Some Contributions to Multi-view Learning

Chao Lan
Assistant Professor
Department of Computer Science
University of Wyoming

Outline

Introduction to multi-view learning

My contributions to multi-view learning

- statistical learning

- matrix recovery

A brief introduction to fairness problem in machine learning

Our current research initiatives

Outline

Introduction to multi-view learning

My contributions to multi-view learning

- statistical learning

- matrix recovery

A brief introduction to fairness problem in machine learning

Our current research initiatives

How to effectively use unlabeled data
in semi-supervised learning?

If subject is describable from multiple views for learning...

Avrim Blum
Professor of Computer Science
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891

Admin: Ann Stetser

On August 1, 2017 I joined TTI-Chicago as its new CAO. My new homepage is http://www.ttic.edu/blum.
My new email address is *[firstname] at ttic.edu.*

My main research interests are in machine learning theory, approximation algorithms, on-line algorithms, algorithmic game theory / mechanism design, the theory of database privacy, and non-worst-case analysis of algorithms. Some time ago I also did work in AI Planning.

I am on the organizing committee for the STOC 2017 Theory Fest, the STOC 2017 Workshop Program, and was recently on Program Committees for STOC 2016 and COLT 2014. I also co-organized the STOC 2013 Workshop on New (Theoretical) Challenges in Machine Learning. and was co-PI for the Indo-US Joint Center for Advanced Research in Machine Learning, Game Theory, and Optimization. For more information on my research, see the publications and research interests links below. My home department is the CMU Computer Science Department, but I am also affiliated with the CMU Machine Learning Department. I am additionally a member of the CS Theory Group.

If subject is describable from multiple views for learning...

Avrim Blum
Professor of Computer Science
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891

Admin: Ann Stetser

On August 1, 2017 I joined TTI-Chicago as its new CAO. My new homepage is http://www.ttic.edu/blum.
My new email address is *[firstname] at ttic.edu.*

Assumptions
1. sufficiency (or, consistency)
2. independence (or relaxations)

My main research interests are in machine learning theory, approximation algorithms, on-line algorithms, algorithmic game theory / mechanism design, the theory of database privacy, and non-worst-case analysis of algorithms. Some time ago I also did work in AI Planning.

I am on the organizing committee for the STOC 2017 Theory Fest, the STOC 2017 Workshop Program, and was recently on Program Committees for STOC 2016 and COLT 2014. I also co-organized the STOC 2013 Workshop on New (Theoretical) Challenges in Machine Learning, and was co-PI for the Indo-US Joint Center for Advanced Research in Machine Learning, Game Theory, and Optimization. For more information on my research, see the publications and research interests links below. My home department is the CMU Computer Science Department, but I am also affiliated with the CMU Machine Learning Department. I am additionally a member of the CS Theory Group.

Representative Algorithm: Co-Training

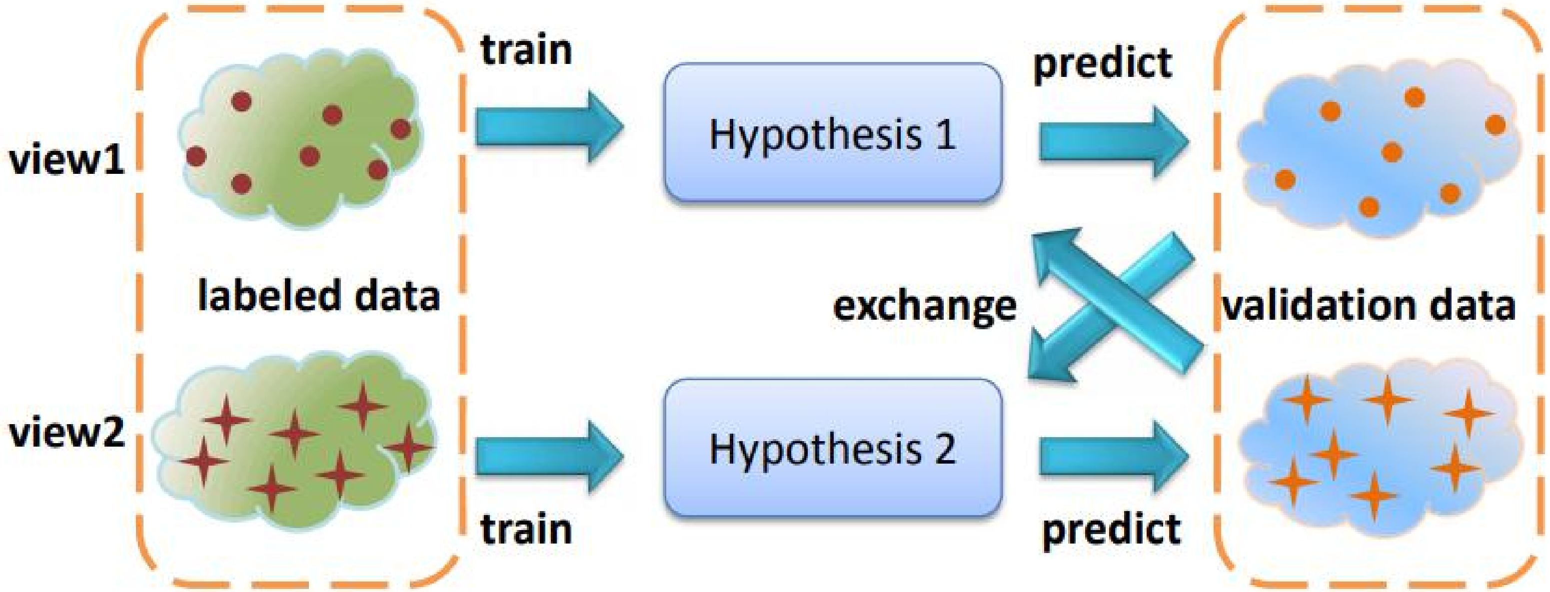
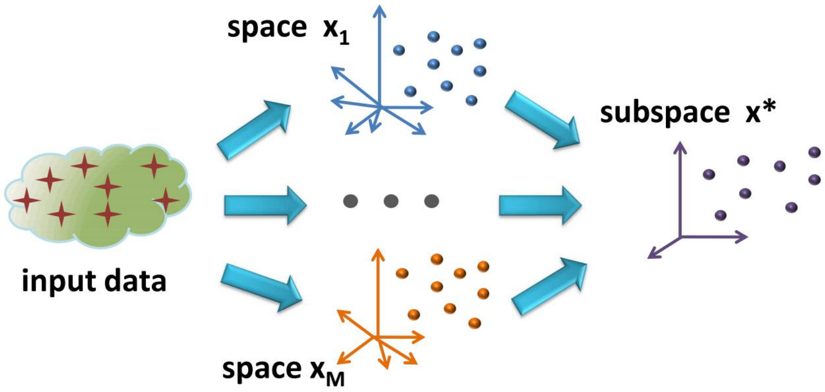


Image Source: Xu, Tao and Xu. A Survey on Multi-view Learning. CoRR 2013.

