

# Final Oral Exam

## Relational Learning and Fairness

Guy W. Cole



November 26, 2019

# Outline

1. Topic Blockmodel

2. Monotonic Fairness

3. Elicited Monotonic Fairness

# Topic Blockmodel: Introduction

- *Problem:* Lots of communications observed over network connections, how do we improve inference about the people based on the communications?

# Topic Blockmodel: Introduction

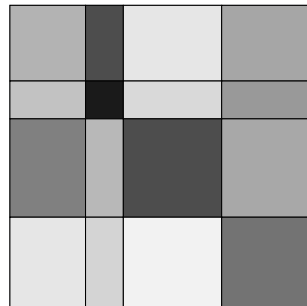
- *Problem*: Lots of communications observed over network connections, how do we improve inference about the people based on the communications?
- *Intuition*: Individuals in one homogeneous group should behave similarly when interacting with individuals of another homogeneous group.

# Topic Blockmodel: Introduction

- *Problem:* Lots of communications observed over network connections, how do we improve inference about the people based on the communications?
- *Intuition:* Individuals in one homogeneous group should behave similarly when interacting with individuals of another homogeneous group.
- *Model:* Use a stochastic blockmodel to assign individuals to communities, and a topic model for the communications from each community to each other.

# Background: Stochastic Blockmodels

- Boolean Stochastic Blockmodel:<sup>a</sup>
  - Each of  $n$  individuals are assigned to one of  $K$  communities
  - Each pair of communities  $(i, j)$  has an edge probability  $P_{ij}$ , probability that node in  $i$  has edge with any node in  $j$ .
- Can move from boolean to other distributions...



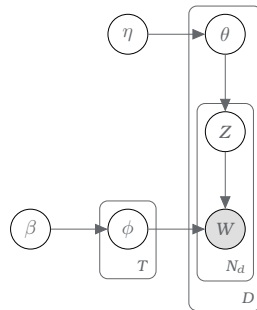
---

<sup>a</sup>Y. Wang and Wong 1987; Snijders and Nowicki 1997.

# Background: Latent Dirichlet Allocation

Topic Modeling:<sup>a</sup> Each document has a topic distribution  $\theta$  from which each topic  $z$  is drawn, and each word  $w$  is drawn from the word distribution  $\phi$  for its corresponding topic.

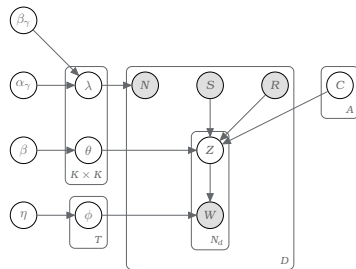
$$\begin{aligned}\theta_d &\sim \text{Dirichlet}(\eta) \\ \phi_t &\sim \text{Dirichlet}(\beta) \\ z_{dn} &\sim \text{Multinomial}(\theta_d) \\ w_{dn} &\sim \text{Multinomial}(\phi_{(z_{dn})})\end{aligned}$$



<sup>a</sup>Blei, Ng, and Jordan 2003.

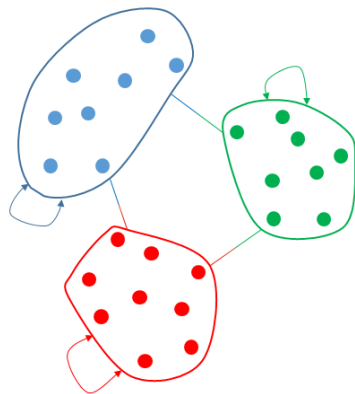
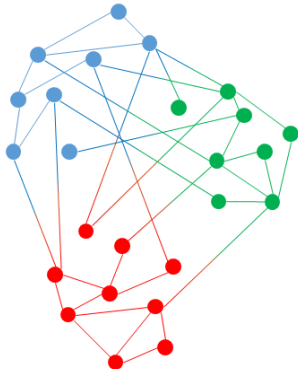
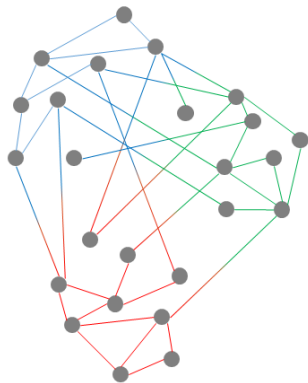
# Topic Blockmodel

- Draw community memberships:  $Z \sim \text{CRP}(\alpha)$
- Draw expected of word counts  
 $\lambda_{ij} \sim \text{Gamma}(\alpha_\gamma, \beta_\gamma)$
- Draw topic distributions:  $\theta_{ij} \sim \text{Dirichlet}(\eta)$
- Draw word distributions:  $\phi_t \sim \text{Dirichlet}(\beta)$
- For each sender  $s$  to receiver  $r$ ,
  - Draw a number of words  $n_{sr} \sim \text{Poisson}(\lambda_{(C_s)(C_r)})$
  - For each of  $n_{ij}$  words:
    - Draw a topic:  $Z_{wsr} \sim \text{Multinomial}(\theta_{(C_s)(C_r)})$
    - Draw a word:  $W_{wsr} \sim \text{Multinomial}(\phi_{(Z_{wsr})})$





# Topic Blockmodel



# Count Modeling

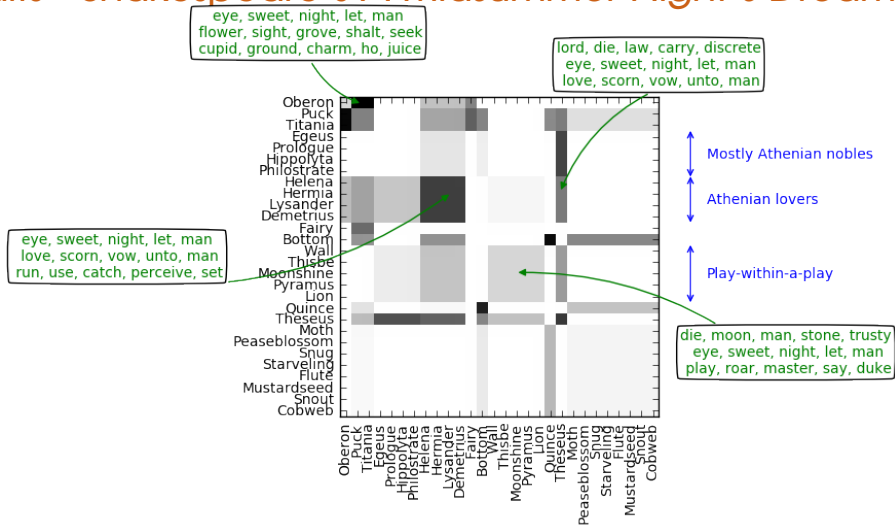
We model not just the *content* but also the *frequency*

