaustively recompute PageRank with the addition of **each** node pair

- **Goal:** find the edge $(x, y), \forall x, y \in \mathcal{V}$, such that

$$\underset{(x,y)}{\mathrm{argmax}} \quad \mathrm{gain}(x, y)$$

- **Question:** how to efficiently compute the gain?

[1] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Best Fair Edge Identification: Fairness Gain

- **Main result:** Adding an edge to the graph is a **rank-1 perturbation** of the transition matrix
- We can estimate the gain as:

Closeness of target node y to $\mathcal{S}$

The average 'closeness' of neighbors of x to $\mathcal{S}$

$$\text{gain}(x, y) = r_{\mathcal{E}}(x) \frac{\frac{c}{1-c}\left(r_{\mathcal{E}}(\mathcal{S}|y) - \frac{1}{d_x}\sum_{u\in\mathcal{N}_x} r_{\mathcal{E}}(\mathcal{S}|u)\right)}{d_x + \frac{c}{1-c}\left(\frac{1}{d_x}\sum_{u\in\mathcal{N}_x} r_{\mathcal{E}}(x|u) - r_{\mathcal{E}}(x|y)\right) + 1}$$

degree of source node

Average proximity of node x's neighbors to x

  - $r_{\mathcal{E}}(x|y)$: personalized PageRank (PPR) score of node x, with query node y, based on edge set $\mathcal{E}$
  - $r_{\mathcal{E}}(\mathcal{S}|y) = \sum_{i\in\mathcal{S}} r_{\mathcal{E}}(i|y)$: total PPR mass of nodes in $\mathcal{S}$, with query node y, based on edge set $\mathcal{E}$
- $r_{\mathcal{E}}(x)$: node x should have high PageRank score
- d_x: node x should have small degree
- $r_{\mathcal{E}}(x|y) - \frac{1}{d_x}\sum_{u\in\mathcal{N}_x} r_{\mathcal{E}}(x|u)$: node y is close to node x
- $r_{\mathcal{E}}(\mathcal{S}|y) - \frac{1}{d_x}\sum_{u\in\mathcal{N}_x} r_{\mathcal{E}}(\mathcal{S}|u)$: node y is closer to $\mathcal{S}$ than the neighborhood of x

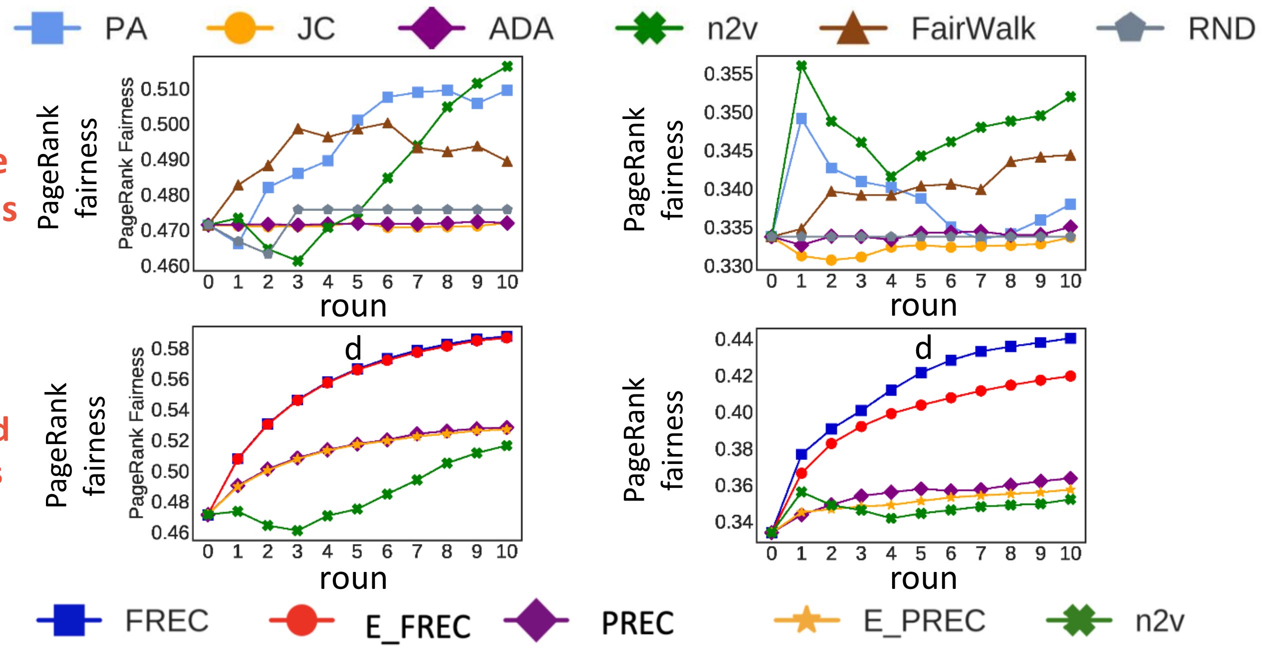
[1] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Best Fair Edge Identification: Experiment

