Machine Learning for Data Science (CS4786)
Lecture 26

Fairness Through Awareness

Fairness Through Awareness

- Setup:

Fairness Through Awareness

- Setup:
  - Variable T: indicated protected class or not

Fairness Through Awareness

- Setup:
  - Variable T: indicated protected class or not
  - Output variable O: Indicates our prediction or outcome

Fairness Through Awareness

- Setup:
  - Variable T: indicated protected class or not
  - Output variable O: Indicates our prediction or outcome
  - Variable Y: indicates true/target/desired outcome (eg. Individual capable/qualified, individual can afford etc.)

Fairness Through Awareness

- Setup:
  - Variable T: indicated protected class or not
  - Output variable O: Indicates our prediction or outcome
  - Variable Y: indicates true/target/desired outcome (eg. Individual capable/qualified, individual can afford etc.)

Demographic Parity

$$P(O=1|T=1) = P(O=1|T=0)$$

Fairness Through Awareness

- Setup:
  - Variable T: indicated protected class or not
  - Output variable O: Indicates our prediction or outcome
  - Variable Y: indicates true/target/desired outcome (eg. Individual capable/qualified, individual can afford etc.)

Demographic Parity

$$P(O=1|T=1) = P(O=1|T=0)$$

Problem: when T=0, O can correlate with Y and if T=1, O can be random

Equalized Odds

Sufficiency or Predictive Rate Parity

Equalized Odds

For all o, y in {0,1}

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

Sufficiency or Predictive Rate Parity

Equalized Odds

For all o, y in {0,1}

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- O is independent of T given Y

Sufficiency or Predictive Rate Parity

Equalized Odds

For all o, y in {0,1}

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- O is independent of T given Y

Sufficiency or Predictive Rate Parity

For all o, y in {0,1}

$$P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)$$

Equalized Odds

For all o, y in {0,1}

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- O is independent of T given Y

Sufficiency or Predictive Rate Parity

For all o, y in {0,1}

$$P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)$$

- Y is independent of T given O

Equalized Odds

For all o, y in {0,1}

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- O is independent of T given Y

Sufficiency or Predictive Rate Parity

For all o, y in {0,1}

$$P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)$$

- Y is independent of T given O

Impossibility Result

Impossibility Result

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive

Impossibility Result

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive
- Key insight: Often the label Y and variable T are correlated

Impossibility Result

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive
- Key insight: Often the label Y and variable T are correlated
  - Eg. Proportion of people in T who qualify might be lower than in the complement of T

Impossibility Result

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive
- Key insight: Often the label Y and variable T are correlated
  - Eg. Proportion of people in T who qualify might be lower than in the complement of T
- Demographic parity & Sufficiency => T and Y are independent

Impossibility Result

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive
- Key insight: Often the label Y and variable T are correlated
  - Eg. Proportion of people in T who qualify might be lower than in the complement of T
- Demographic parity & Sufficiency => T and Y are independent
- Demographic parity and Equal Odds=> if T depends on Y then O has to be independent of Y (outcome independent of label!)

Impossibility Result

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive

- Key insight: Often the label Y and variable T are correlated
  - Eg. Proportion of people in T who qualify might be lower than in the complement of T

- Demographic parity & Sufficiency => T and Y are independent

- Demographic parity and Equal Odds=> if T depends on Y then O has to be independent of Y (outcome independent of label!)

- Equal Odds & Sufficiency => T and Y are independent

Achieving Fairness

- Preprocessing: Eg. Demographic Parity
  - preprocess to remove information about T from input features X to create feature Z, use Z as new input
  - Example:
    - Find all directions in data matrix X that correlate with T
    - Remove these directions and let Z be the data matrix projected on remaining directions
    - If X is gaussian distributed this will make T and Z independent

Achieving Fairness

- While training: Find model that minimizes training error subject to fairness constraints

Eg. Fair K-means Clustering (very naive)

Eg. Fair K-means Clustering (very naive)

$$\text{Objective} = \sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \mathbf{r}_j \right\|_2^2$$

$$\text{where } \mathbf{r}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_t \in C_j} \mathbf{x}_t$$

Eg. Fair K-means Clustering (very naive)

$$\text{Objective} = \sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \mathbf{r}_j \right\|_2^2$$

$$\text{where } \mathbf{r}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_t \in C_j} \mathbf{x}_t$$

$$\text{Fairness constraints:} \quad \forall j \in [K], \qquad \sum_{t : c_t = j} \mathbf{1}_{x_t \in T} = \sum_{t : c_t = j} \mathbf{1}_{x_t \notin T}$$

Eg. Fair K-means Clustering (very naive)

$$\text{Objective} = \sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \mathbf{r}_j \right\|_2^2$$

$$\text{where } \mathbf{r}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_t \in C_j} \mathbf{x}_t$$

$$\text{Fairness constraints:} \quad \forall j \in [K], \qquad \sum_{t: c_t = j} \mathbf{1}_{x_t \in T} = \sum_{t: c_t = j} \mathbf{1}_{x_t \notin T}$$

Number of protected in cluster j = Number of unprotected in cluster j

Fair Classification

A view from a mile above:

Fair Classification

A view from a mile above:

Minimize Classification objective
(or whatever other surrogate loss you use usually)

Fair Classification

A view from a mile above:

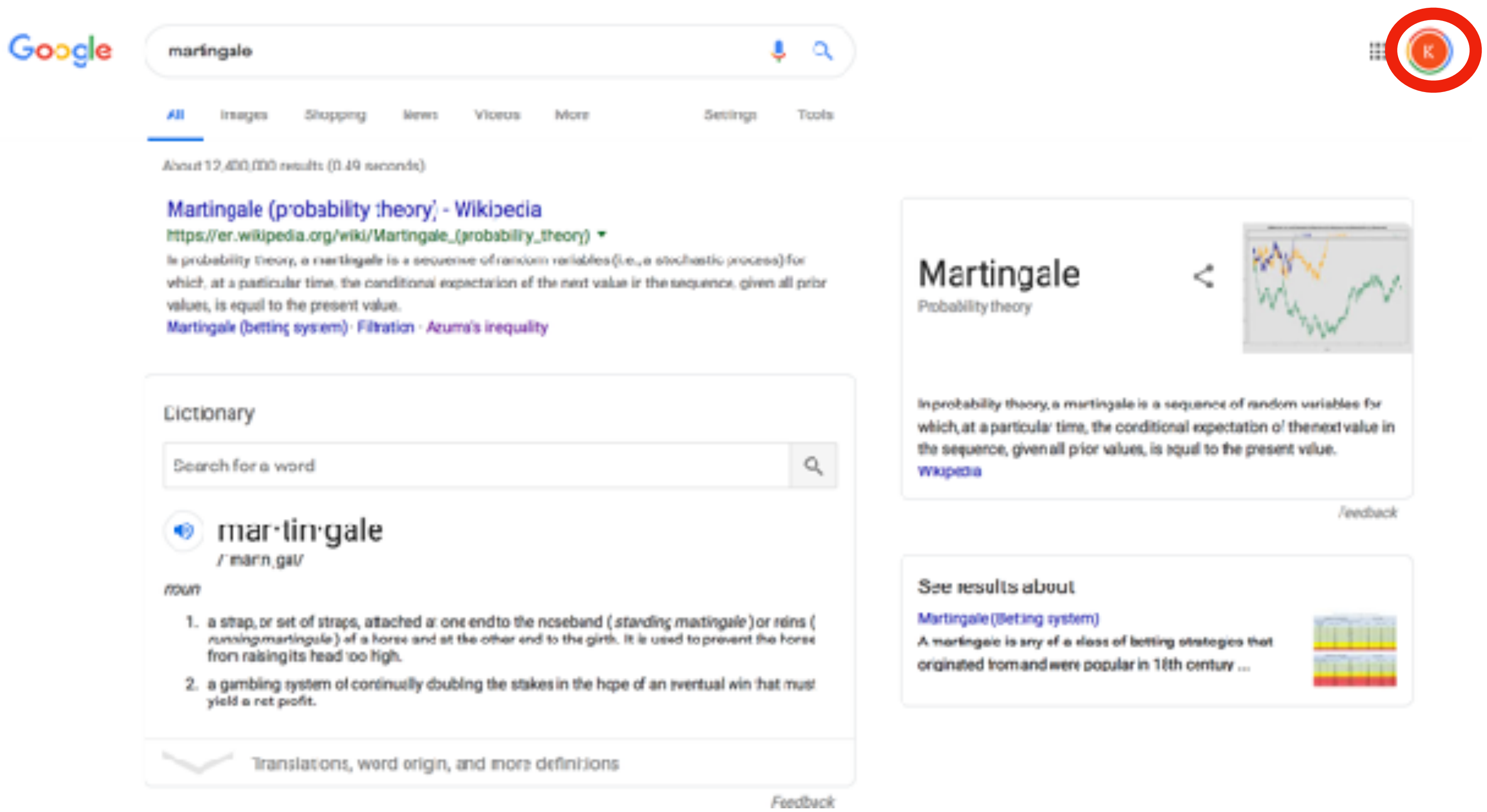
Minimize Classification objective
(or whatever other surrogate loss you use usually)

Added Constraint: subject to proportion of labels in each class being same for protected and unprotected population