$$\lambda_{ij} \sim \text{Gamma}(\alpha_\gamma, \beta_\gamma)$$

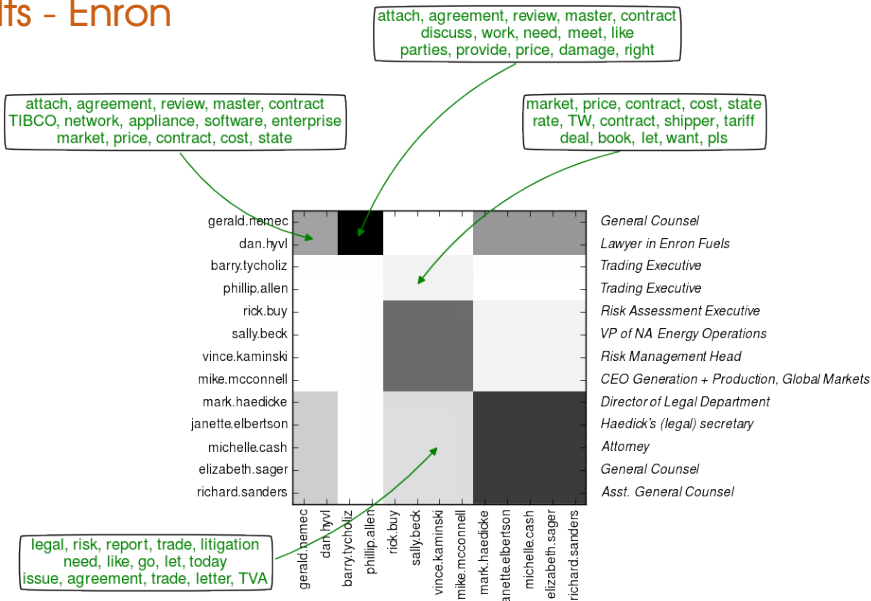
$$N_{sr} \sim \text{Poisson}(\lambda_{(C_s)(C_r)})$$

Captures information from communication frequency / intensity, not just what they discuss

# Results - Shakespeare's *A Midsummer Night's Dream*



# Results - Enron



# Results - Quantitative Tasks

**Table:** Log predictive likelihood ( $\pm$  one standard error) of document text, conditioned on sender and recipient where applicable.

Model	ENRON	Shakespeare
LDA	-410,110.2 $\pm$ 50.8	-48,716.2 $\pm$ 4.6
ART	-365,600.5 $\pm$ 47.7	-47,495.5 $\pm$ 4.8
CNT	-368,983.5 $\pm$ 89.2	<b>-46,076.6 <math>\pm</math> 3.9</b>
<i>Topic Blockmodel</i>	<b>-345.632.5 <math>\pm</math> 4.1</b>	-46,275.9 $\pm$ 4.0

## Results - Quantitative Tasks

**Table:** Log predictive likelihood ( $\pm$  one standard error) of document recipient, conditioned on document content and sender where applicable.

Model	ENRON	Shakespeare
ART	-204,585.3 $\pm$ 6.4	-19,809.7 $\pm$ 1.1
CNT	-216,278.9 $\pm$ <0.1	-19,703.3 $\pm$ <0.1
Poisson-SBM	-160,984.7 $\pm$ 148.6	-14,587.2 $\pm$ 35.9
<i>Topic Blockmodel</i>	<b>-137,199.8 <math>\pm</math> 53.2</b>	<b>-12,997.8 <math>\pm</math> 20.6</b>

## Results - Quantitative Tasks

**Table:** Log predictive likelihood ( $\pm$  one standard error) of document sender and recipient, conditioned on document content where applicable.

Model	ENRON	Shakespeare
ART	-416,588.6 $\pm$ 6.8	-39,580.0 $\pm$ 1.0
CNT	-432,557.7 $\pm$ <0.1	-39,406.7 $\pm$ <0.1
Poisson-SBM	-347,479.6 $\pm$ 148.6	-31,400.3 $\pm$ 35.9
<i>Topic Blockmodel</i>	<b>-321,127.8 <math>\pm</math> 53.3</b>	<b>-29,614.0 <math>\pm</math> 20.6</b>

# Results - Quantitative Tasks

**Table:** Log predictive likelihood ( $\pm$  one standard error) of sender and recipient counts.

Model	ENRON	Shakespeare
Poisson-SBM	$-92,851.2 \pm 12.1$	$-103,411.4 \pm 0.6$
<i>Topic Blockmodel</i>	<b><math>-88,730.4 \pm 3.1</math></b>	<b><math>-102,549.8 \pm 0.2</math></b>



# Conclusion

Questions?

# Monotonic Fairness: Introduction

- *Problem:* Although we can create fair(er) prediction and classification systems, they tend to create resentment which undermines their support.

# Monotonic Fairness: Introduction

- *Problem:* Although we can create fair(er) prediction and classification systems, they tend to create resentment which undermines their support.
- *Intuition:* If we define resentment as seeing someone "worse" get a "better" outcome, we can form models that avoid that outcome.