Representative Algorithm: Co-Regularization

$$\min_{f_1,f_2} \underbrace{\sum\nolimits_{x,i} \| f_i(x_{[i]}) - y \|}_{\text{predictive loss on labeled data } x} + \lambda \underbrace{\sum\nolimits_{u} \| f_1(u_{[1]}) - f_2(u_{[2]}) \|}_{\text{view agreement on unlabeled data } u} + \Lambda(f_1, f_2),$$

Other Multi-View Learning Techniques



Multi-View Subspace Learning

Multi-View Kernel Method

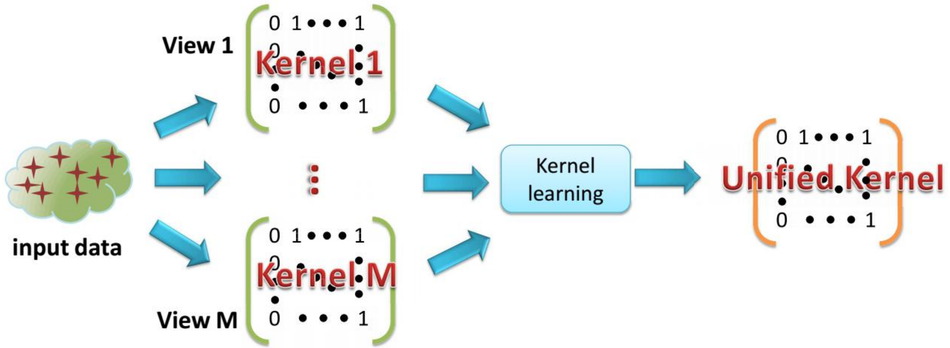


Image Source: Xu, Tao and Xu. A Survey on Multi-view Learning. CoRR 2013.

Many applications of multi-view learning...

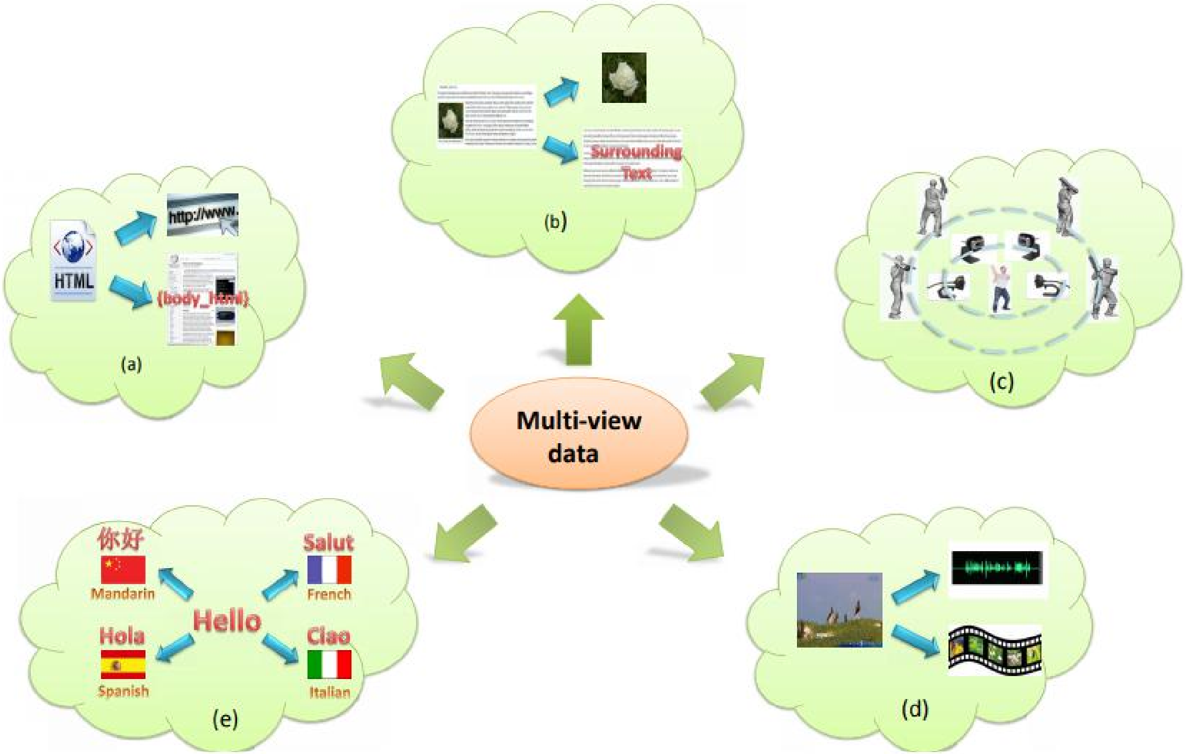


Image Source: Xu, Tao and Xu. A Survey on Multi-view Learning. CoRR 2013.