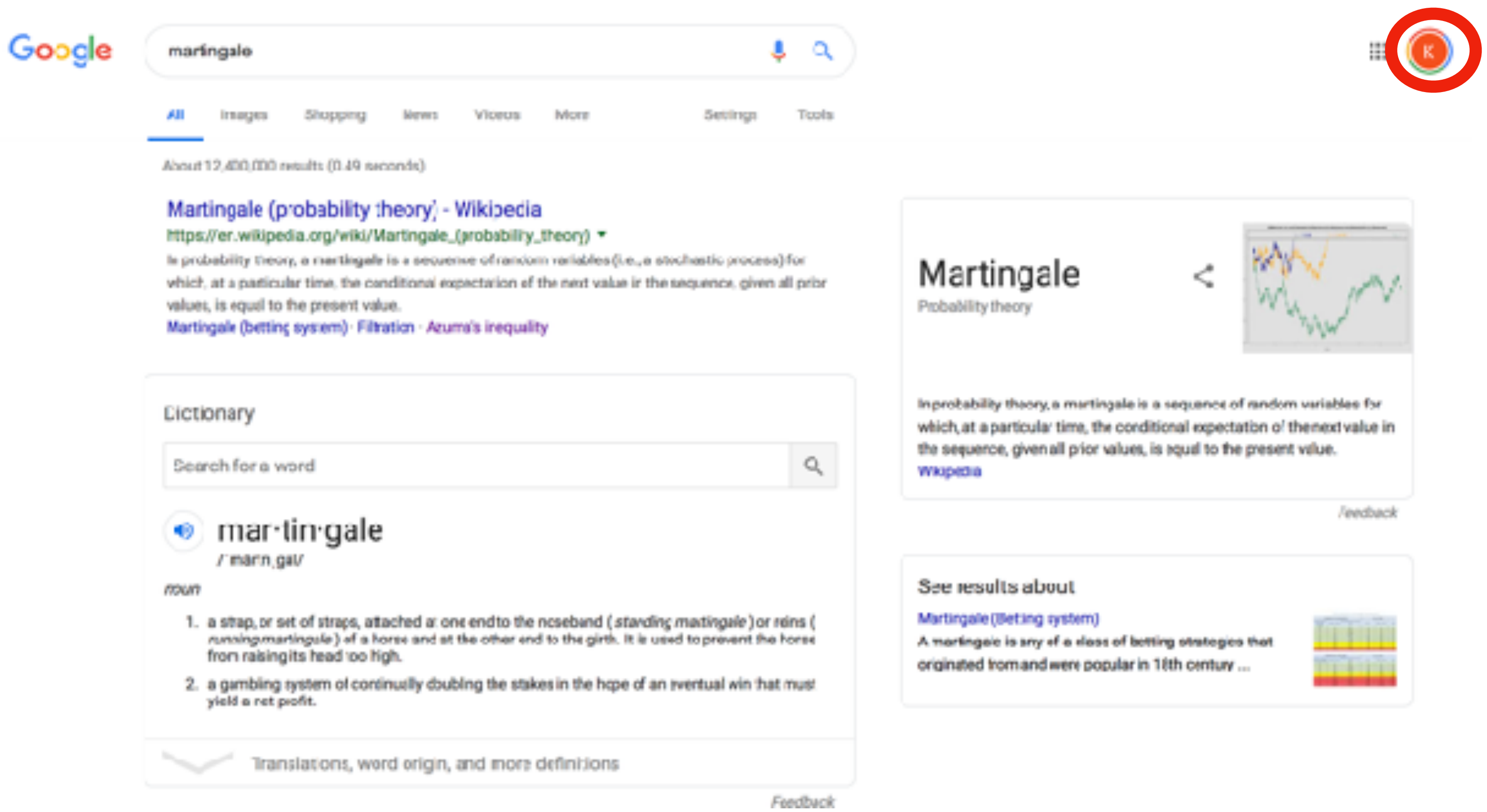
ACHIEVING FAIRNESS

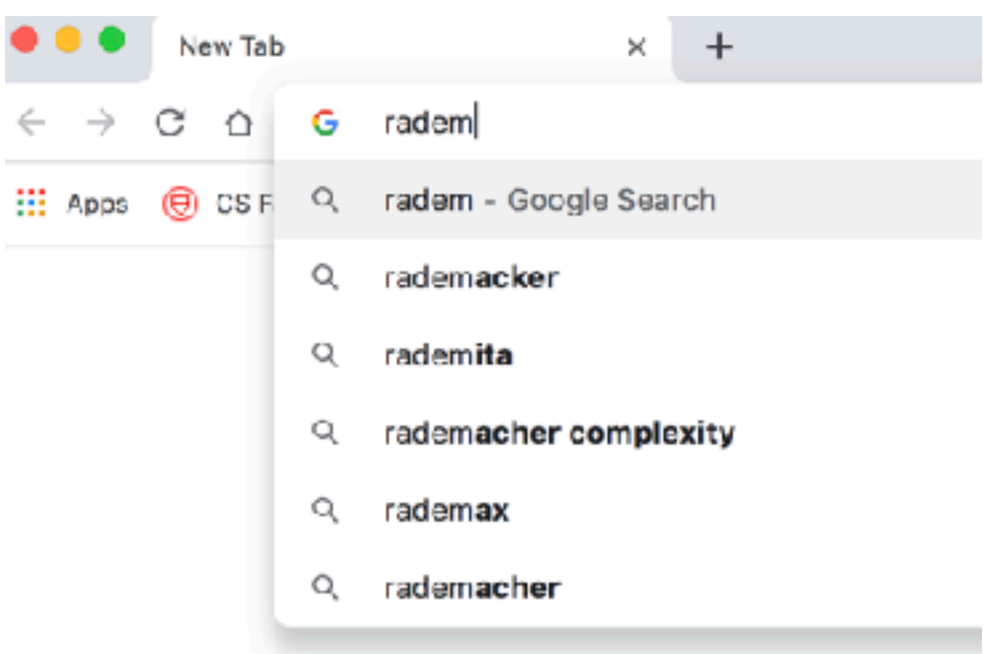
- Post-processing:
  - Learn model as before on training data,
  - As post processing use fresh training data to learn a bias parameter to correct for fairness
- Eg. Equal Odds (Binary classification)
  - Learn mapping f from training set such that from input to reals such that Y = 1 if f(X) >0 and Y = 0 if not
  - Now on fresh dataset, learn new threshold theta such that for protected class, Y = 1 if f(X) > theta and Y = 0 if not
  - Theta is chosen so as to ensure Equal odds