# Monotonic Fairness: Introduction

- *Problem:* Although we can create fair(er) prediction and classification systems, they tend to create resentment which undermines their support.
- *Intuition:* If we define resentment as seeing someone "worse" get a "better" outcome, we can form models that avoid that outcome.
- *Model:* Modify existing fair neural network models with a monotonic neural network to guarantee that resentment doesn't occur

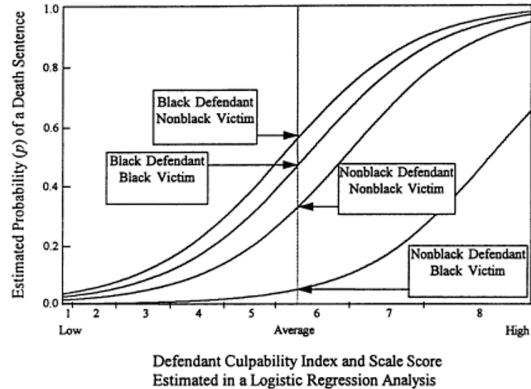
# Fairness in machine learning

Machine learning models are powerful, but imperfect

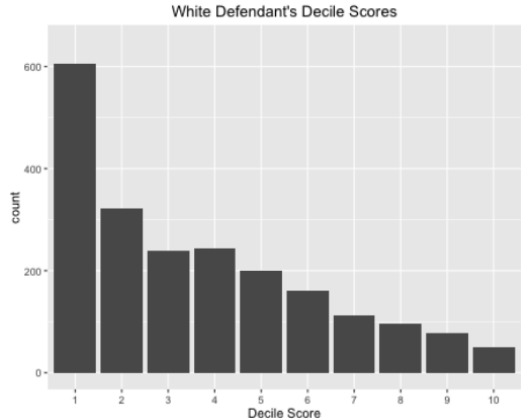
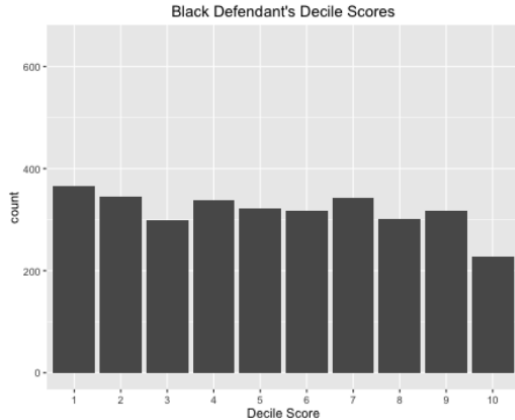
- Tend to reproduce historic biases
- Can overfit, producing functions that are locally nonsensical

# Old Systemic Bias

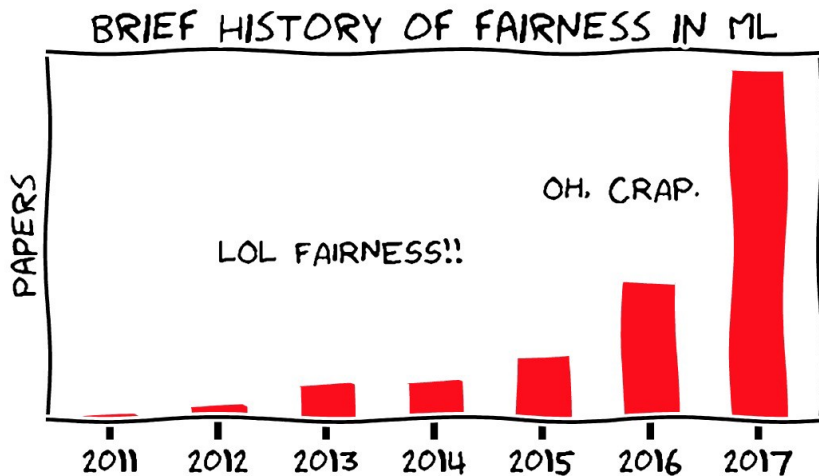
Estimated Race of Defendant and Race of Victim Effects  
in Jury Death Sentencing Decisions Among All Death Eligible Cases  
Philadelphia 1983-93



# New Systemic Bias



## Growing Awareness





# Fairness in machine learning

Fairness in ML generally has 3 steps:

- Conceptualize



# Fairness in machine learning

Fairness in ML generally has 3 steps:

- Conceptualize
- Measure



# Fairness in machine learning

Fairness in ML generally has 3 steps:

- Conceptualize
- Measure
- Prevent



# Fairness in machine learning

Fairness in ML generally has 3 steps:

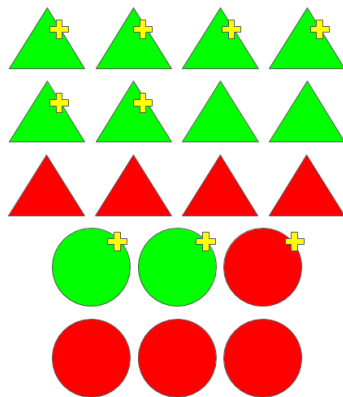
- Conceptualize
- Measure
- Prevent
  - Cheaply, hopefully



# Concepts of fairness: Equality of Outcome

Equality of Outcome: "Each group should have the same outcome on average"

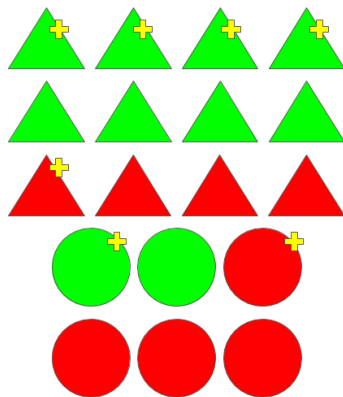
$$\mathbb{E}[\hat{Y}|A = a] = \mathbb{E}[\hat{Y}|A = a']$$



# Concepts of fairness: Equality of Odds

Equality of Odds: "The average prediction should be independent of protected class for people with the same outcome."

$$\mathbb{E}[\hat{Y}|A = a, Y = y] = \mathbb{E}[\hat{Y}|A = a', Y = y]$$

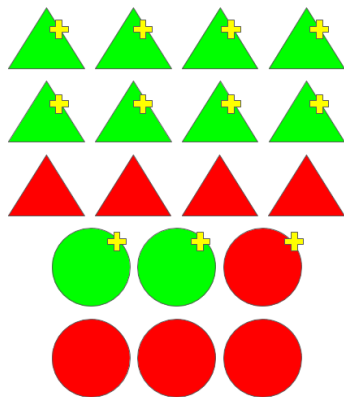


# Concepts of fairness: Equality of Opportunity

Equality of Opportunity: "For people who deserve the favorable outcome, the probability of receiving the favorable prediction should be independent of class"

$$\Pr[\hat{Y} = 1 | A = a, Y = 1] =$$

$$\Pr[\hat{Y} = 1 | A = a', Y = 1]$$



# Concepts of fairness: Individual Fairness

Individual fairness: similar individuals should be treated similarly

- $d(\hat{f}(X_i), \hat{f}(X_j)) \leq D(X_i, X_j)$
- Essentially a requirement of Lipschitz continuity with bounded smoothness.