- **Observation:** the proposed method find the best edges to improve PageRank fairness

Baseline Methods

Proposed Methods



Books Dataset

Blogs Dataset

- FREC: select edge (x, y) with highest $\text{gain}(x, y) = \Lambda(x, y) p_{\mathcal{E}}(x)$
- PREC: select edge (x, y) with highest $\text{gain}(x, y \mid x) = \Lambda(x, y) p_{\mathcal{E}}(x|x)$
- E_FREC: select edge (x, y) with highest $\text{gain}(x, y) p_{\text{acc}}(x, y)$
- E_PREC: select edge (x, y) with highest $\text{gain}(x, y \mid x) p_{\text{acc}}(x, y)$

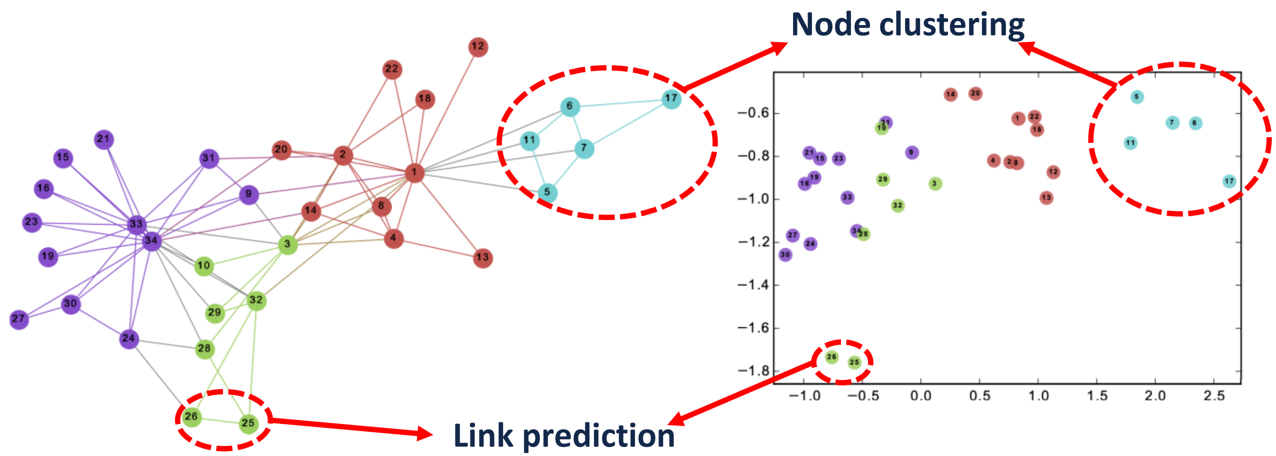
[1] Tsioutsiouliklis, S., Pitoura, E., Semertzidis, K., & Tsaparas, P. (2022). Link Recommendations for PageRank Fairness. WWW 2022.

Roadmap

- Network Centrality Fairness
- Fair Graph Embeddings

Preliminary: Node Embedding

- **Motivation:** learn low-dimensional node representations that preserve structural/attributive information
- **Applications**
  - –Node classification
  - –Link prediction
  - –Node visualization



(a) Input: Karate Graph (b) Output: Representation

Visualization of Node Embedding

[1] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). Deepwalk: Online Learning of Social Representations. KDD 2014.
[2] Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. KDD 2016.
[3] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. NeurIPS 2013.

Graph embeddings

- Graph embeddings utilize only the graph structure to derive the node representation
- In broad terms, the embedding of a node depends on the embeddings of the k-hop neighborhood of the node
- Since neighboring nodes tend to have similar (sensitive) attributes, the embeddings are likely to encode information about the sensitive attributes
  - Therefore, they are **biased**

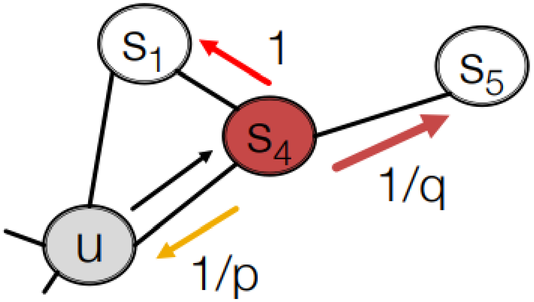
- How can we remove these biases?