Personalization and Recommender Systems

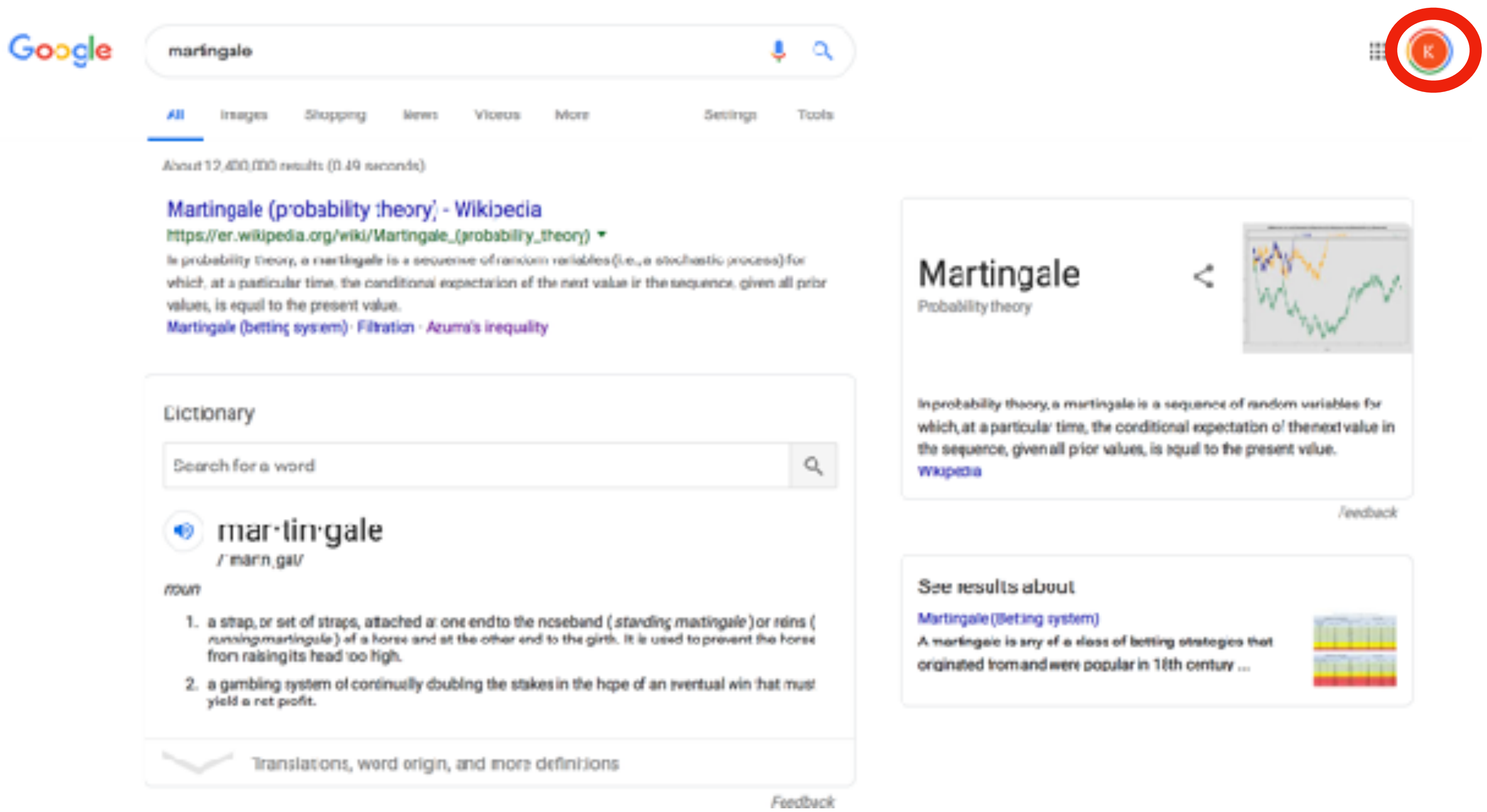


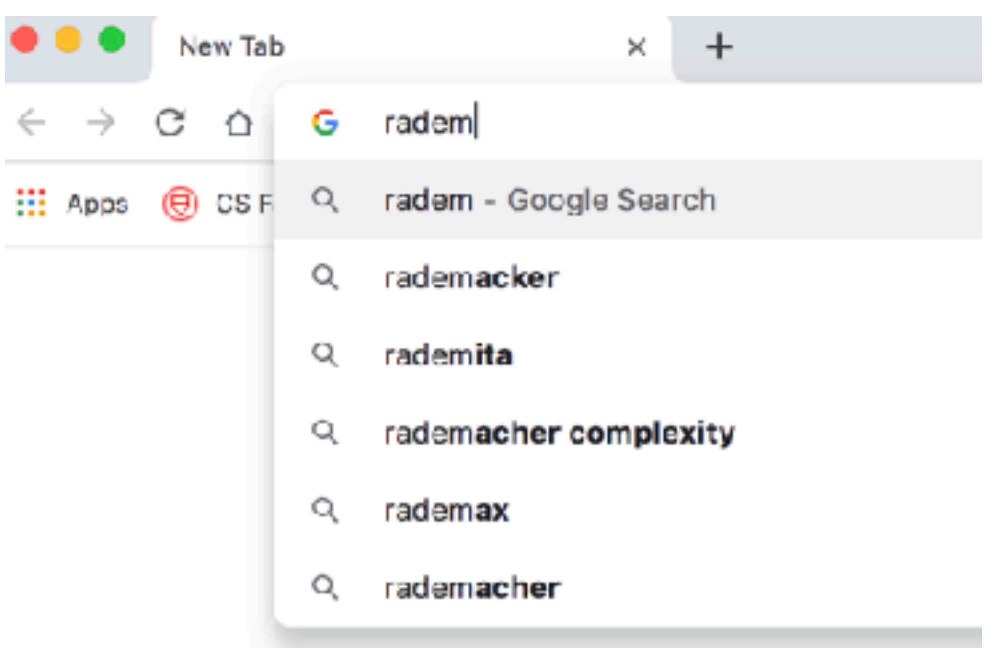
PERSONALIZATION AND RECOMMENDER SYSTEMS



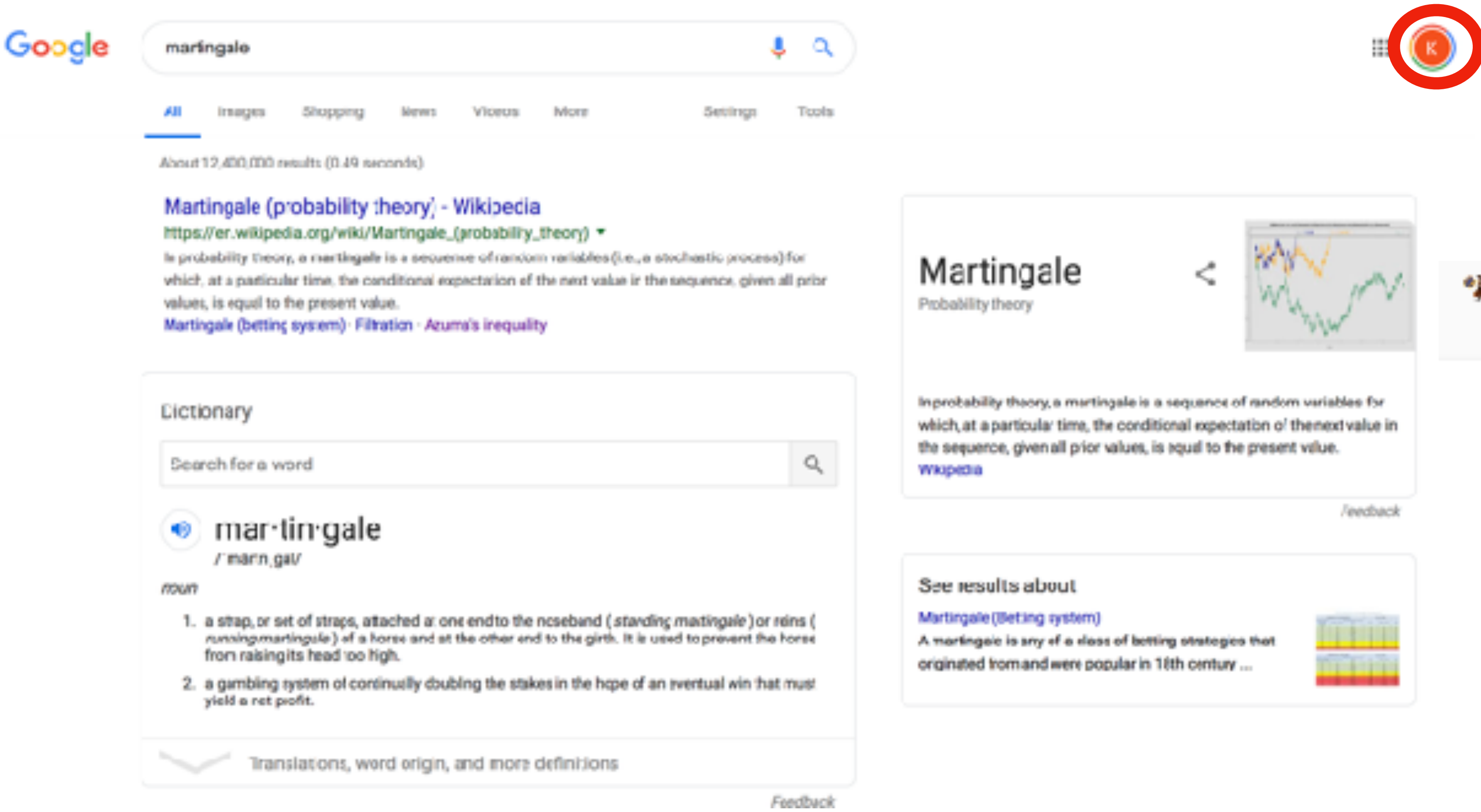


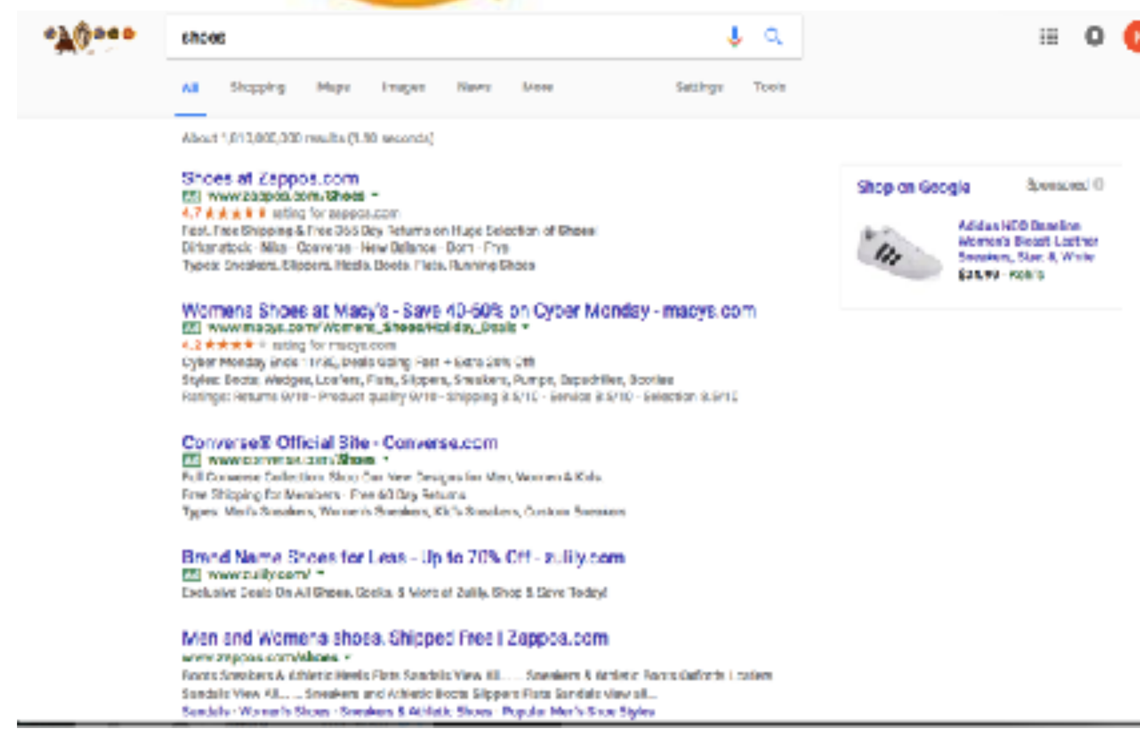
Personalization and Recommender Systems

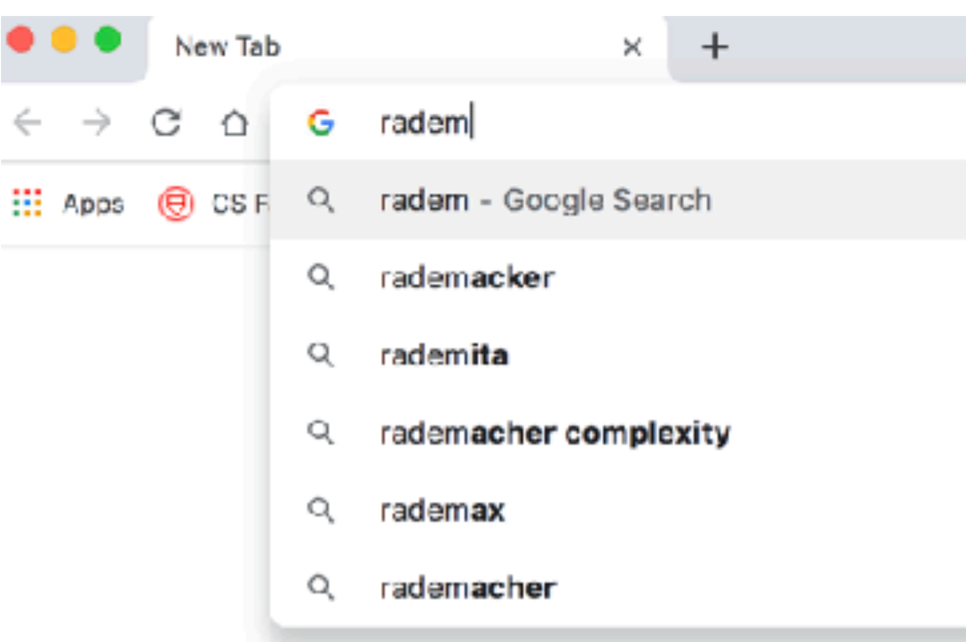




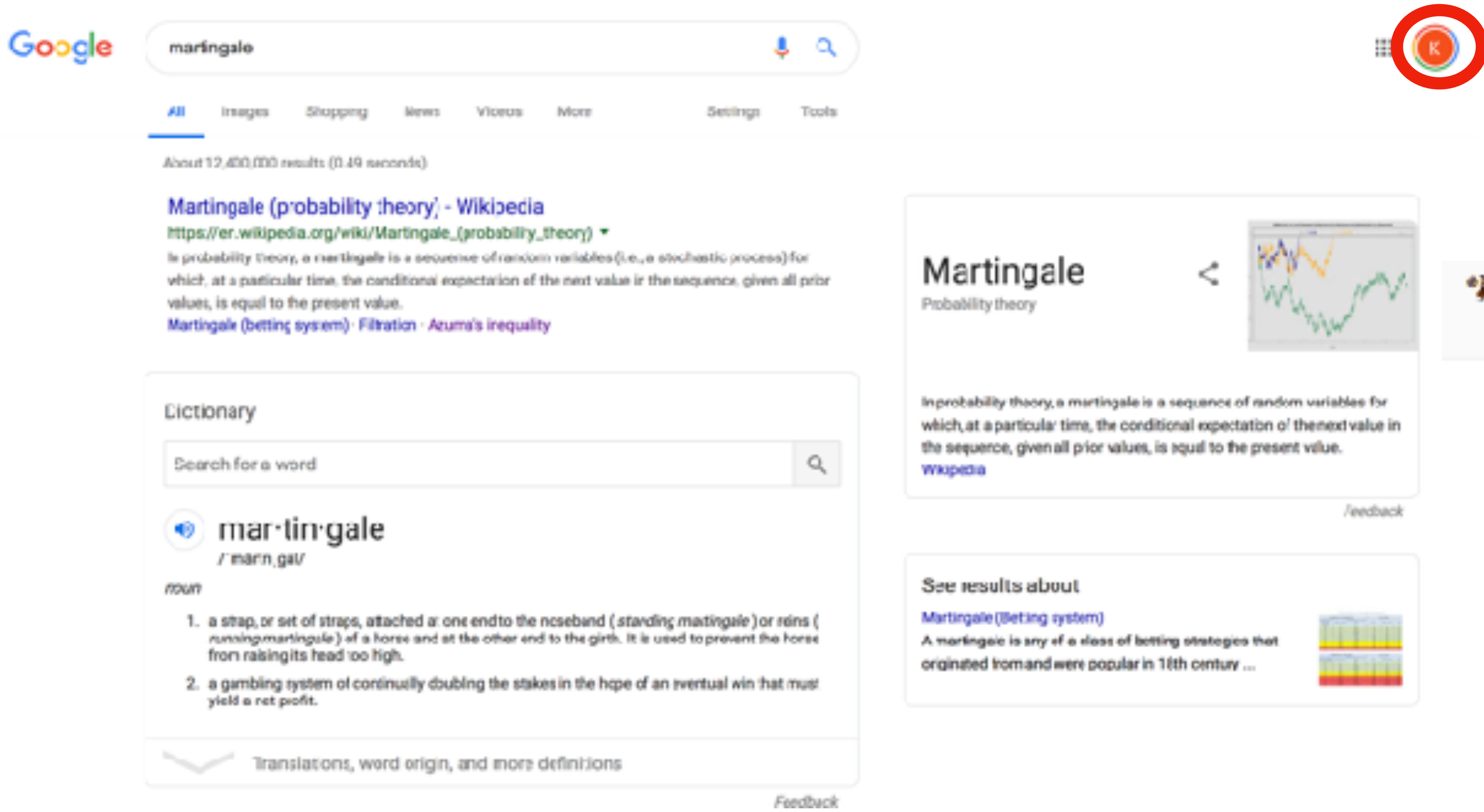
Personalization and Recommender Systems

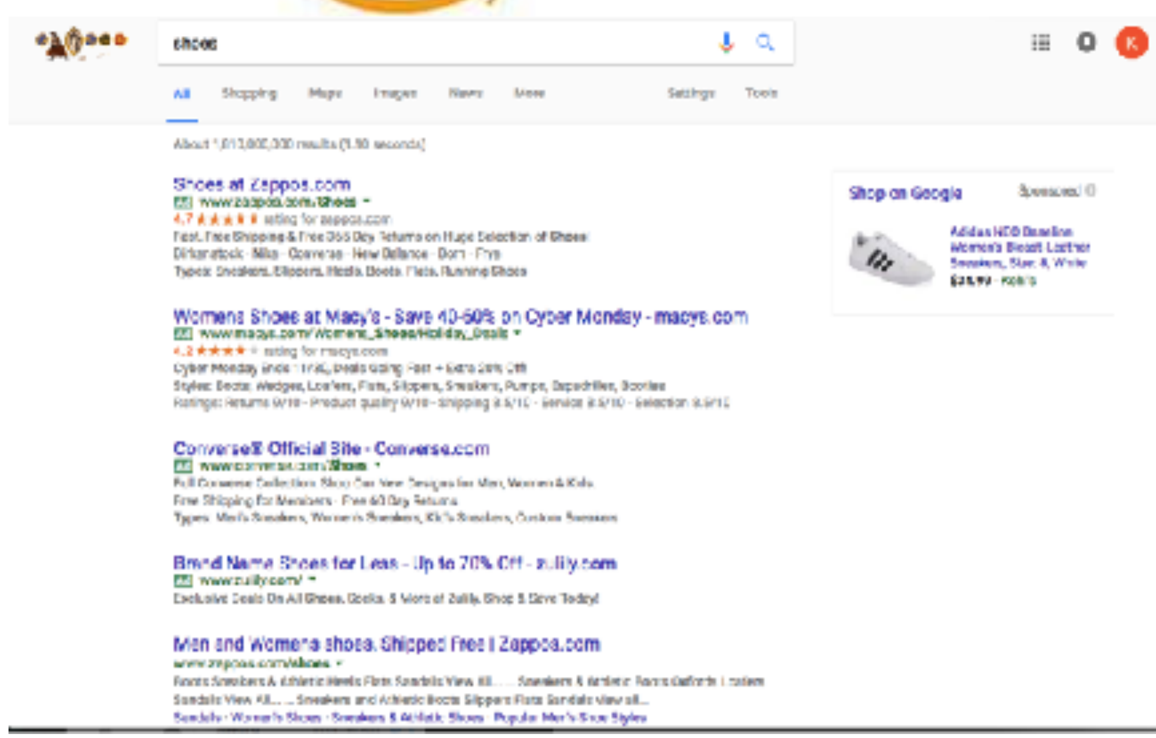


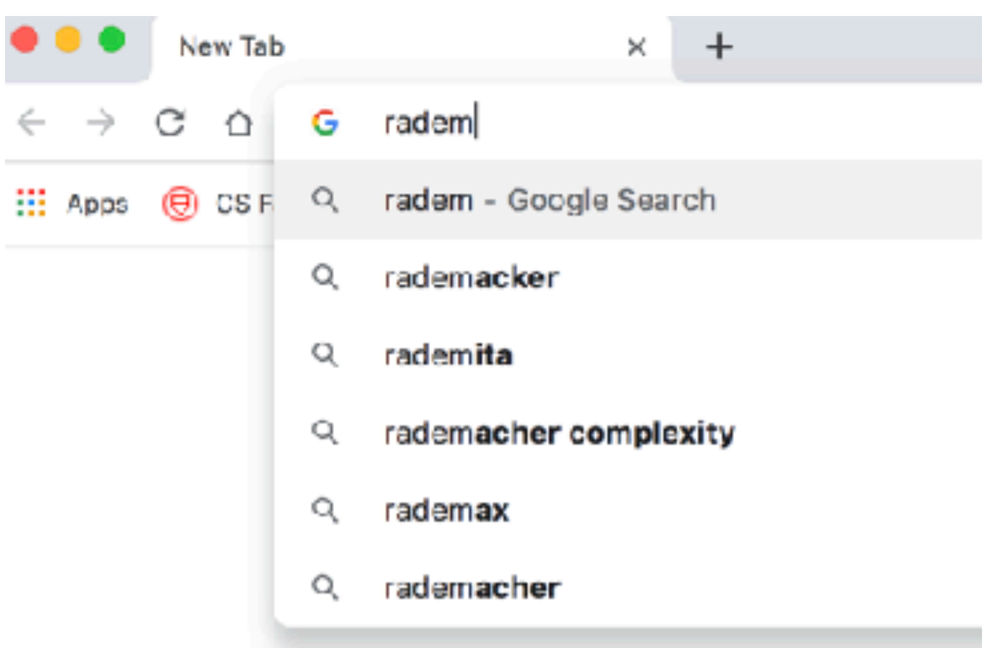




Personalization and Recommender Systems







Word of Caution!

With Big Data comes Bigger Responsibilities …

Word of Caution!

"That's what's happening with these Facebook pages where more and more people are getting their news from. At a certain point you just live in a bubble," he said. "And that's part of why our politics is so polarized right now. I think it is a solvable problem but it's one we have to spend a lot of time thinking about."

Word of Caution!

"That's what's happening with
these Facebook pages where
more and more people are

☰ Menu Search **Bloomberg** Sign In

Politics

Trump Says 'Do Something' About Alleged Social Media Bias

By Jennifer Jacobs
March 19, 2019, 2:31 PM EDT

LIVE ON BLOOMBERG
Watch Live TV >
Listen to Live Radio >

Word of Caution!

"That's what's happening with

Menu Search Bloomberg Sign In

Politics

Trump Says 'Do Something' About Alleged

The Atlantic Popular Latest Sections Magazine

POLITICS

Two Universes, One Report

The release of Robert Mueller's findings was a choose-your-own-adventure moment for political punditry.

On CNN, the headline from the attorney general's press conference gestured toward presidential malfeasance: **ag barr: mueller looked at "10 episodes" involving trump and obstruction.**

Fox News, meanwhile, declared **presidential vindication: ag barr: special counsel found no collusion.**

There is nothing new, of course, about the American media's descent into a choose-your-own-adventure dystopia of information bubbles and confirmation bias. But this week's coverage of the Mueller report stood out as a stark example of our fracturing media landscape—and the dysfunctional discourse it's produced.

WORD OF CAUTION!

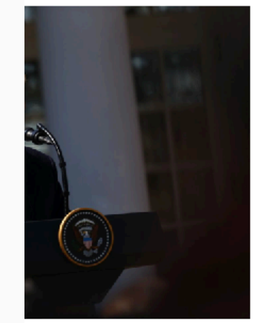
"That's what's happening with

Bloomberg

Politics

Trump Says 'Do Something' About Alleged