# Individual Resentment

An individual may still experience *resentment* when:

- They receive a less favorable outcome,
- Another person receives a more favorable outcome,
- And either:
  - The other person is identical except for a protected attribute
  - The other person has "worse" non-protected attributes

# Individual Resentment

## Definition

**Protected Attribute Resentment (Class Resentment):** Individual  $u$  experiences *class resentment* under function  $f$  if  $\exists A' \in \mathcal{A}$  s.t.  $f(X_u, A') > f(X_u, A_u)$ .

# Individual Resentment

## Definition

**Protected Attribute Resentment (Class Resentment):** Individual  $u$  experiences *class resentment* under function  $f$  if  $\exists A' \in \mathcal{A}$  s.t.  $f(X_u, A') > f(X_u, A_u)$ .

## Definition

**Non-Protected Attribute Resentment (Score Resentment):** Individual  $u$  experiences *score resentment* under function  $f$  if  $\exists (X', A') \in (\mathcal{X}, \mathcal{A})$  such that  $X_u$  is objectively "better" than  $X'$  but  $f(X', A') > f(X_u, A_u)$ . ( $A'$  may be  $A_u$ .)

# COMPAS Resentment

*Defendant A*

Caucasian

25 y.o

4 priors

No juvenile charges

Felony, Violent charge

Robbery, no weapon



*Defendant B*

African-American

25 y.o

3 priors

No juvenile charges

Felony, Non-Violent charge

Grand Theft, 3rd Deg.



# COMPAS Resentment

*Defendant A*

Caucasian

25 y.o

4 priors

No juvenile charges

Felony, Violent charge

Robbery, no weapon



Decile Score: 4

*Defendant B*

African-American

25 y.o

3 priors

No juvenile charges

Felony, Non-Violent charge

Grand Theft, 3rd Deg.



Decile Score: 10

# Preventing Resentment

In order to prevent resentment, we propose a system which:

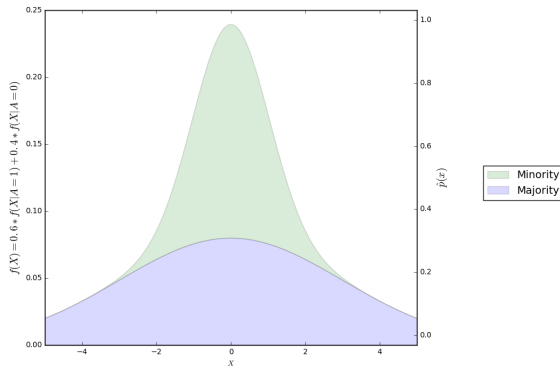
- Takes only non-protected attributes  $X$  (i.e. exclude protected attributes  $A$ ) as input for prediction.
- Use a neural network which has an output function that is monotonic w.r.t. those dimensions  $X$  that are user specified

# Preventing Resentment

We then train that function to minimize a weighted sum of prediction loss and group fairness loss.

$$\text{Loss} = (1 - \alpha) \text{Loss}_{Acc} + \alpha \text{Loss}_{Fair}$$

# Resentment Graphically



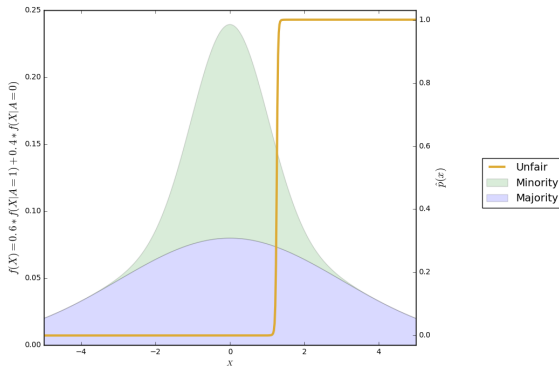


# Resentment Graphically

Learn  $f : \mathbb{R} \rightarrow [0, 1]$

Maximize  $\sum f(X_i)X_i / \sum f(X_i)$

s.t.  $\sum f(X_i)/n = 0.25$



# Resentment Graphically

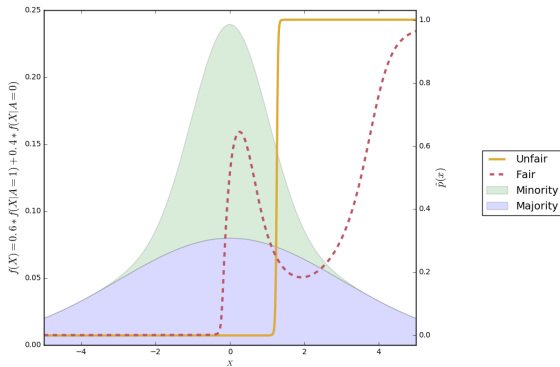
Learn  $f : \mathbb{R} \rightarrow [0, 1]$

Maximize  $\sum f(X_i)X_i / \sum f(X_i)$

s.t.  $\sum f(X_i)/n = 0.25$

Add fairness:

$$\mathbb{E}[f|A=0] = \mathbb{E}[f|A=1]$$



# Resentment Graphically

Learn  $f : \mathbb{R} \rightarrow [0, 1]$

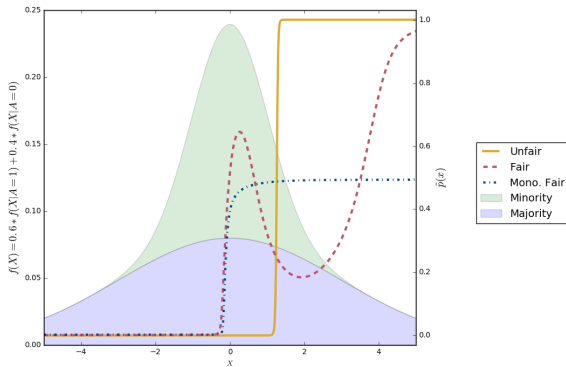
Maximize  $\sum f(X_i)X_i / \sum f(X_i)$

s.t.  $\sum f(X_i)/n = 0.25$

Add fairness:

$$\mathbb{E}[f|A=0] = \mathbb{E}[f|A=1]$$

...and no resentment



# Monotonic Neural Networks

As proposed by Sill 1998, redefine hidden nodes as:

$$h_{j,l} = \sigma \left( \sum_{i \in 1 \dots |H_{l-1}|} \tau(w_{i,j}^l) h_{j,l-1} + b_j^l \right)$$