Outline

Introduction to multi-view learning

My contributions to multi-view learning

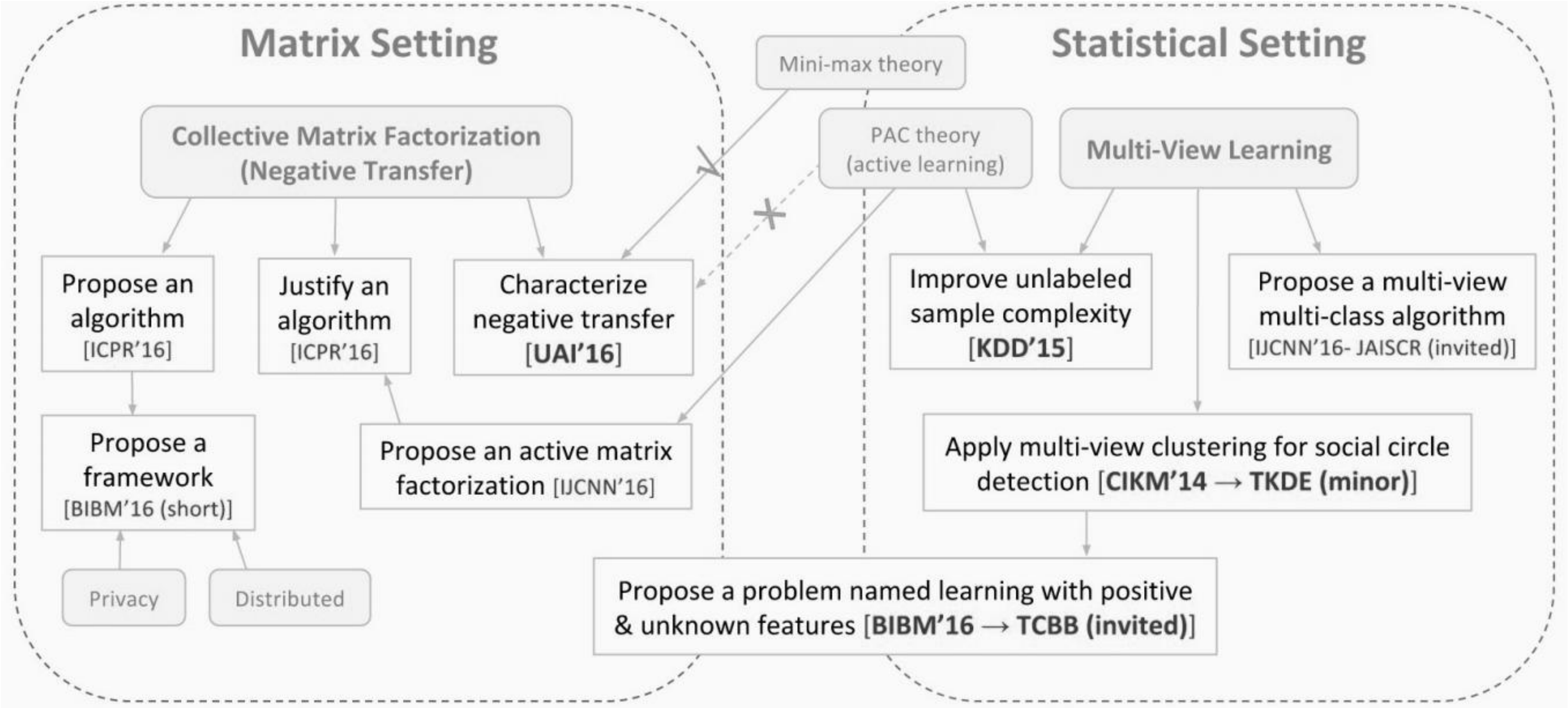
- statistical learning

- matrix recovery

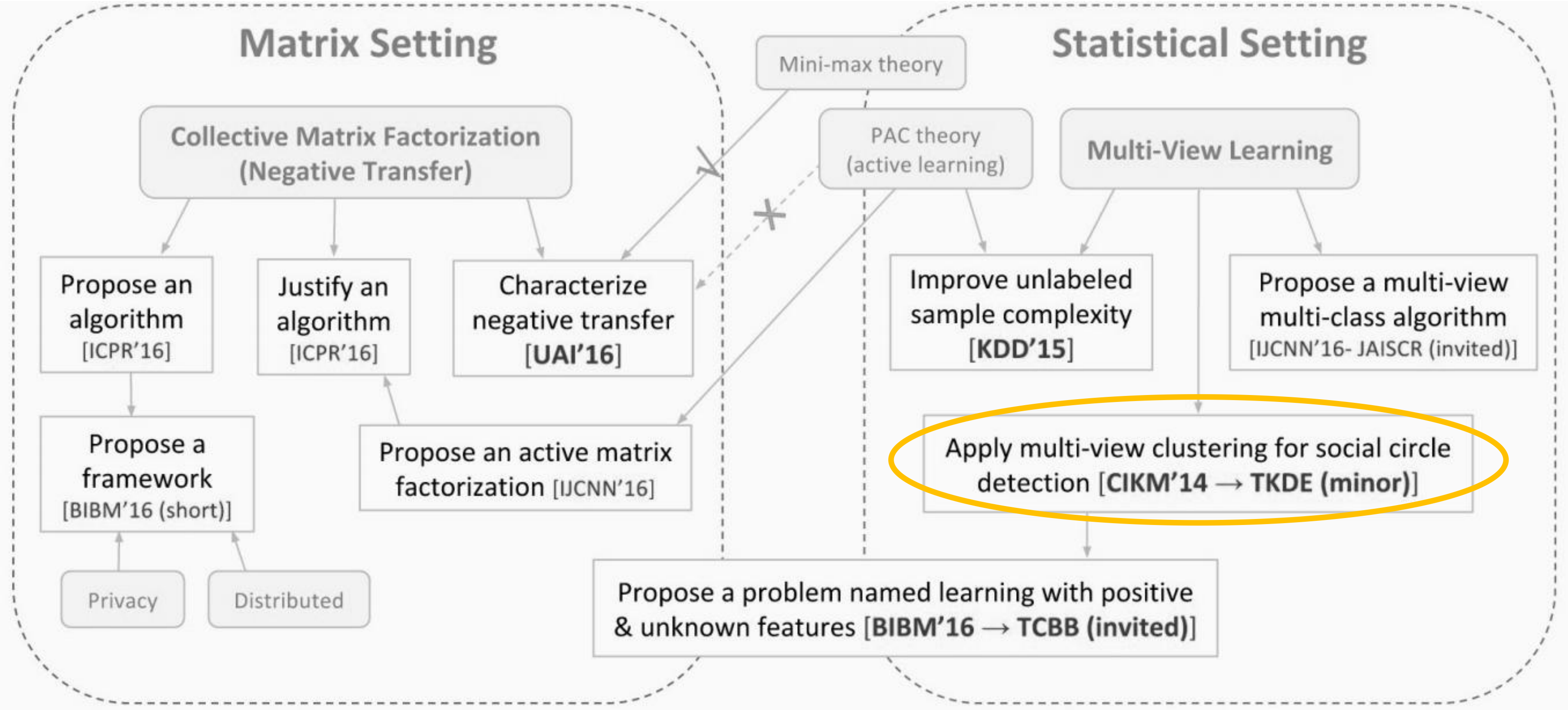
A brief introduction to fairness problem in machine learning

Our current research initiatives

My PhD Research Overview

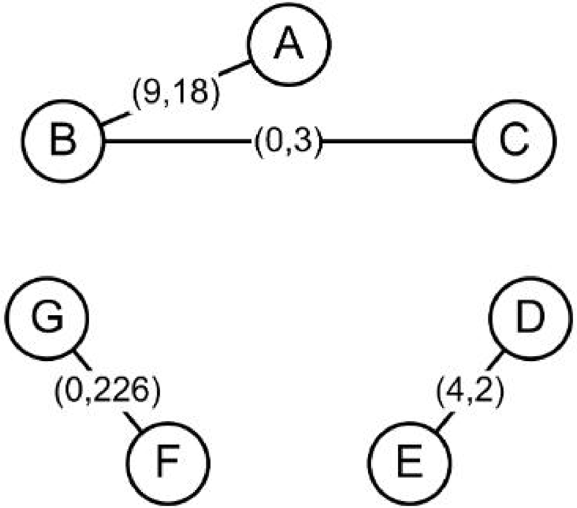


My PhD Research Overview



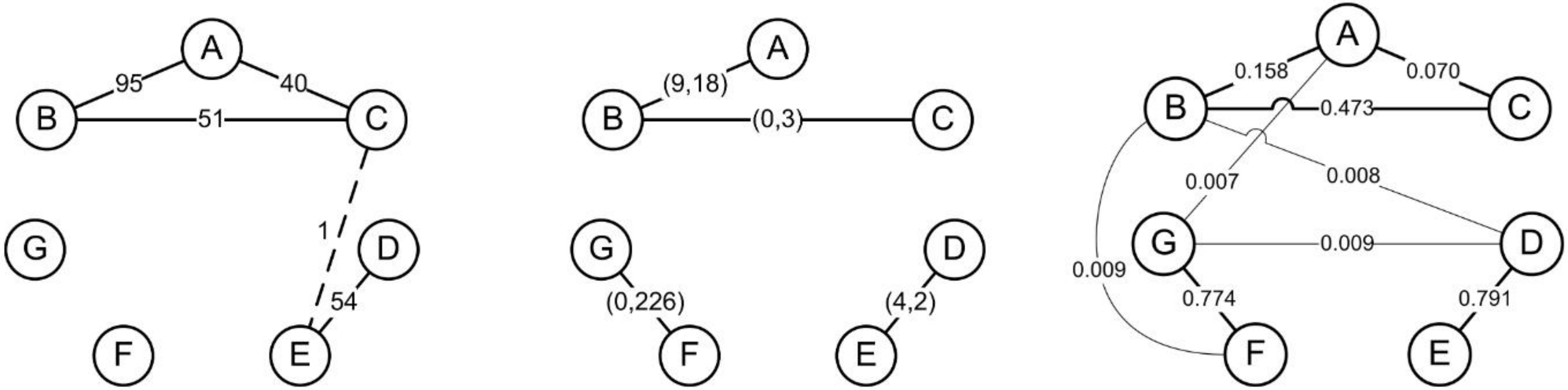
Background: Social Circle Detection

A social circle is a group of one's similar friends in a social network.