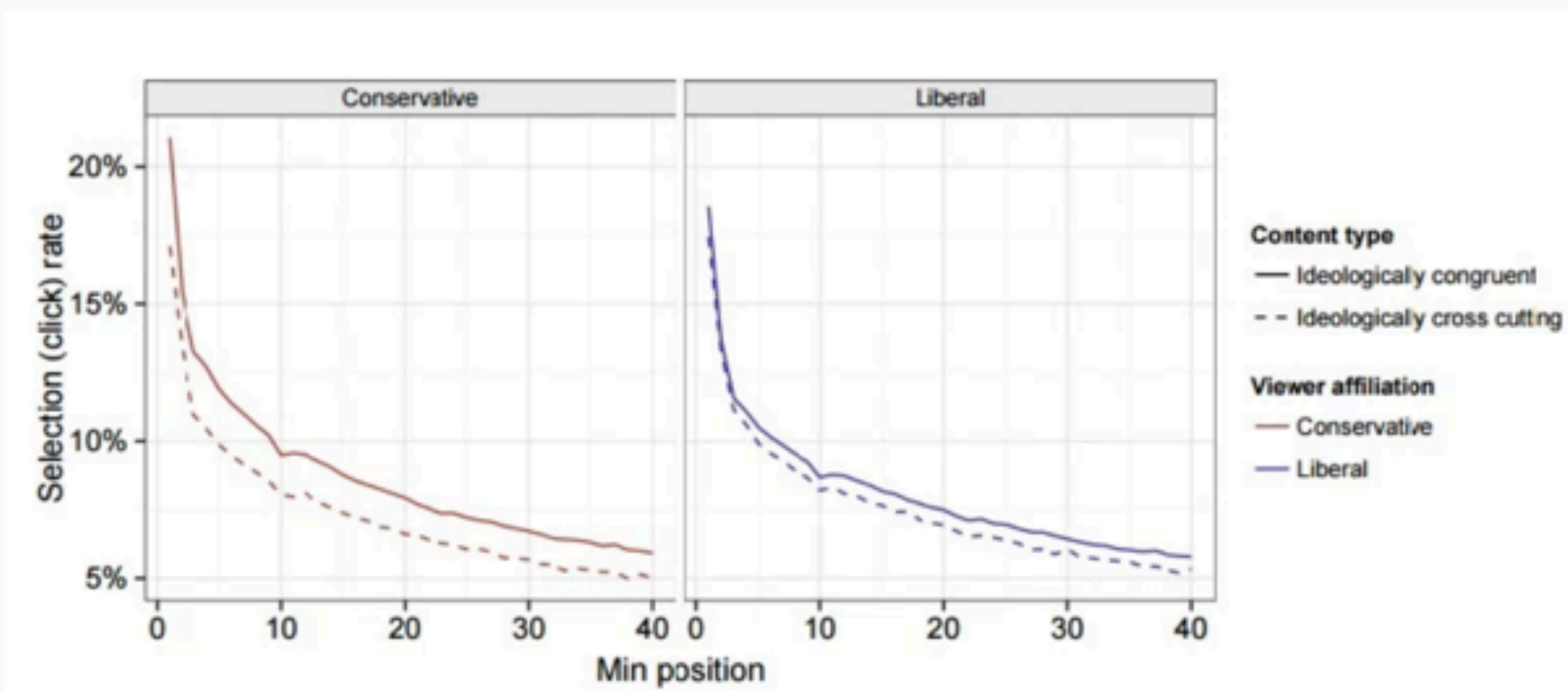
IS



https://www.brookings.edu/blog/techtank/2015/05/13/political-polarization-on-facebook/

The Facebook News feed does limit the amount of cross-cutting links that viewers choose to read. The News feed algorithm ranks stories based on a variety of factors including their history of clicking on links for particular websites. If a user regularly clicks on stories from sources with a partisan leaning then the chances of seeing a similar story increases. The News feed algorithm functions in this way to make the experience of using the website more enjoyable. This approach also has some unintended negative consequences. The authors find that the News feed algorithm reduces the politically cross-cutting content by 5 percent for conservatives and 8 percent for liberals.

THE FACEBOOK NEWS FEED ALGORITHM



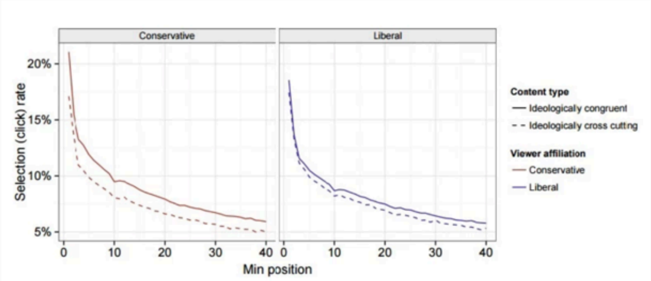
Word of Caution!

"That's what's happening with these Facebook pages where more and more people are getting their news from. At a certain point you just live in a bubble," he said. "And that's part of why our politics is so polarized right now. I think it is a solvable problem but it's one we have to spend a lot of time thinking about."

https://www.brookings.edu/blog/techtank/2015/05/13/political-polarization-on-facebook/

The Facebook News feed does limit the amount of cross-cutting links that viewers choose to read. The News feed algorithm ranks stories based on a variety of factors including their history of clicking on links for particular websites. If a user regularly clicks on stories from sources with a partisan leaning then the chances of seeing a similar story increases. The News feed algorithm functions in this way to make the experience of using the website more enjoyable. This approach also has some unintended negative consequences. The authors find that the News feed algorithm reduces the politically cross-cutting content by 5 percent for conservatives and 8 percent for liberals.

THE FACEBOOK NEWS FEED ALGORITHM



Menu Search **Bloomberg** Sign In

Politics

Trump Says 'Do Something' About Alleged Social Media Bias

By Jennifer Jacobs
March 19, 2019, 2:31 PM EDT

The Atlantic Popular Latest Sections Magazine

POLITICS

Two Universes, One Report

The release of Robert Mueller's findings was a choose-your-own-adventure moment for political punditry.