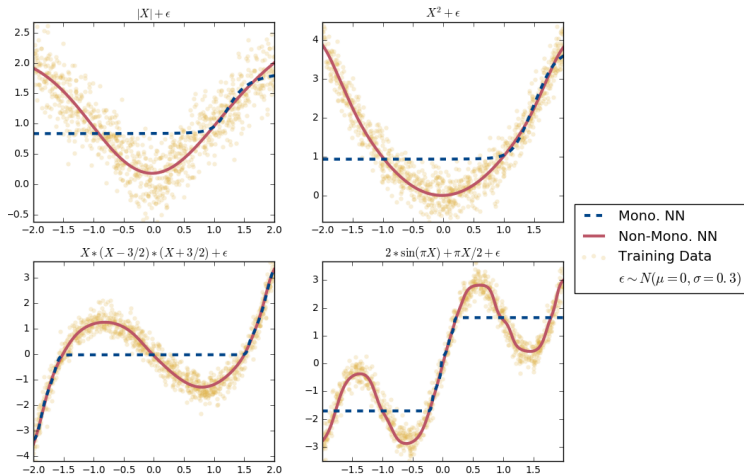
where  $\tau : \mathbb{R} \rightarrow \mathbb{R}_+$  and  $\sigma$  is monotonically non-decreasing

# Mixed Monotonicity

If we *don't apply*  $\tau$  to the weights in the first layer corresponding to dimension of  $X_j$ , then the function *will not be monotonic* w.r.t.  $X_j$ , even if we transform the weights in subsequent layers.

If we replace  $\tau(w_{i,j}^1)$  with  $-\tau(w_{i,j}^1)$  for some input dimension  $j$ , then the first layer (and all subsequent layers) will have a monotonic non-increasing relationship with that dimension.

# Monotonicity Demo



# Datasets

We evaluate our model on three fairness-related data sets:

- *Law school* : 17,400 law school applicants, trying to predict law school grade as function of LSAT and undergrad GPA, protecting gender.

# Datasets

We evaluate our model on three fairness-related data sets:

- *Law school* : 17,400 law school applicants, trying to predict law school grade as function of LSAT and undergrad GPA, protecting gender.
- *COMPAS* : 7,000 individuals arrested in Broward County Florida, trying to predict re-arrest on bail based on prior record, while preventing gender and racial discrimination. Age is a non-monotonic, non-protected attribute.

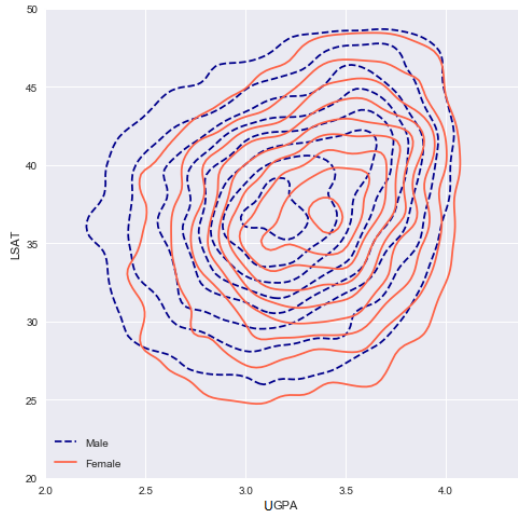


# Datasets

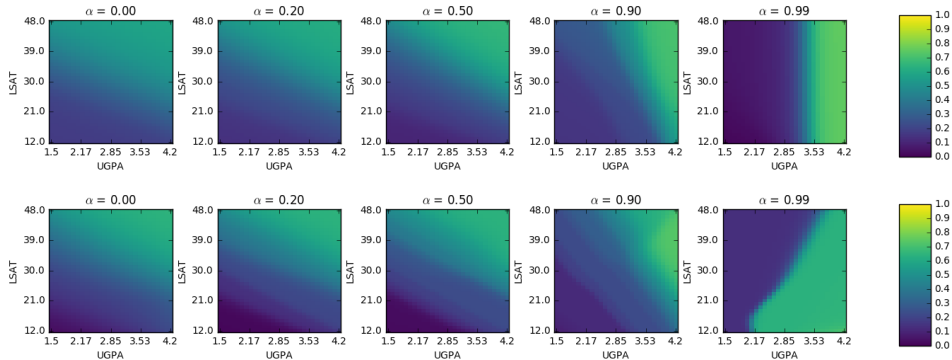
We evaluate our model on three fairness-related data sets:

- *Law school* : 17,400 law school applicants, trying to predict law school grade as function of LSAT and undergrad GPA, protecting gender.
- *COMPAS* : 7,000 individuals arrested in Broward County Florida, trying to predict re-arrest on bail based on prior record, while preventing gender and racial discrimination. Age is a non-monotonic, non-protected attribute.
- *German Credit Data* : 1,000 West German credit applications. Predict loan repayment based on employment, financial, and residency information while protecting (binary) age. 58 attributes with mixture of monotonicity.

# Law School Gender Density



# Law School Acceptance Functions



( $\alpha$  is fraction of loss from fairness)

# Evaluation metrics

We follow the evaluation metrics of Zemel 2013 *Learning Fair Representations*.

- Discrimination: disparate impact i.e. absolute difference in expectation:

$$\left| \frac{\sum_{n:s_n=1} \hat{y}_n}{\sum_{n:s_n=1} 1} - \frac{\sum_{n:s_n=0} \hat{y}_n}{\sum_{n:s_n=0} 1} \right|$$

# Evaluation metrics

We follow the evaluation metrics of Zemel 2013 *Learning Fair Representations*.