Our Contribution: Multi-View Social Circle Detection

Existing detection techniques focus on a single view of network structure, and we propose to detect social circles based on multi-view network structure.



(a) Structural Views (b) Interaction Views (C) Content View

Core Technique: Multi-View Spectral Clustering

1. Solve spectral clustering on individual graphs to get the discriminative eigenvectors in each view, say $\mathbf{U}_1$ and $\mathbf{U}_2$.
2. Cluster points using $\mathbf{U}_1$ and use this clustering to *modify* the graph structure in view 2.
3. Cluster points using $\mathbf{U}_2$ and use this clustering to *modify* the graph structure in view 1.
4. Go to Step 1 and repeat for a number of iterations.

Image Source: Kumar and Dalume. A co-training approach for multi-view spectral clustering. ICML, 2011.

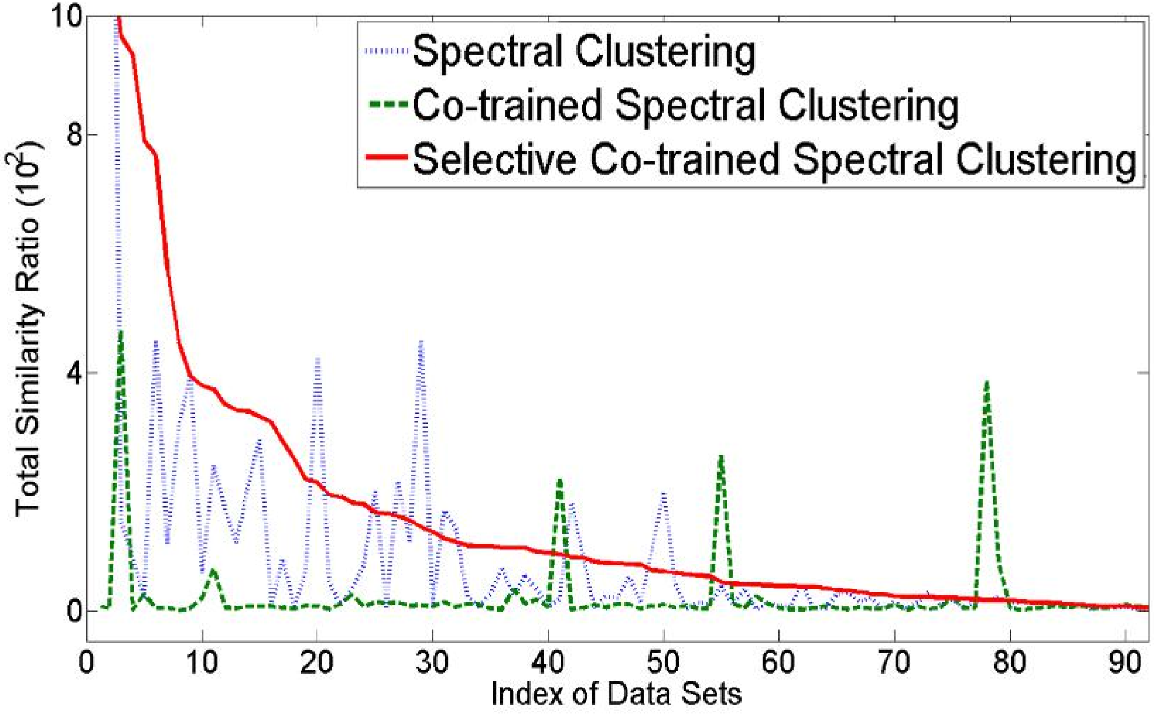
Experiment Design

Crawl ~90 ego-networks from Twitter, each contraining ~250 people.

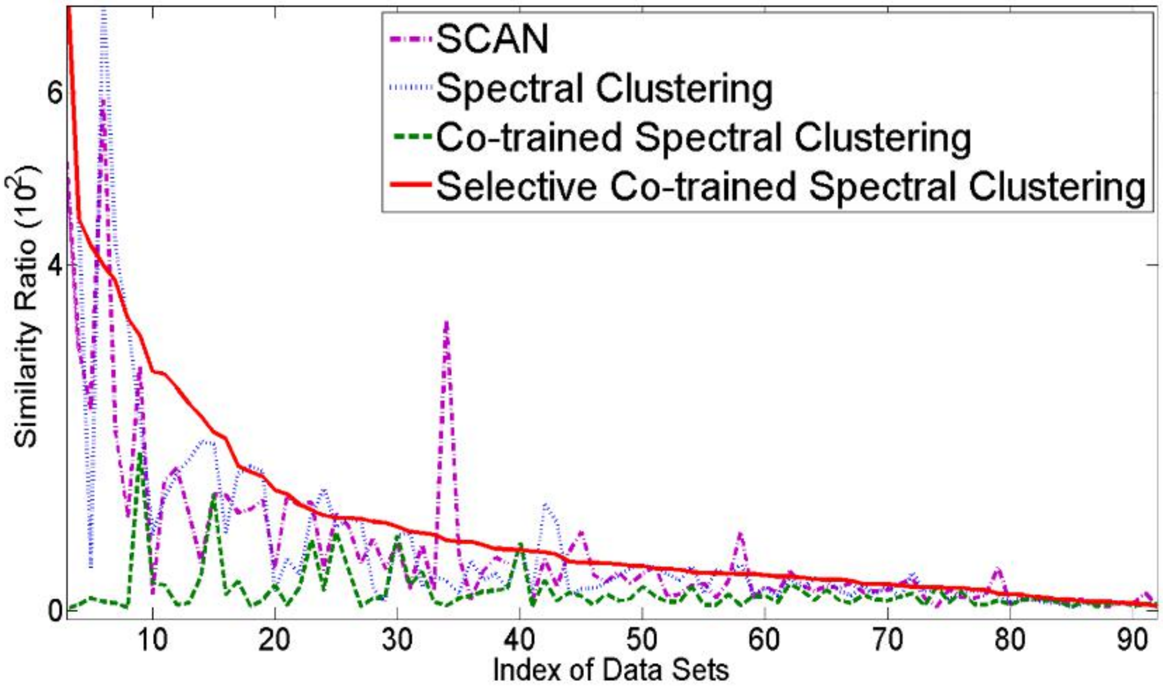
Construct six views from each data set.
- friend
- common friend
- reply
- co-reply
- retweet
- content

Evaluation Metric: total similarity ratio $\gamma = \frac{\sum_{t\in[T]} S_w^{(t)}}{\sum_{t\in[T]} S_b^{(t)}}$