Graph unfairness

- Homophily-based metrics:
  - E.g., the fraction of edges that link nodes with the same sensitive attribute value
- Neighborhood metrics:
  - The entropy of the label distribution of the neighborhood of a node.
- Preprocessing approach:
  - Change the graph (e.g., via edge rewiring or edge additions) to improve fairness

Preliminary: Random Walk-based Node Embedding

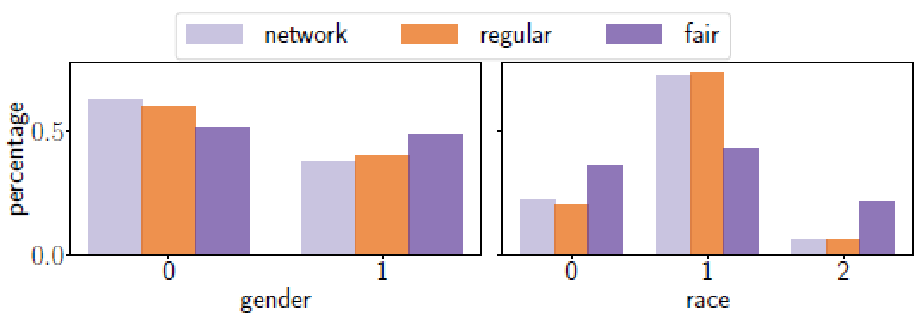
- **Goal:** learn node embeddings that are predictive of nodes in its neighborhood
- **Key idea**
  - Simulate random walk as a sequence of nodes
  - Apply skip-gram technique to predict the context node
- **Example**
  - **DeepWalk:** random walk for sequence generation
  - **Node2vec:** biased random walk for sequence generation
    - **Return parameter** p: how fast the walk **explores** the neighborhood of the starting node
    - **In-out parameter** q: how fast the walk **leaves** the neighborhood of the starting node



[1] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). Deepwalk: Online Learning of Social Representations. KDD 2014.
[2] Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. KDD 2016.

Fairwalk: Solution

- **Key idea:** modify the random walk procedure in node2vec
- **Steps of Fairwalk**
  - Partition neighbors into demographic groups
  - Uniformly sample a demographic group to walk to
  - Randomly select a neighboring node within the chosen demographic group
- **Example:** ratio of each demographic group
  - Original network vs. regular random walk vs. fair random walk



[1] Rahman, T., Surma, B., Backes, M., & Zhang, Y. (2019). Fairwalk: Towards Fair Graph Embedding. IJCAI 2019.

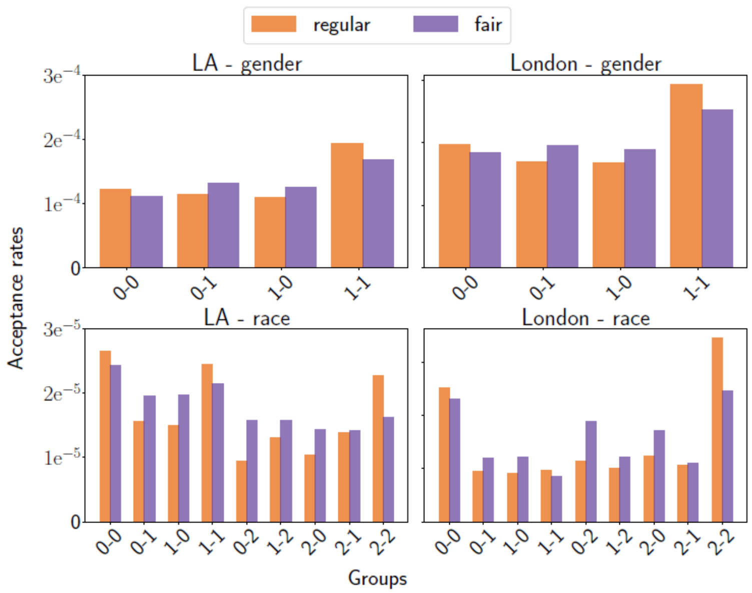
Fairwalk vs. Existing Works

- **Fairwalk vs. node2vec**
  - **Node2vec:** skip-gram model + walk sequences by original random walk
  - **Fairwalk:** skip-gram model + walk sequences by fair random walk
- **Fairwalk vs. fairness-aware PageRank**
  - **Fairness-aware PageRank:** the minority group should have a certain proportion of PageRank probability mass
  - **Fairwalk:** all demographic group have the same random walk transition probability mass

[1] Rahman, T., Surma, B., Backes, M., & Zhang, Y. (2019). Fairwalk: Towards Fair Graph Embedding. IJCAI 2019.
[2] Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. KDD 2016.
[3] Tsioutsiouliklis, S., Pitoura, E., Tsaparas, P., Kleftakis, I., & Mamoulis, N. (2021). Fairness-Aware PageRank. WWW 2021.

Fairwalk: Results on Statistical Parity

- **Observations**
  - Fairwalk achieves a more balanced acceptance rates among groups
  - Fairwalk increases the fraction of cross-group recommendations



[1] Rahman, T., Surma, B., Backes, M., & Zhang, Y. (2019). Fairwalk: Towards Fair Graph Embedding. IJCAI 2019.