- Discrimination: disparate impact i.e. absolute difference in expectation:

$$\left| \frac{\sum_{n:s_n=1} \hat{y}_n}{\sum_{n:s_n=1} 1} - \frac{\sum_{n:s_n=0} \hat{y}_n}{\sum_{n:s_n=0} 1} \right|$$

- Accuracy: inverse of mean absolute error

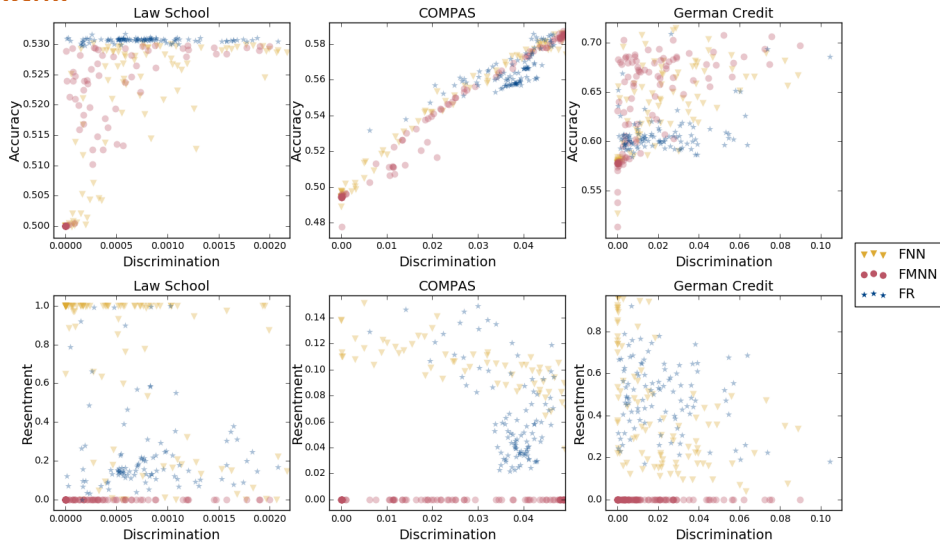
$$1 - \frac{1}{N} \sum_{n=1}^N |y_n - \hat{y}_n|$$

# Evaluation metrics

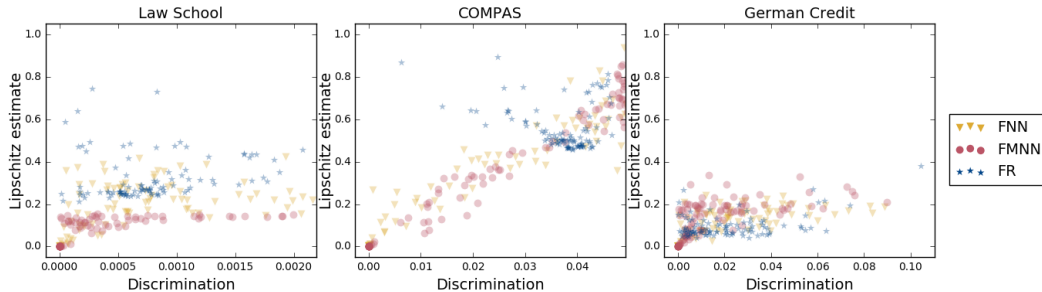
We add Resentment: fraction of people that find example in sample they resent

$$\frac{1}{N} \sum_{i=1}^N \max_{j \in \mathcal{N}_i} (1_{\hat{y}_i < \hat{y}_j})$$

# Results



# Lipschitz Smoothness





# Eliciting Monotonic Fairness: Introduction

- *Problem:* Axis-based monotonicity is too strict for fairness, and oracle info about which individuals are "better" is hard to systematize (even if we had an oracle).

# Eliciting Monotonic Fairness: Introduction

- *Problem:* Axis-based monotonicity is too strict for fairness, and oracle info about which individuals are "better" is hard to systematize (even if we had an oracle).
- *Intuition:* Lay people experience resentment, so we should be able to survey them and use that data to figure out what is a "better" set of attributes.

# Eliciting Monotonic Fairness: Introduction

- *Problem:* Axis-based monotonicity is too strict for fairness, and oracle info about which individuals are "better" is hard to systematize (even if we had an oracle).
- *Intuition:* Lay people experience resentment, so we should be able to survey them and use that data to figure out what is a "better" set of attributes.
- *Model:* Collect non-expert arbiter ratings on what the relative treatment of pairs of individuals should be. Combine this with historical data as input to a conditional neural network that can post hoc adjust between accuracy and resentment prevention.

# Introduction

In practice we often wish to capture more complex definitions of "better" attribute sets that consider multiple attributes at once.

Imagine you're evaluating whether two defendants should get bail:

- Defendant A: 0 prior felonies, 1 prior misdemeanors
- Defendant B: 10 prior felonies, 0 prior misdemeanors

Monotonicity as previously discussed would say A and B are incomparable, but most people would agree that B should be less likely to get bail.

# Preference Learning

Ultimately we have a problem of *preference learning*. We have two options,  $X_i$  and  $X_j$ , and want to learn a preference function between them.

Many approaches in the literature aim to learn personalized preference functions to recommend the best product *for an individual*. We wish to learn a population-wide preference function which will be applied universally.

# Model Structure

Re-encode  $Y_i$  into pairwise data  $Z_{ij}$ :

$$Z_{ij}^{obs} = \begin{cases} 1 & \text{if } Y_i^{obs} = 1 \text{ and } Y_j^{obs} = 0 \\ 2 & \text{if } Y_i^{obs} = 0 \text{ and } Y_j^{obs} = 1 \\ 3 & \text{if } Y_i^{obs} = Y_j^{obs} \end{cases} .$$

Will use the same encodings for survey data for "more likely", "less likely", and "similarly likely"

# Model Structure

Define a pairwise loss function using  $Z$ :

$$\mathcal{L}_Z(Z, \hat{p}, \mathcal{Z}) = -\frac{1}{|\mathcal{Z}|} \sum_{(i,j) \in \mathcal{Z}} \left( \begin{array}{l} \mathbf{1}_{Z_{ij}=1} \log(\hat{p}_i(1 - \hat{p}_j)) + \\ \mathbf{1}_{Z_{ij}=2} \log((1 - \hat{p}_i)\hat{p}_j) + \\ \mathbf{1}_{Z_{ij}=3} \log(\hat{p}_i\hat{p}_j + (1 - \hat{p}_i)(1 - \hat{p}_j)) \end{array} \right)$$

# Model Loss

Define our neural network:

$$\text{logit}(\hat{p}_i) = f_{\theta}(X_i, c)$$

Define the loss the minimize:

$$\mathcal{L} = \underbrace{\mathcal{L}_Z(Z_{ij}^{obs}, \hat{p}_i = f(X_i, c = 0), \mathcal{O})}_{\mathcal{L}_Z^{obs}} + \underbrace{\mathcal{L}_Z(Z_{ij}^{arb}, \hat{p}_i = f(X_i, c = 1), \mathcal{A})}_{\mathcal{L}_Z^{arb}} + g(\theta)$$

Optimize  $\hat{\theta} = \arg \min_{\theta} \mathcal{L}$ .