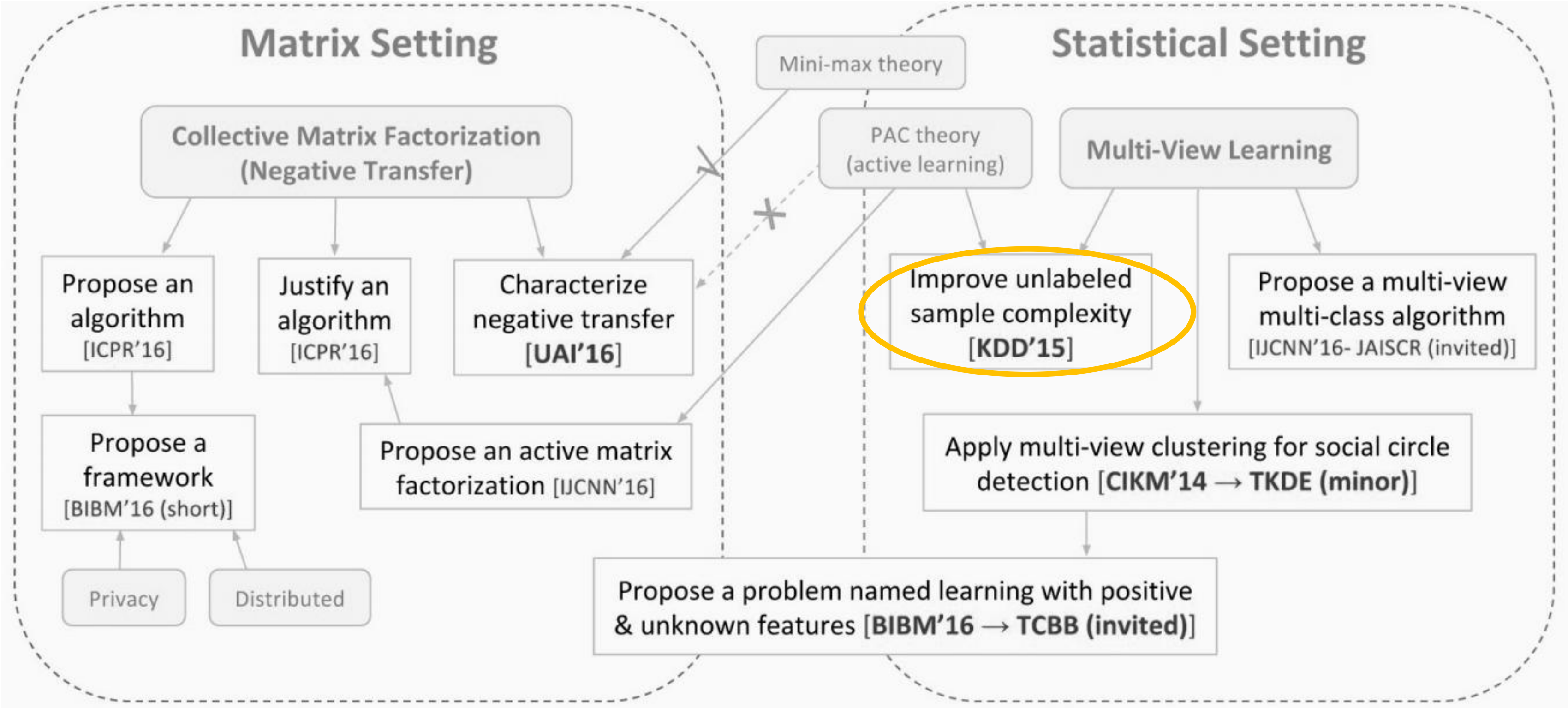
Result over All Views



Result on Friend View



My PhD Research Overview



Background: Unlabeled Sample Complexity

Sample complexity is the number of training data for learning an accurate model.

Labeled sample complexity of supervised learning is well-studied.

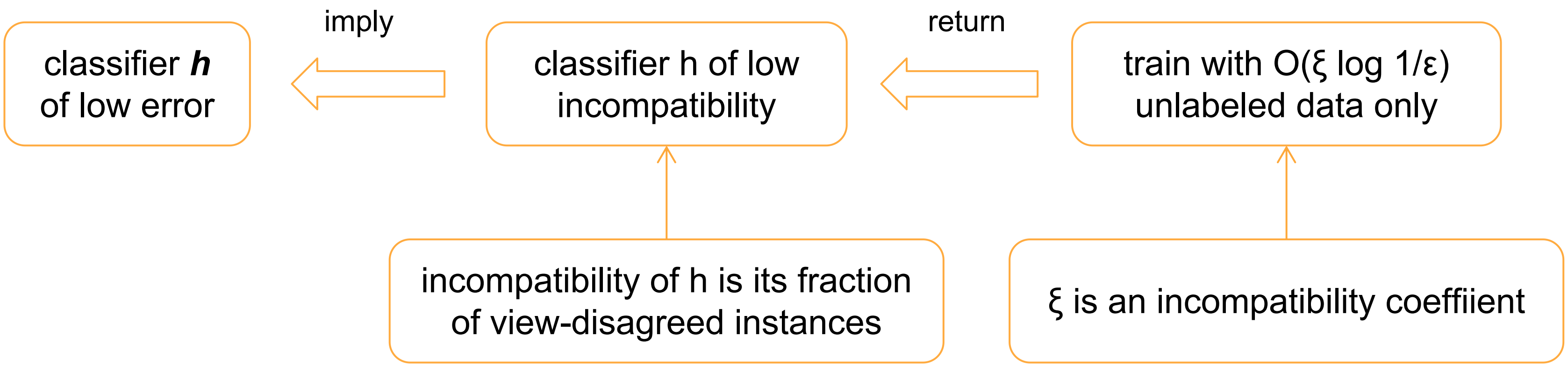
- single-view learning is $O(1/\varepsilon)$ for passive and $O(\log 1/\varepsilon)$ for active
- multi-view learning is $O(\log 1/\varepsilon)$ for active

Unlabeled sample complexity of semi-supervised learning is not well-studied.

- single-view learning is $O(1/\varepsilon)$
- multi-view learning is $O(1/\varepsilon)$

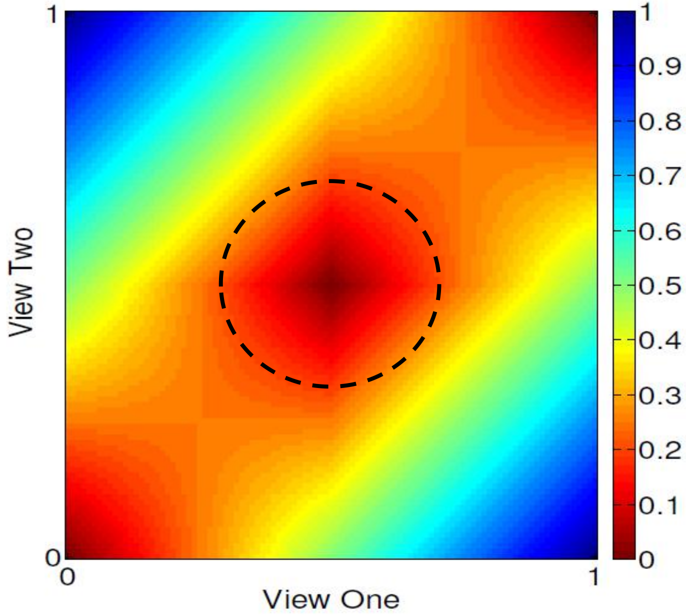
Our Contribution: Improve Unlabeled Sample Complexity

We improve the unlabeled sample complexity of multi-view semi-supervised learning from O(1/ε) to O(log1/ε), under the PAC framework.