On CNN, the headline from the attorney general's press conference gestured toward presidential malfeasance: **ag barr: mueller looked at "10 episodes" involving trump and obstruction.**

Fox News, meanwhile, declared **presidential vindication: ag barr: special counsel found no collusion.**

There is nothing new, of course, about the American media's descent into a choose-your-own-adventure dystopia of information bubbles and confirmation bias. But this week's coverage of the Mueller report stood out as a stark example of our fracturing media landscape—and the dysfunctional discourse it's produced.

WORD OF CAUTION!

DeepMind is asking how AI helped turn the internet into an echo chamber

Researchers found that the more accurately a recommendation engine pegs your interests, the faster it traps you in an information bubble.

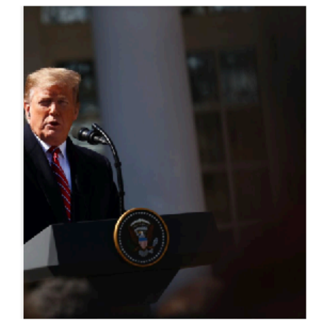
by **Karen Hao** Mar 7

One of the most common applications of machine learning today is in recommendation algorithms. Netflix and YouTube use them to push you new shows and videos; Google and Facebook use them to rank the content in your search results and news feed. While these algorithms offer a great deal of convenience, they have some undesirable side effects. You've probably heard of them before: filter bubbles and echo chambers.

out as a stark example of our fracturing media landscape—and the dysfunctional discourse it's produced.

Bloomberg Sign In

'Do Something' About Alleged a Bias



LIVE ON BLOOMBERG
Watch Live TV >
Listen to Live Radio >

Word of Caution!

Word of Caution!

Word of Caution!

User 1

User 2

Word of Caution!

User 1

Apples are extremely rich in important antioxidants, flavanoids, and dietary fiber. The phytonutrients and antioxidants in **apples** may help reduce the risk of developing cancer, hypertension, diabetes, and heart

User 2

For fewer calories per fruit, **oranges** have higher levels of Vitamin C, folate, potassium, and protein.

Word of Caution!

User 1

Apples are extremely rich in important antioxidants, flavanoids, and dietary fiber. The phytonutrients and antioxidants in **apples** may help reduce the risk of developing cancer, hypertension, diabetes, and heart

User 2

For fewer calories per fruit, **oranges** have higher levels of Vitamin C, folate, potassium, and protein.

WORD OF CAUTION!

User 1

Apples are extremely rich in important antioxidants, flavanoids, and dietary fiber. The phytonutrients and antioxidants in **apples** may help reduce the risk of developing cancer, hypertension, diabetes, and heart

User 2

For fewer calories per fruit, **oranges** have higher levels of Vitamin C, folate, potassium, and protein.

Word of Caution!

Vs

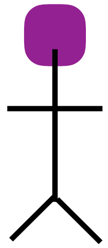
User 1

Apples are extremely rich in important antioxidants, flavanoids, and dietary fiber. The phytonutrients and antioxidants in **apples** may help reduce the risk of developing cancer, hypertension, diabetes, and heart

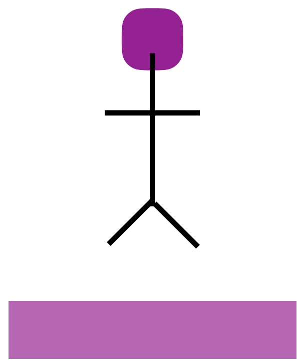
User 2

For fewer calories per fruit, **oranges** have higher levels of Vitamin C, folate, potassium, and protein.

Opinion Formation Model

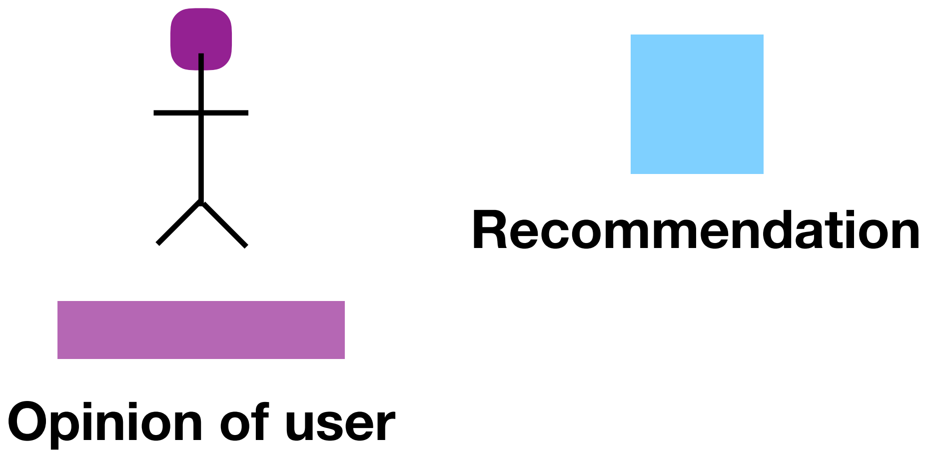


Opinion Formation Model

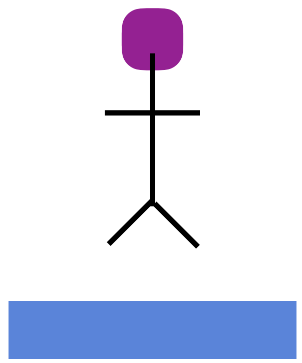


Opinion of user

Opinion Formation Model



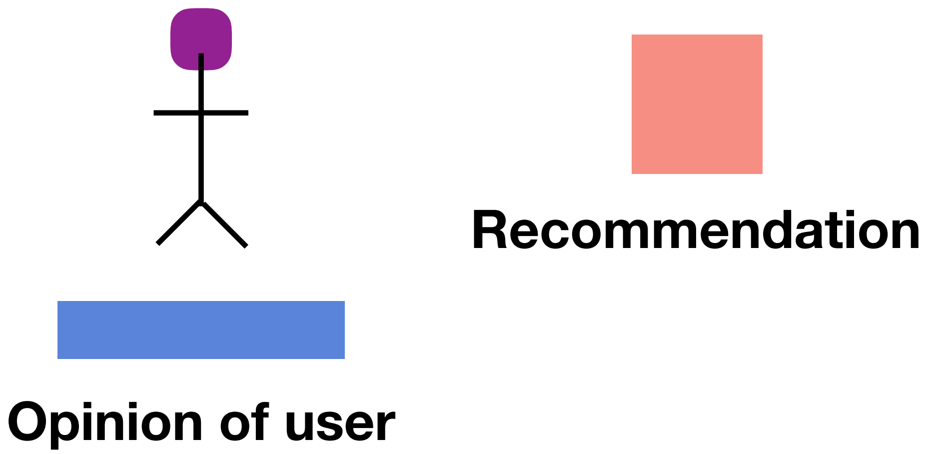
Opinion Formation Model



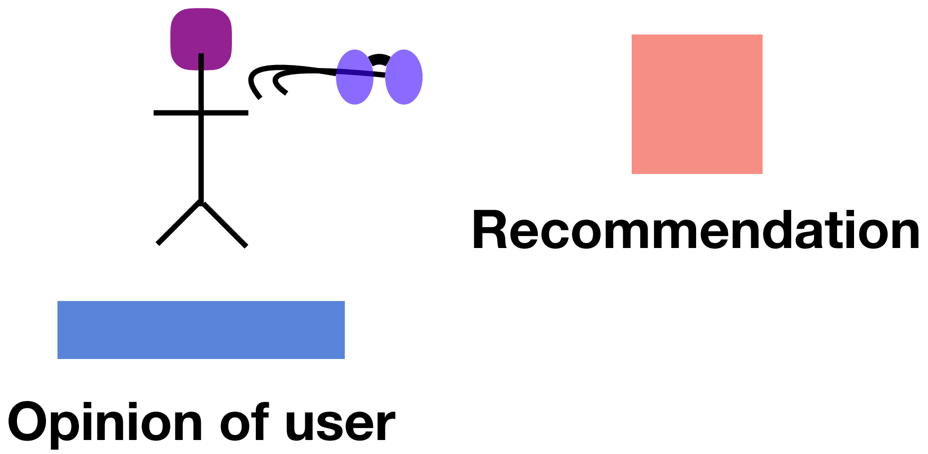
Recommendation

Opinion of user

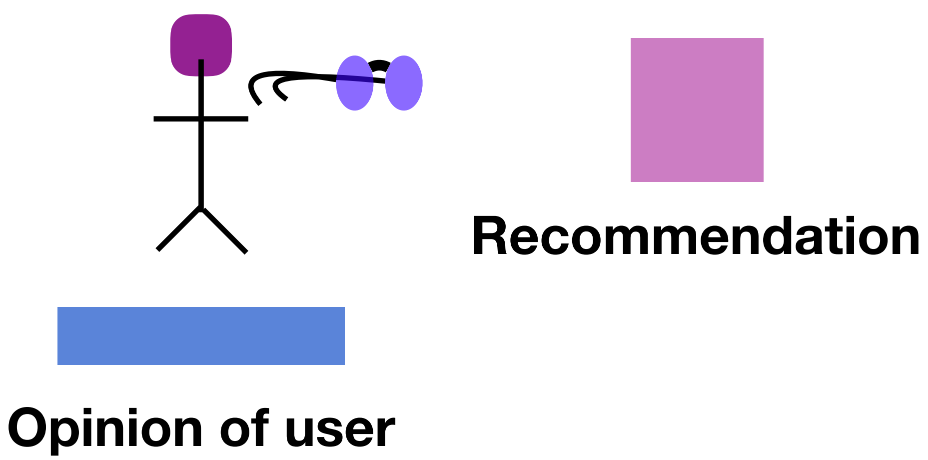
Opinion Formation Model



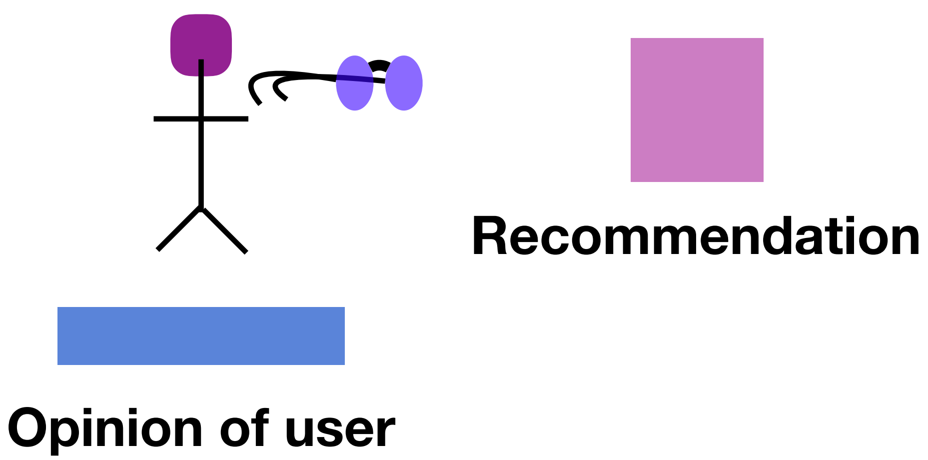
Opinion Formation Model



Opinion Formation Model

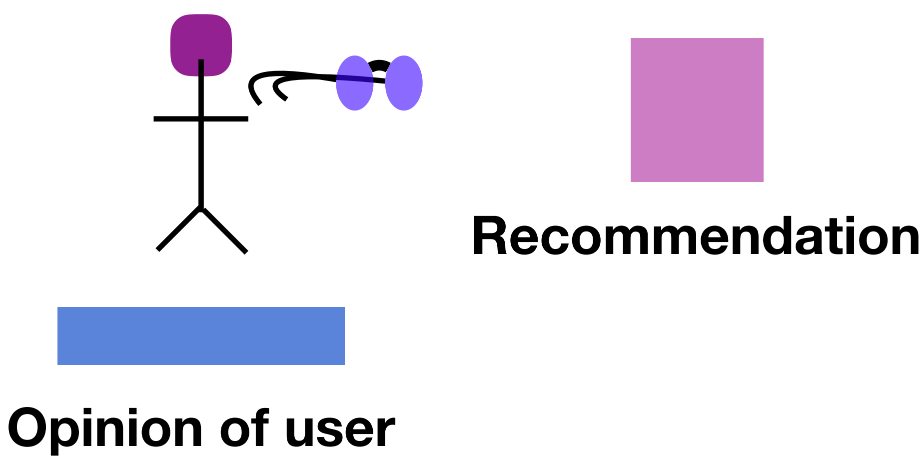


Opinion Formation Model



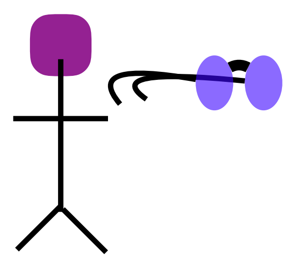
With confirmation bias, recommendations have to be neutral

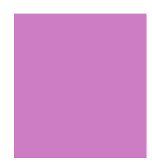
Opinion Formation Model



Neutral articles might be boring! What if user had friends to talk to?