# Synthetic Experiment Setup

$$X_i \sim N(0, 1)^2$$

$$\beta_{obs} = [0.9, 1.1]$$

$$p_i^{obs} = \frac{1}{1 + e^{-(X_i \beta^{obs} - 1)}}$$

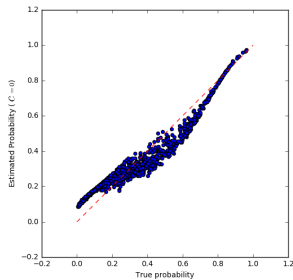
$$Y_i \sim \text{Bernoulli}(p_i^{obs})$$

$$\beta_{arb} = [1.1, 0.9]$$

$$Z_{ij}^{arb} = \begin{cases} 1 & \text{if } X_i \beta^{arb} > X_j \beta^{arb} + 0.25 \\ 2 & \text{if } X_j \beta^{arb} > X_i \beta^{arb} + 0.25 \\ 3 & \text{if } |X_i \beta^{arb} - X_j \beta^{arb}| < 0.25 \end{cases}$$

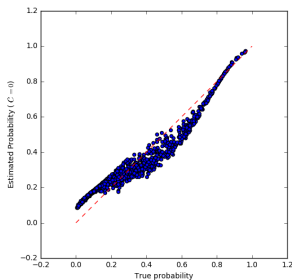
# Synthetic Experiment Results

Recover  $\hat{p}$

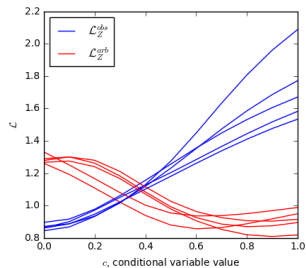


# Synthetic Experiment Results

Recover  $\hat{p}$

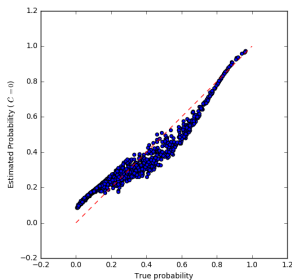


Losses respond to  $c$

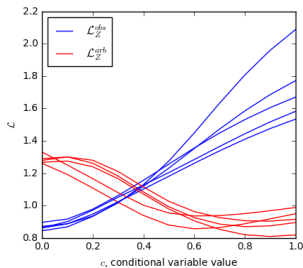


# Synthetic Experiment Results

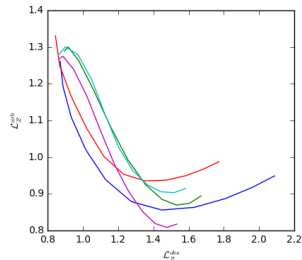
Recover  $\hat{p}$



Losses respond to  $c$



Tradeoff



# COMPAS Experiment Setup

Five arbiters each shown 100 pairs on individuals' attributes: age, (adult) priors count, juv. felony count, juv. misdemeanor count, juv. other counts, charge degree (fel. or mis.), and violent charge (T/F). Asked to rate as:

- "A is at least as likely to (re)offend" ( $Z = 1$ )
- "B is at least as likely to (re)offend" ( $Z = 2$ )
- "A and B are similarly likely to (re)offend" ( $Z = 3$ )
- "No preference / any of the others are fair" (Ignored)

# COMPAS Arbiter Results

- 298 dissimilar ( $Z \in \{1, 2\}$ ) ratings, 185 similar ( $Z = 3$ ) responses, 18 ratings ignored

# COMPAS Arbiter Results

- 298 dissimilar ( $Z \in \{1, 2\}$ ) ratings, 185 similar ( $Z = 3$ ) responses, 18 ratings ignored
- Surprisingly accurate: 78% of dissimilar ratings correct, similar to COMPAS decile score difference of 3 ( $\sim 54\%$  of pairs have decile score difference  $\geq 3$ )

# COMPAS Arbiter Results

- 298 dissimilar ( $Z \in \{1, 2\}$ ) ratings, 185 similar ( $Z = 3$ ) responses, 18 ratings ignored
- Surprisingly accurate: 78% of dissimilar ratings correct, similar to COMPAS decile score difference of 3 ( $\sim 54\%$  of pairs have decile score difference  $\geq 3$ )
- Disparate impact: when comparing African-American to Caucasian, rate former more likely to re-offend 65% of the time



# COMPAS Experiment Loss

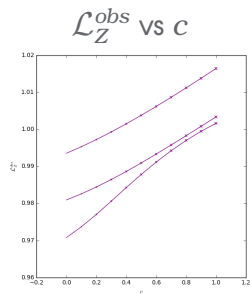
Add Equality of Odds loss:

$$\mathcal{L}_F = \sum_y \sum_a (\bar{\hat{y}}_{ay} - \bar{\hat{y}}_{\cdot y})^2$$

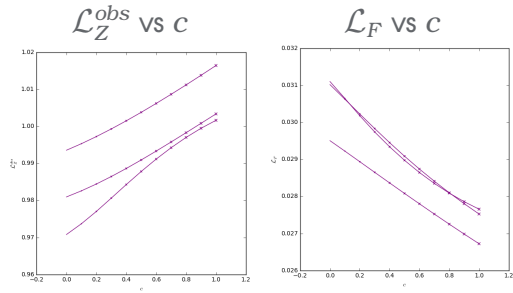
$$\mathcal{L} = \mathcal{L}_Z^{obs} + \mathcal{L}_Z^{arb} + \lambda_F \mathcal{L}_F + g(\theta)$$

Set  $\lambda_F = 0.001$  for experiments.