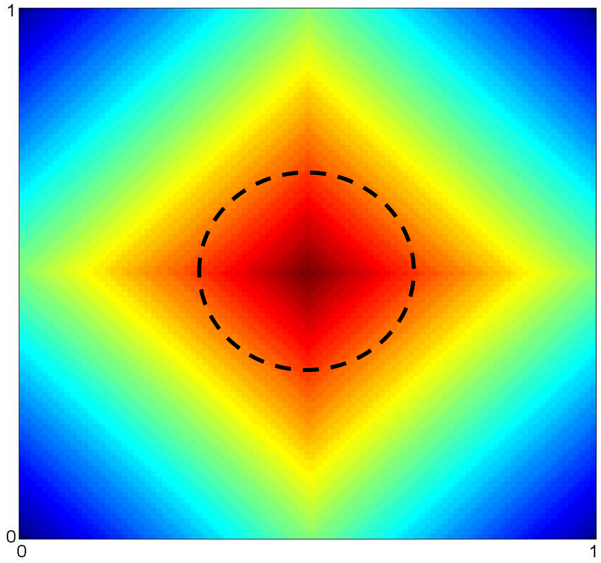
Relation between Incomplatibility and Error

Incompatibility



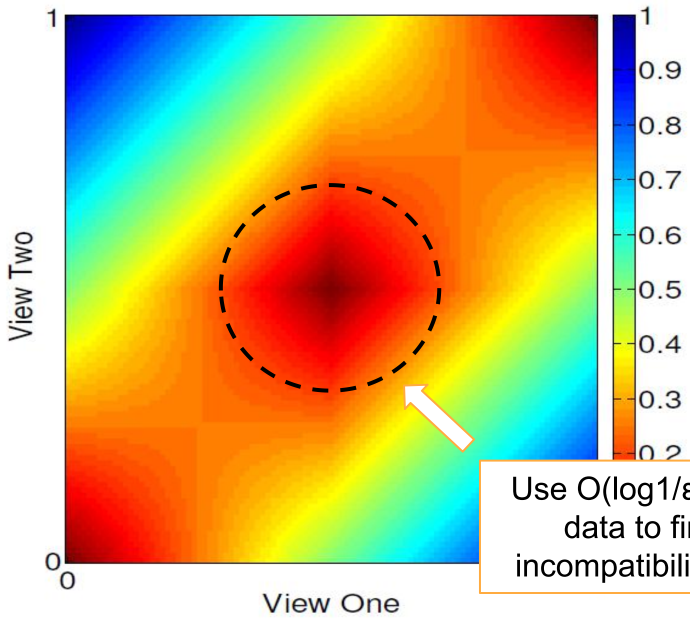
when error is small,
low incompatibility
implies low error

Error



Efficient Identification of Compatible Classifier

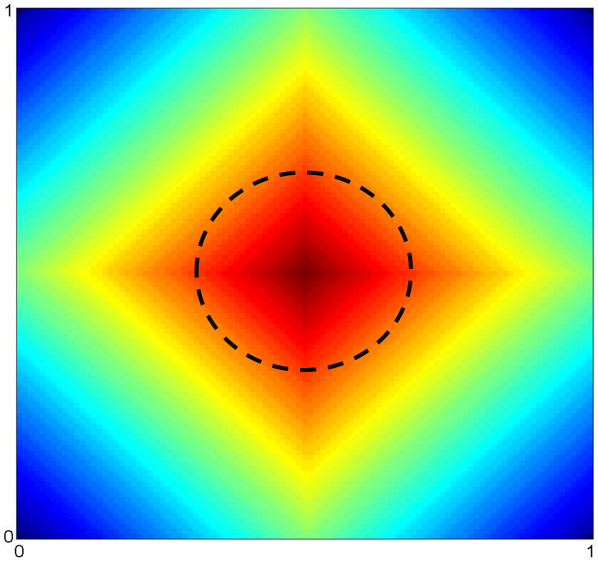
Incompatibility



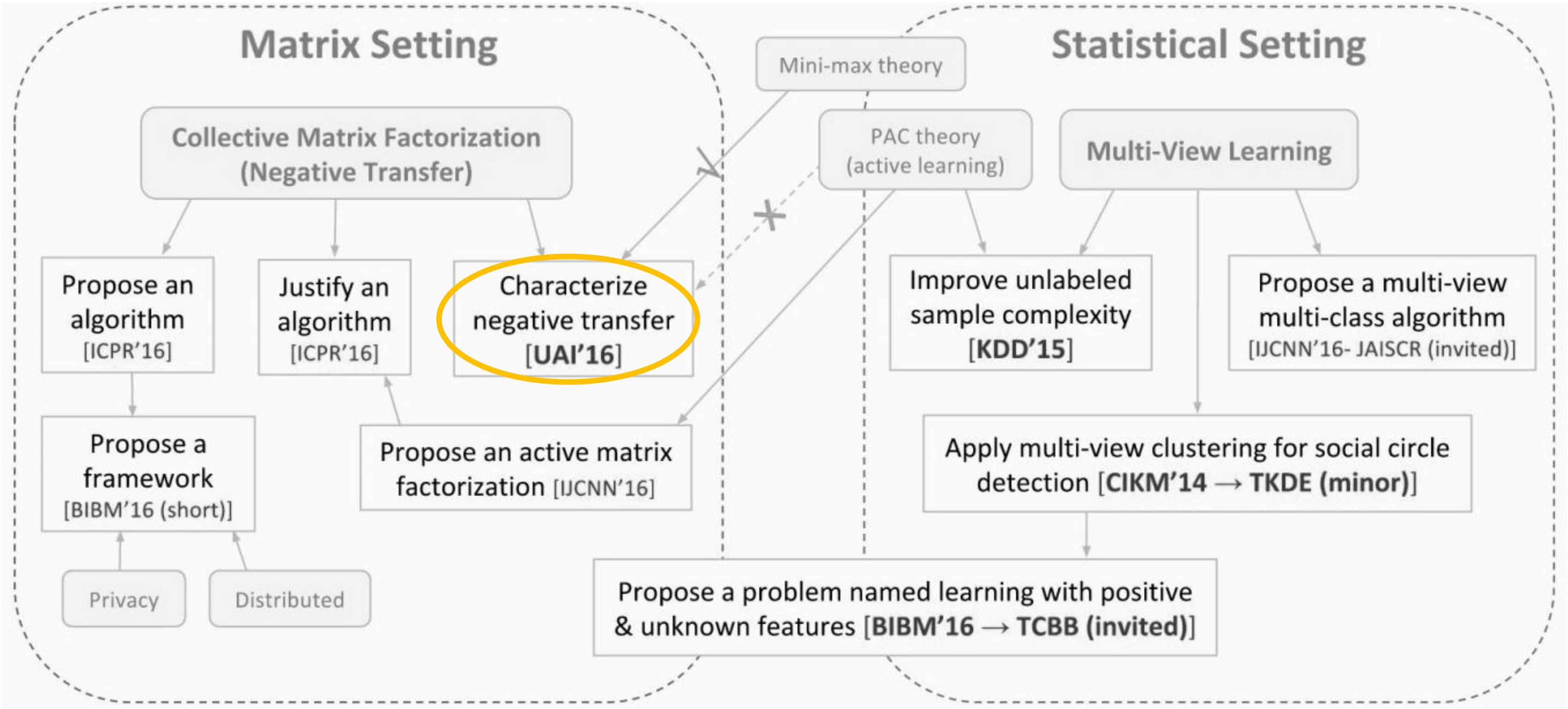
Use O(log1/ε) unlabeled data to find an ε-incompatibility classifier