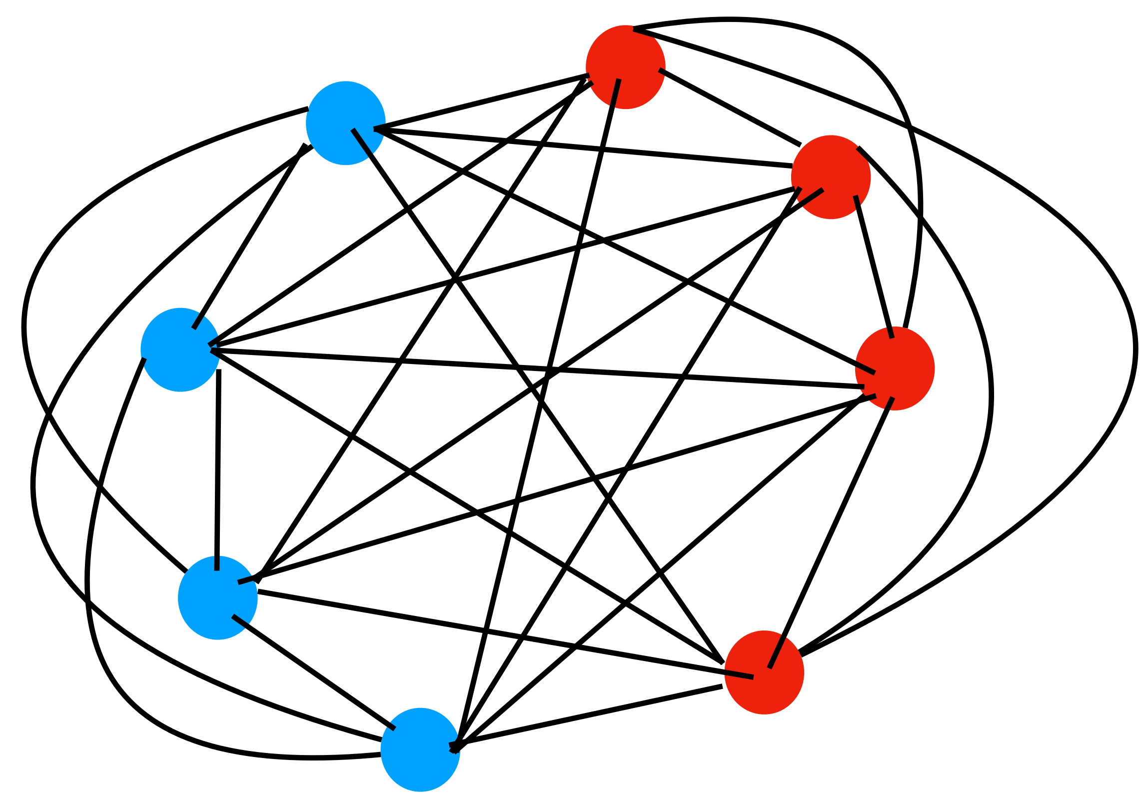
Opinion Formation Model



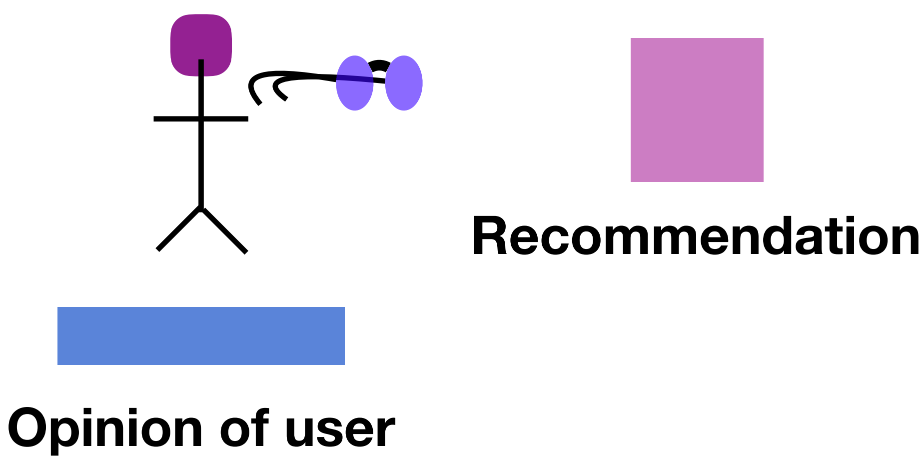


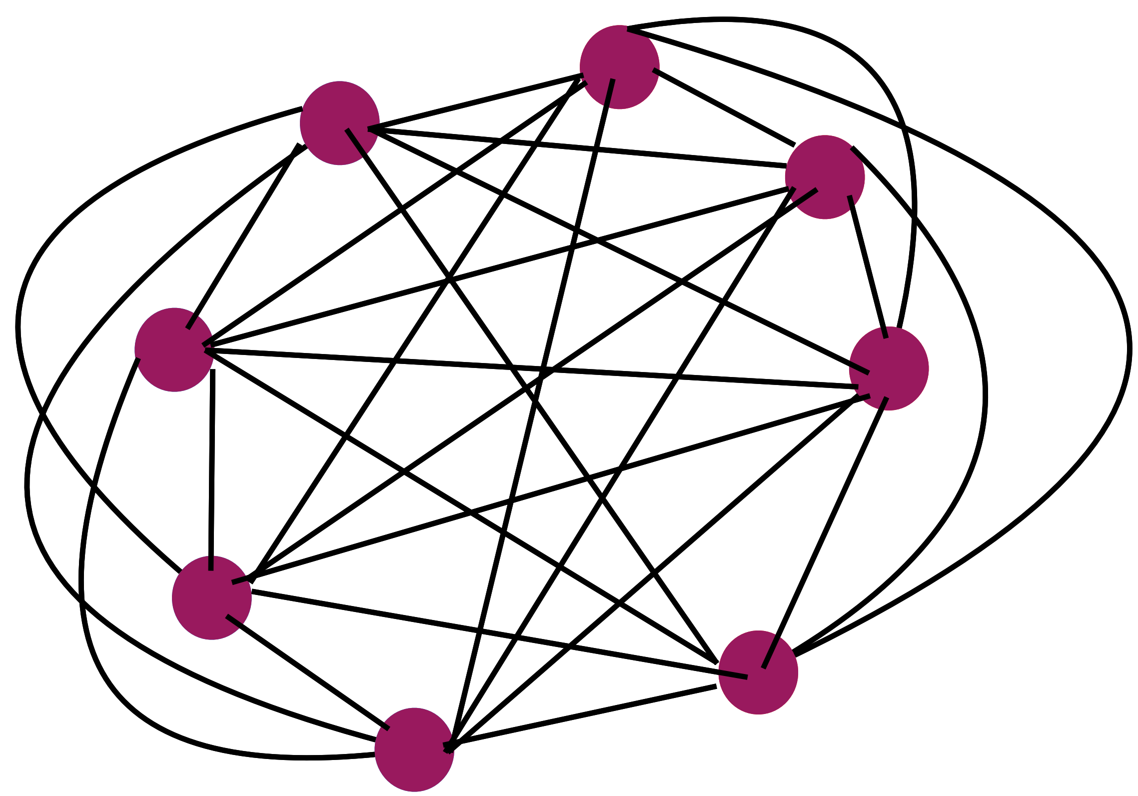
Recommendation

Opinion of user

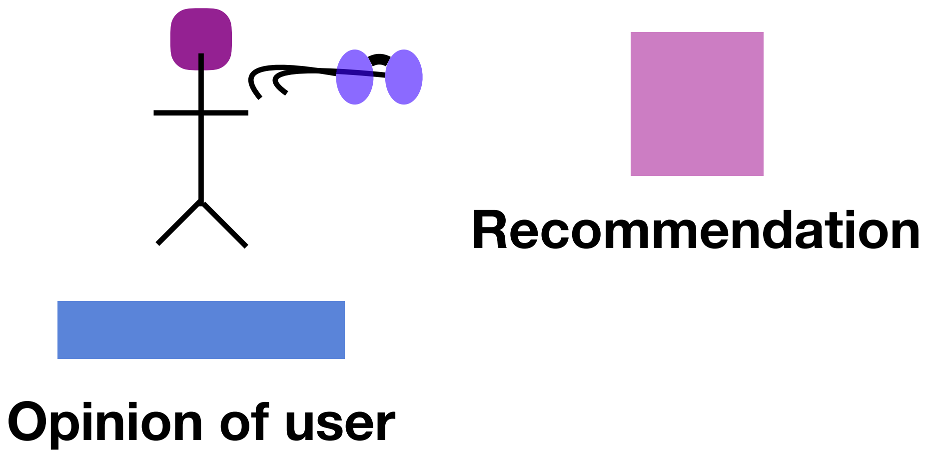


Opinion Formation Model

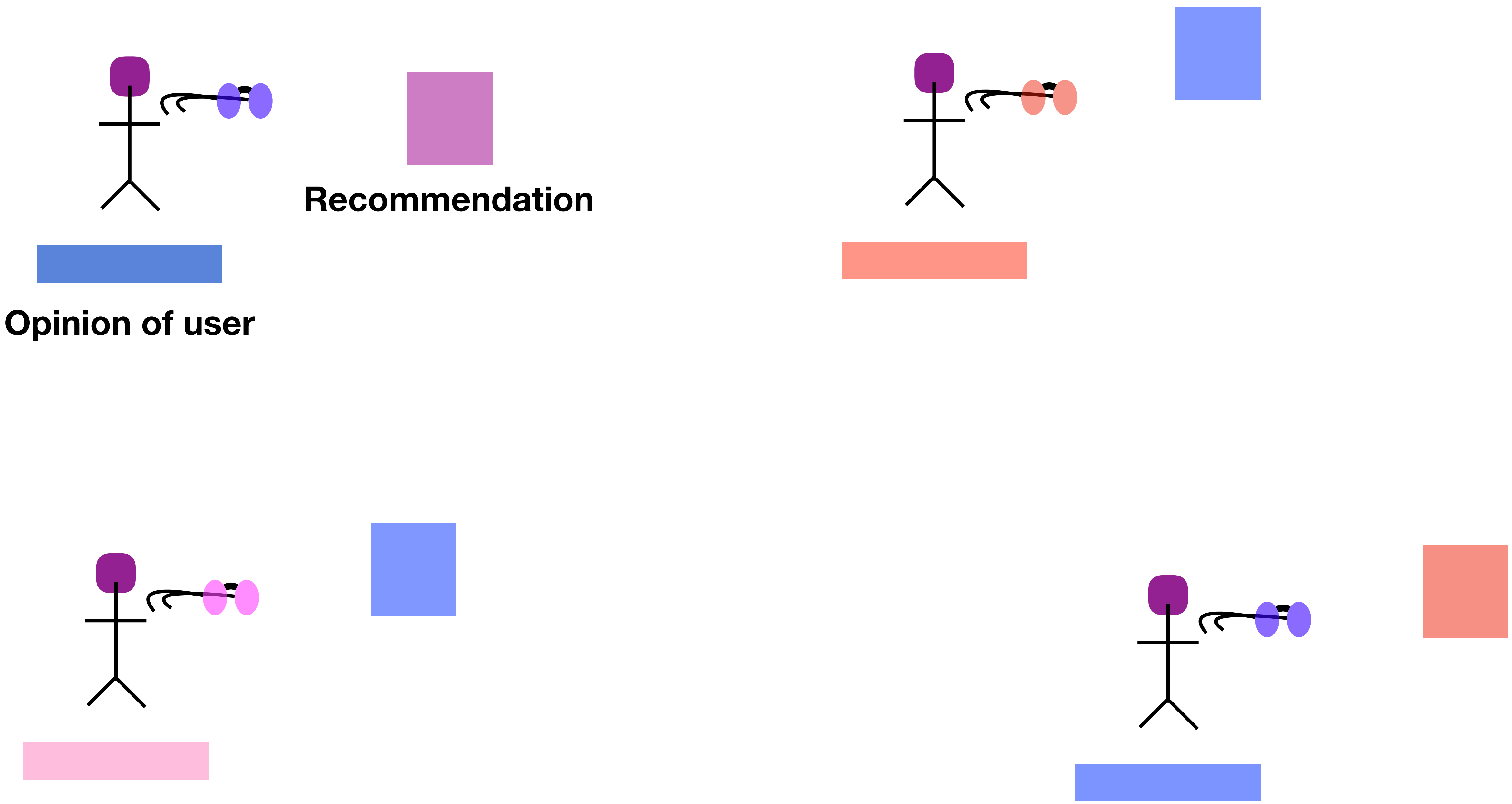




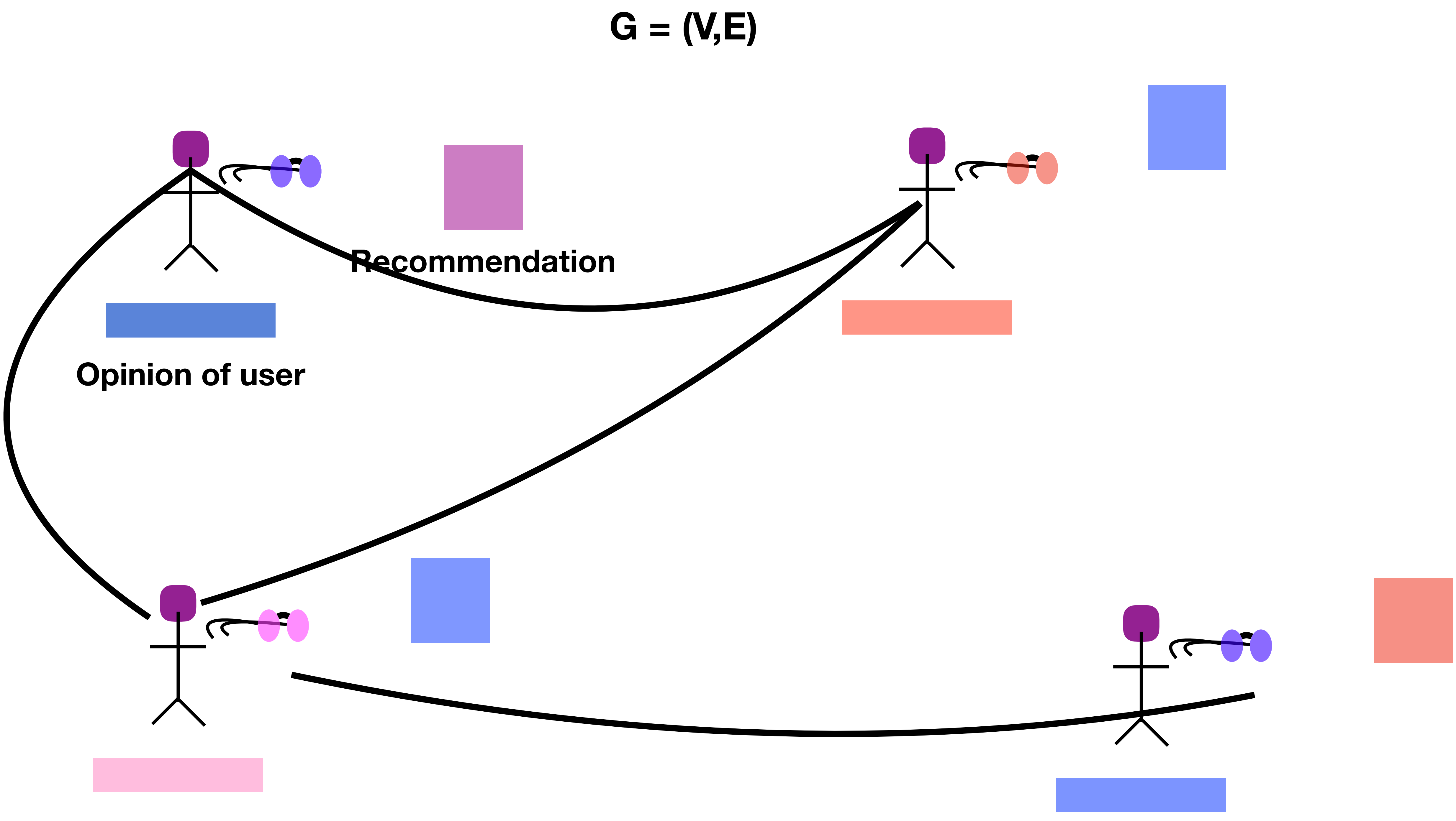
Opinion Formation Model



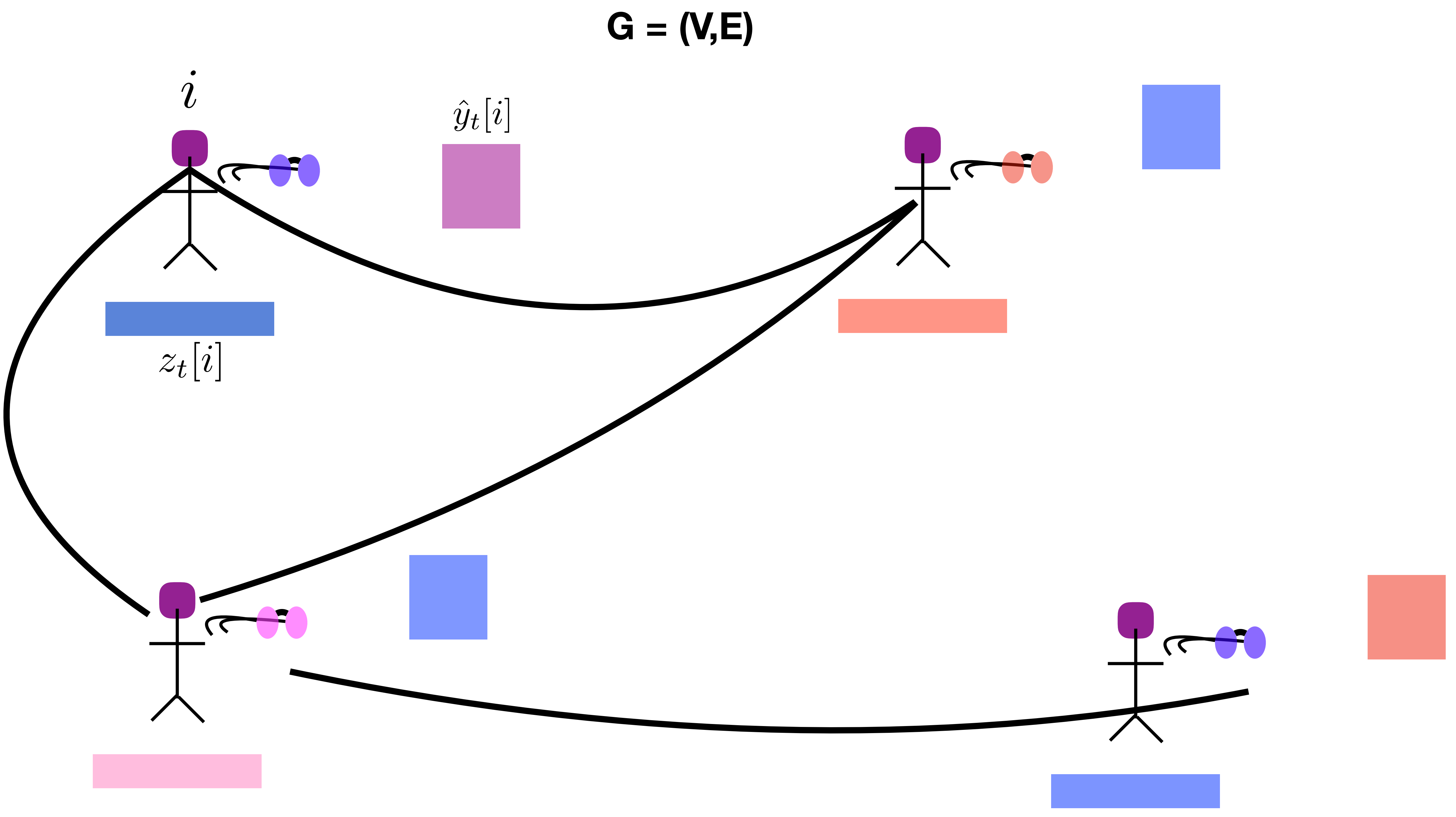
Opinion Formation Model



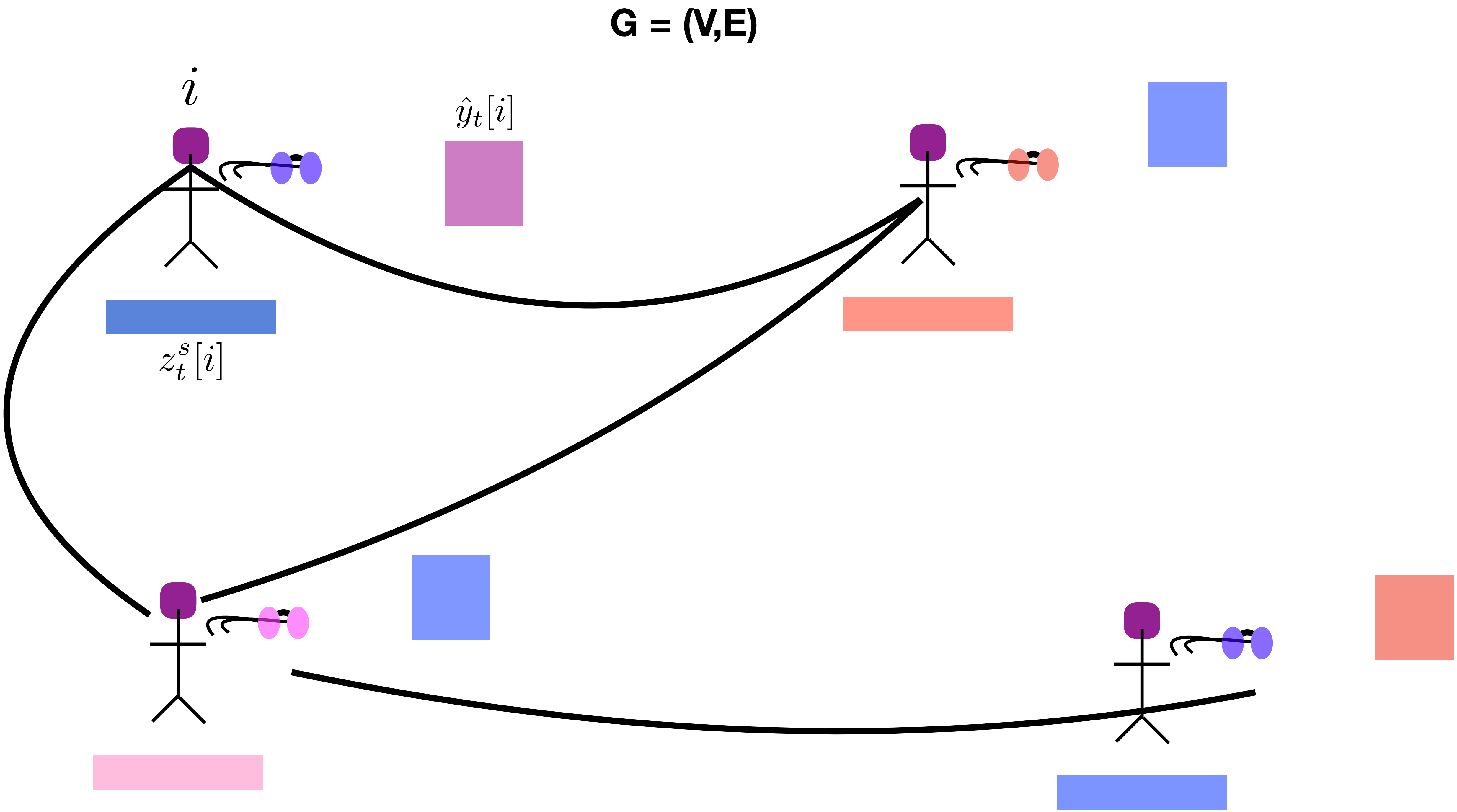
Opinion Formation Model



Opinion Formation Model



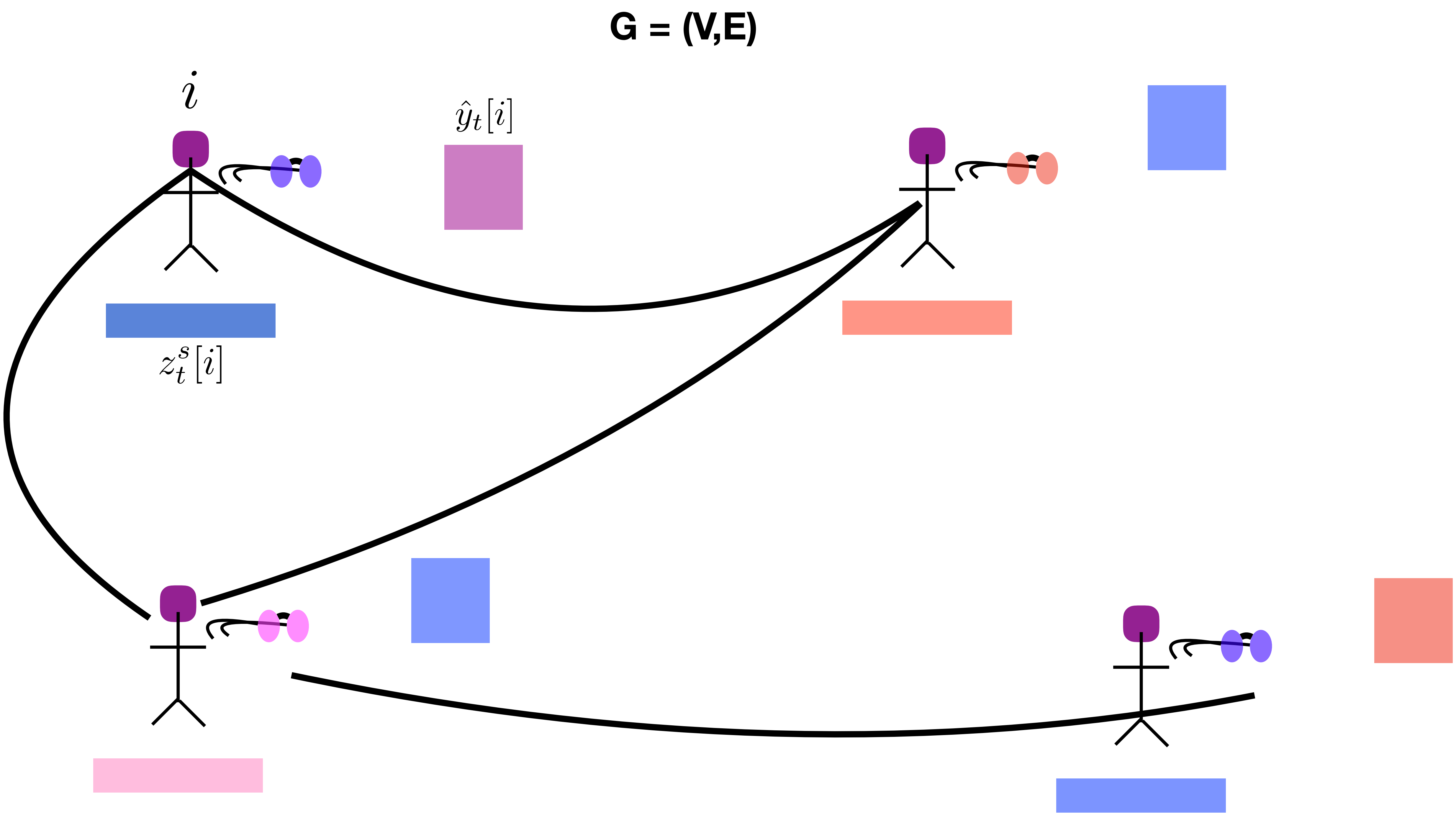
Opinion Formation Model



Opinion Formation Model

Assuming users start at neutral,

Measure of polarization for day t : $\|z_t\|_\infty$



Opinion Formation Model: With Friends

- On each day we model users as interacting with each other for multiple rounds. We will denote by vector z_t^s the opinion of the n users on day t after s'th round of interaction.

Opinion Formation Model: With Friends

- On each day we model users as interacting with each other for multiple rounds. We will denote by vector z_t^s the opinion of the n users on day t after s'th round of interaction.

- Beginning of day t provide recommendations $\hat{y}_t$:

$$z_t^0 = \beta_t z_{t-1} + (1 - \beta_t)\mathrm{diag}(\sigma(z_{t-1} \odot \hat{y}_t))\hat{y}_t$$

where $z_t = \lim_{s \to \infty} z_t^s$ and $\sigma : [-1, 1] \mapsto [0, 1]$ is increasing.

Opinion Formation Model: With Friends

- On each day we model users as interacting with each other for multiple rounds. We will denote by vector z_t^s the opinion of the n users on day t after s'th round of interaction.

- Beginning of day t provide recommendations $\hat{y}_t$:

$$z_t^0 = \beta_t z_{t-1} + (1 - \beta_t)\text{diag}(\sigma(z_{t-1} \odot \hat{y}_t))\hat{y}_t$$

where $z_t = \lim_{s \to \infty} z_t^s$ and $\sigma : [-1, 1] \mapsto [0, 1]$ is increasing.

Opinion Formation Model: With Friends

- On each day we model users as interacting with each other for multiple rounds. We will denote by vector z_t^s the opinion of the n users on day t after s'th round of interaction.

- Beginning of day t provide recommendations $\hat{y}_t$:

$$z_t^0 = \beta_t z_{t-1} + (1-\beta_t)\operatorname{diag}(\sigma(z_{t-1} \odot \hat{y}_t))\hat{y}_t$$