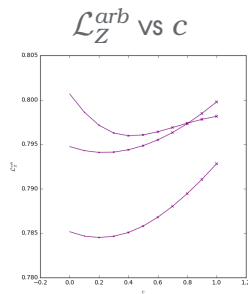
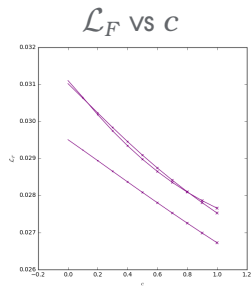
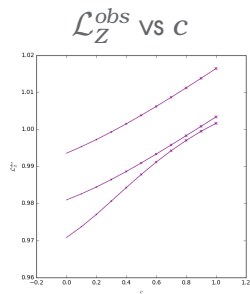
# COMPAS Experiment Results



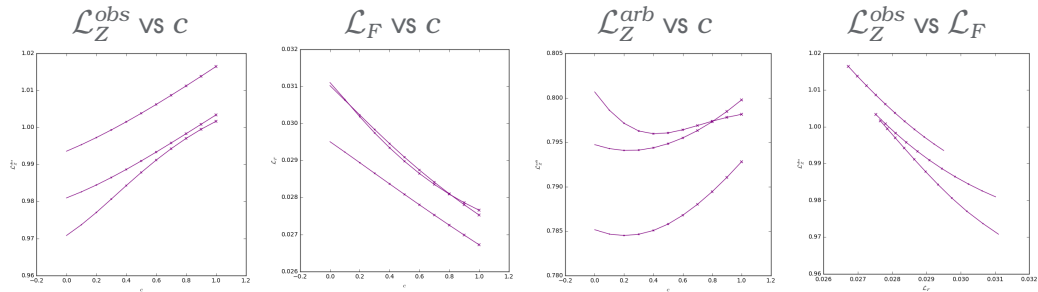
# COMPAS Experiment Results



# COMPAS Experiment Results



# COMPAS Experiment Results



# Conclusion






Today we covered:

- How to model communities in networks based on the content and frequency of their members communications,
- How to create fair functions that avoid resentment between users, and
- How to use arbiter ratings to determine who would resent whom (and avoid it).

# Conclusion






Questions?

# Abridged Bibliography I

-  Blei, D.M., A.Y. Ng, and M.I. Jordan (2003). "Latent Dirichlet allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022.
-  Larson, Jeff et al. (2016). "How we analyzed the COMPAS recidivism algorithm". In: *ProPublica* (5 2016) 9.
-  Lichman, M (2013). *UCI machine learning repository*. URL: [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)).
-  McCallum, A., A. Corrada-Emmanuel, and X. Wang (2005). "The author-recipient-topic model for topic and role discovery in social networks, with application to Enron and academic email". In: *Workshop on Link Analysis, Counterterrorism and Security*.
-  Sill, Joseph (1998). "Monotonic networks". In: *Advances in Neural Information Processing Systems*, pp. 661–667.



# Abridged Bibliography II

-  Snijders, T.A.B. and T. Nowicki (1997). "Estimation and prediction for stochastic blockmodels for graphs with latent block structure". In: *Journal of Classification* 14.1, pp. 75–100.
-  Tu, Y. et al. (2010). "Citation author topic model in expert search". In: *ACL International Conference on Computational Linguistics*.
-  Wang, Y.J. and G.Y. Wong (1987). "Stochastic blockmodels for directed graphs". In: *Journal of the American Statistical Association* 82.397, pp. 8–19.
-  Wightman, Linda F and Henry Ramsey (1998). *LSAC national longitudinal bar passage study*. Law School Admission Council.
-  Zemel, Rich et al. (2013). "Learning fair representations". In: *International Conference on Machine Learning*, pp. 325–333.

# Proving Monotonicity

Define a *monotonic non-decreasing (MND)* function  $f : \mathbb{R} \rightarrow \mathbb{R}$   
s.t.  $f(x + dx) - f(x) = df \geq 0 \forall dx \geq 0$ .

Assume  $f, g$  are MND,  $h$  is monotone non-increasing (MNI), then:

Recursion:  $f \circ g$  is MND

Negation:  $f \circ h$  and  $h \circ f$  are MNI

Linearity: if  $a > 0$ ,  $af(x) + b$  is MND

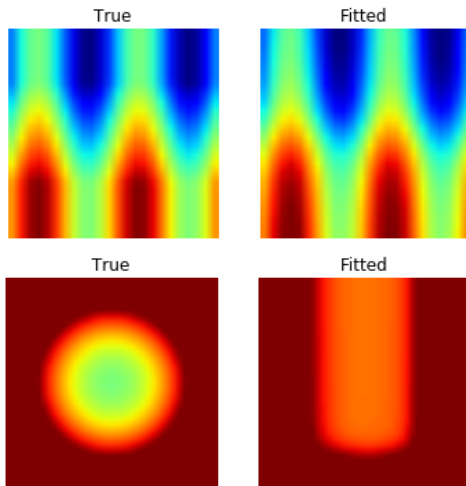
Addition:  $f + g$  is MND

# Proving Monotonicity

We can then prove the properties of our network:

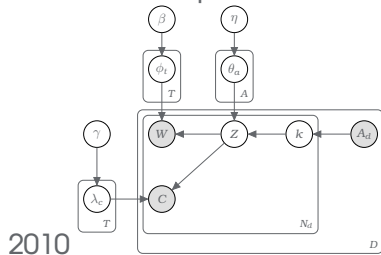
- By Linearity, if  $\tau(w_{i,j}^l) > 0$  then  $\tau(w_{i,j}^l)h_{j,l-1} + b_j^l$  is MND w.r.t.  $h_{j,l-1}$
- By Addition,  $\sum_{i \in 1 \dots |H_{l-1}|} \tau(w_{i,j}^l)h_{j,l-1} + b_j^l$  is MND w.r.t. each of  $h_{j,l-1}$
- By Recursion, if  $\sigma$  is MND then  $h_{i,j}^l \sigma \left( \sum_{i \in 1 \dots |H_{l-1}|} \tau(w_{i,j}^l)h_{j,l-1} + b_j^l \right)$  is MND w.r.t. each of  $h_{j,l-1}$

# Mixed Monotonicity Demo



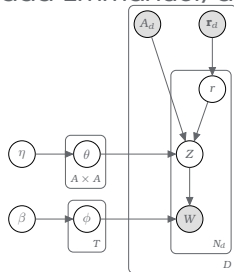
# Related Models

Citation-Author Topic Model Tu et al.



2010

Author-Recipient Topic Model  
McCallum, Corrada-Emmanuel, and



X. Wang 2005