when error is small, low incompatibility implies low error

Error

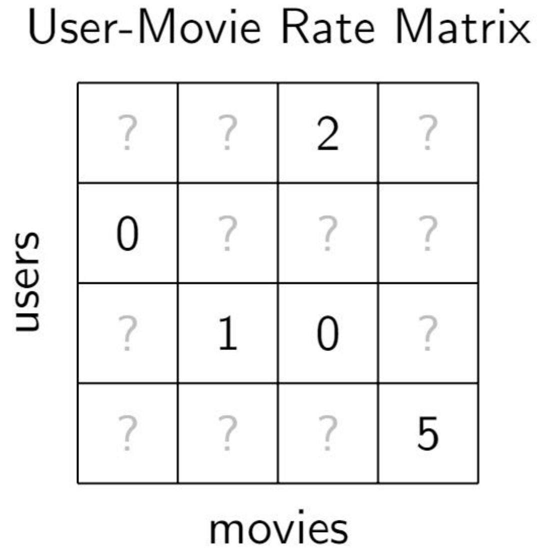


My PhD Research Overview



Background: Matrix Recovery

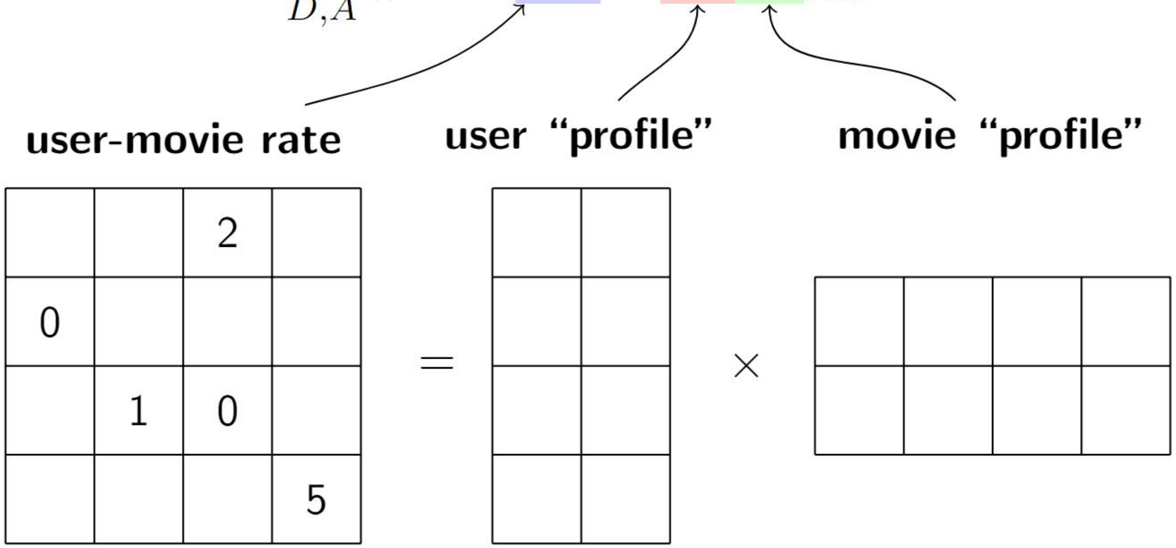
Matrix recovery problem studies how to recover missing values of a matrix.



Background: Matrix Factorization Technique

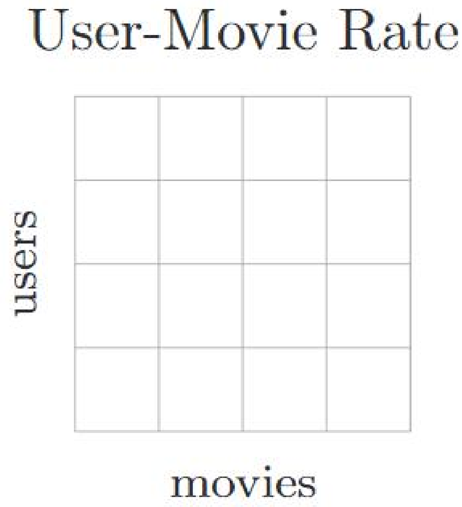
A classic recovery technique is matrix factorization.

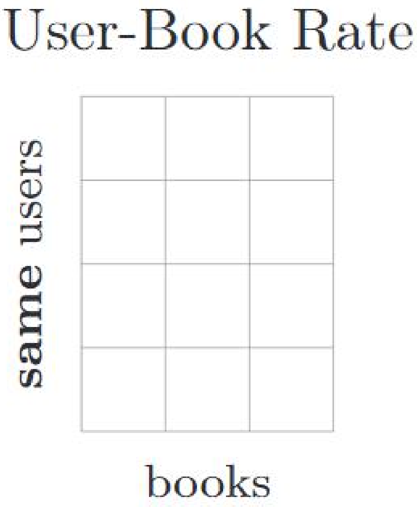
$$\min_{D,A} ||W \circ (M - DA)||_F^2.$$



Background: Multi-View Matrix Recovery Problem

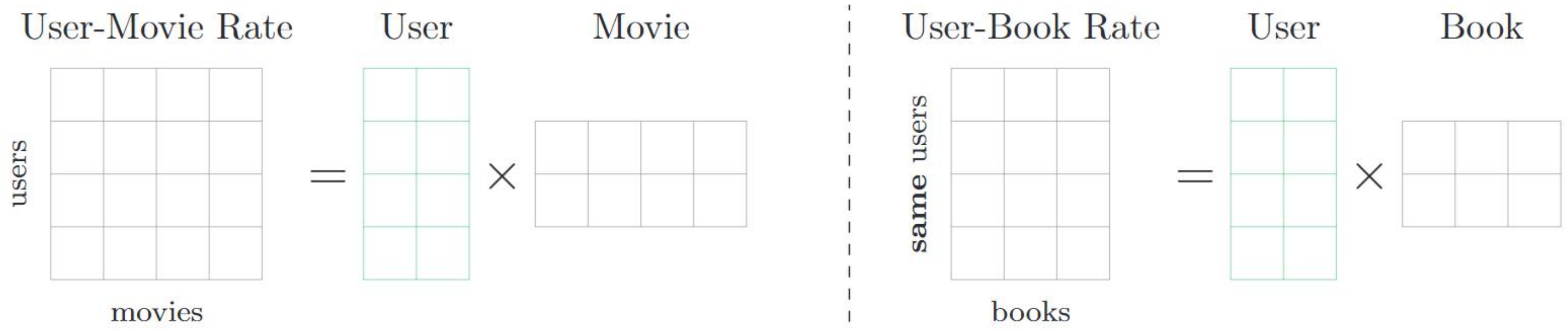
Each view has a sample matrix. How to recover missing values in all matrices?