where $z_t = \lim_{s \to \infty} z_t^s$ and $\sigma : [-1,1] \mapsto [0,1]$ is increasing.

- Opinions through the day in multiple rounds evolve as:

$$z_t^s \leftarrow \alpha(I+D)^{-1}(I+A)z_t^{s-1} + (1-\alpha)\operatorname{diag}(\sigma(z_t^{s-1} \odot z_t^0))z_t^0$$

Opinion Formation Model: With Friends

- On each day we model users as interacting with each other for multiple rounds. We will denote by vector z_t^s the opinion of the n users on day t after s'th round of interaction.

- Beginning of day t provide recommendations $\hat{y}_t$:

$$z_t^0 = \beta_t z_{t-1} + (1 - \beta_t)\mathrm{diag}(\sigma(z_{t-1} \odot \hat{y}_t))\hat{y}_t$$

where $z_t = \lim_{s \to \infty} z_t^s$ and $\sigma : [-1, 1] \mapsto [0, 1]$ is increasing.

- Opinions through the day in multiple rounds evolve as:

$$z_t^s \leftarrow \alpha(I + D)^{-1}(I + A)z_t^{s-1} + (1 - \alpha)\mathrm{diag}(\sigma(z_t^{s-1} \odot z_t^0))z_t^0$$

- Without confirmation biases, its the Freidkin-Johensen model

Possible Fix

- As long as we keep the users neutral in the end of everyday, their confirmation bias wont be too bad

Possible Fix

- As long as we keep the users neutral in the end of everyday, their confirmation bias wont be too bad
- If users were neutral in the start of a day, then we can think about their opinions by end of the day via a random walk view

Possible Fix

- As long as we keep the users neutral in the end of everyday, their confirmation bias wont be too bad
- If users were neutral in the start of a day, then we can think about their opinions by end of the day via a random walk view
- Imagine every user with some probability α stops sharing and with remaining probability shares their opinion to all their neighbors

Possible Fix

- As long as we keep the users neutral in the end of everyday, their confirmation bias wont be too bad
- If users were neutral in the start of a day, then we can think about their opinions by end of the day via a random walk view
- Imagine every user with some probability α stops sharing and with remaining probability shares their opinion to all their neighbors
- Each person takes in average of all the opinions they receive

Possible Fix

- As long as we keep the users neutral in the end of everyday, their confirmation bias wont be too bad
- If users were neutral in the start of a day, then we can think about their opinions by end of the day via a random walk view
- Imagine every user with some probability α stops sharing and with remaining probability shares their opinion to all their neighbors
- Each person takes in average of all the opinions they receive
- It turns out that the final opinion of all the users converge to $M\hat{y}$ where $M = \left(I + D + \frac{1-\alpha}{\alpha} L\right)^{-1} (I + D)$ (is a stochastic matrix)

Possible Fix

- As long as we keep the users neutral in the end of everyday, their confirmation bias wont be too bad
- If users were neutral in the start of a day, then we can think about their opinions by end of the day via a random walk view
- Imagine every user with some probability α stops sharing and with remaining probability shares their opinion to all their neighbors
- Each person takes in average of all the opinions they receive
- It turns out that the final opinion of all the users converge to $M\hat{y}$ where $M = \left(I + D + \frac{1-\alpha}{\alpha} L\right)^{-1} (I + D)$ (is a stochastic matrix)
- Key idea: Pick articles such that $M\hat{y}$ is small on every coordinate

Synergy of Ideas

- Opinion formation and opinion dynamic models are studied extensively in social psychology

Synergy of Ideas

- Opinion formation and opinion dynamic models are studied extensively in social psychology
- We need to combine ideas from this field with ML techniques to deal with such issues

Synergy of Ideas

- Opinion formation and opinion dynamic models are studied extensively in social psychology
- We need to combine ideas from this field with ML techniques to deal with such issues
- The idea in previous slide can we shown to work provably only in limited scenarios

Synergy of Ideas

- Opinion formation and opinion dynamic models are studied extensively in social psychology
- We need to combine ideas from this field with ML techniques to deal with such issues
- The idea in previous slide can we shown to work provably only in limited scenarios
- Topic needs more research