Background: Collective Matrix Factorization (CMF)

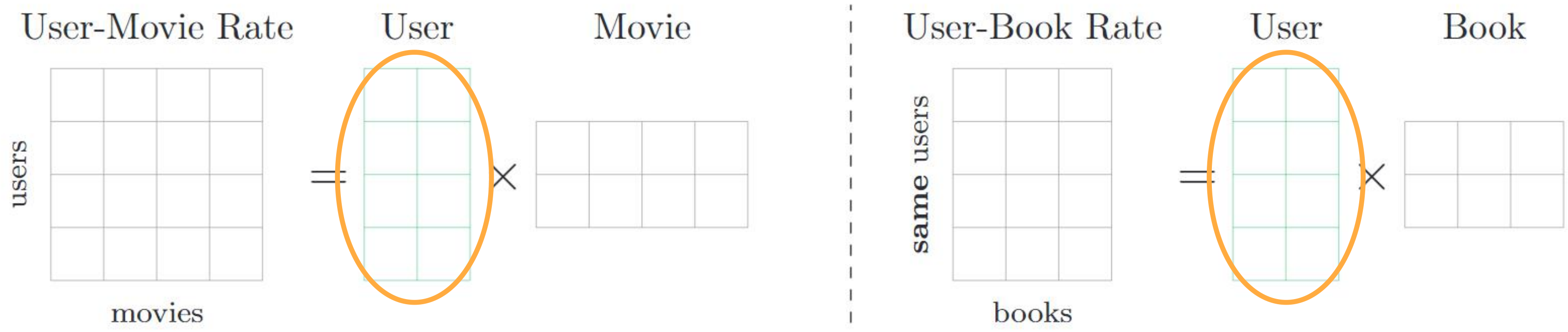
CMF recovers both matrices simultaneously by factorizing them jointly.



Signh and Gordon. Relational Learning via Collective Matrix Factorization. SIGKDD, 2008.

Background: Collective Matrix Factorization (CMF)

But the shared-factor assumption is hard to guarantee in reality.



Our Contribution

We present a first theoretical characterization of CMF performance when there is no guarantee on its shared-factor assumption.

1. if the shared-factor assumption is not guaranteed, there is a bias introduced in the mini-max lower bound of CMF estimator risk (and the bias depends only on the structure of hypothesis space, but not the estimator or sample).

2. if the shared-factor assumption is not guaranteed, there is a high-order root function introduced over the mini-max learning rate of the CMF estimator (and the order of the function depends on the structure of the hypothesis space).

Key Definitions

Let M, M' be two matrices, and M=DA, M'=D'A' be their factorizations.

A CMF estimator is $\hat{\theta} : \{\vec{M}_{\vec{\omega}}\} \to \mathbb{G}_k^n$, where $\mathbb{G}_k^n$ is hypothesis space.

Quality of the estimator is

$$\ell_{\vec{\omega}}(\hat{\theta}|\vec{M}) = \frac{1}{2}\left[\rho(\hat{\theta}_{\vec{\omega}}, \theta(M)) + \rho(\hat{\theta}_{\vec{\omega}}, \theta(M'))\right]$$

Maximum risk of the estimator is

$$\mathfrak{M}(\hat{\theta}) = \sup_{\vec{M}} \mathbb{E}_{\vec{\omega}} \ell_{\vec{\omega}}(\hat{\theta}|\vec{M})$$

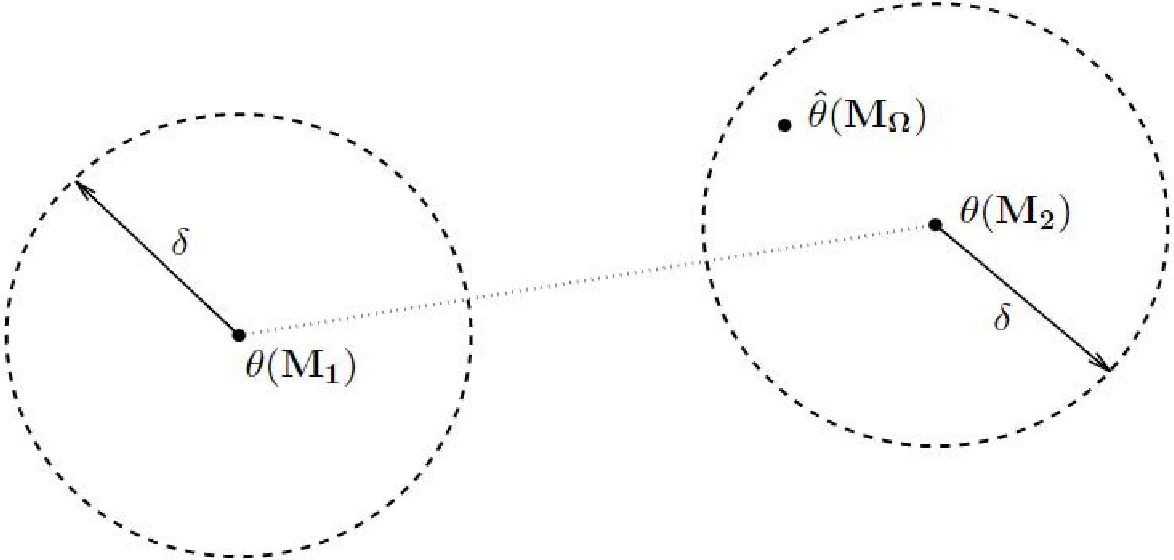
Result 1

Proposition 1. *Suppose $\mathbb{G}_k^n$ admits a 2δ-packing indexed by a finite set $\mathcal{V}$, and V is a uniform random variable on $\mathcal{V}$. Then, any CMF estimator $\hat{\theta}$ satisfies*

$$\mathfrak{M}(\hat{\theta}) \geq \frac{\delta}{2} \cdot \left(C_\delta + \frac{1}{|\mathcal{V}|} \Pr\{\hat{V}(\vec{M}_{\vec{\omega}}) \neq V\} \right), \qquad (9)$$

where $C_\delta = 1 - |\mathcal{V}|^{-1}$ and the probability is defined over the random choice of V and $\vec{M}_{\vec{\omega}}$ [3].

Insight to Prove Proposition 1



Result 2

Theorem 9. *Every CMF estimator $\hat{\theta}$ satisfies*

$$\mathfrak{M}(\hat{\theta}) \geq c \cdot \tau(\mathbb{G}_k^n)^{1-1/d}(|\vec{\omega}|\Sigma_A/\sigma^4)^{-1/d},$$

No guarantee of shared-factor assumption, mini-max rate is $\Omega(|\vec{\omega}|^{-1/d})$

Guarantee of shared-factor assumption, mini-max rate is $\Omega(|\vec{\omega}|^{-1})$

Techniques to Prove Theorem 9

Lemma 5. *Let* $\{M_v \in \mathbb{M}_k^{n\cdot}\}_{v\in\mathcal{V}} \subseteq \mathcal{P}$ *be a collection of matrices indexed by* $\mathcal{V}$ *such that for any* $v \neq v'$,

$$\rho(\theta(M_v), \theta(M_{v'})) \geq 2\delta. \tag{22}$$

Further, suppose

$$I(V; \vec{M}_{\vec{\omega}}) \leq \beta, \tag{23}$$

where V *is a uniform random variable on* $\mathcal{V}$. *Then*

$$\begin{aligned} \max_{v,v'\in\mathcal{V}} \mathbb{E}_\omega \frac{1}{2}\left(d(\hat{\theta}, \theta(M_v)) + d(\hat{\theta}, \theta(M_{v'}))\right) \\ \geq \frac{\delta}{2}\left(1 - \frac{\beta + \log 2}{|\mathcal{V}| \log |\mathcal{V}|}\right). \end{aligned} \tag{24}$$

Lemma 6. *There exist universal constants* $c_1, c_2 > 0$ *such that for any* $\delta \in (0, \tau(\mathbb{G}_k^n)]$,

$$(c_1\tau(\mathbb{G}_k^n)/\delta)^d \leq M(\mathbb{G}_k^n, \rho, \delta) \leq (c_2\tau(\mathbb{G}_k^n)/\delta)^d. \tag{26}$$

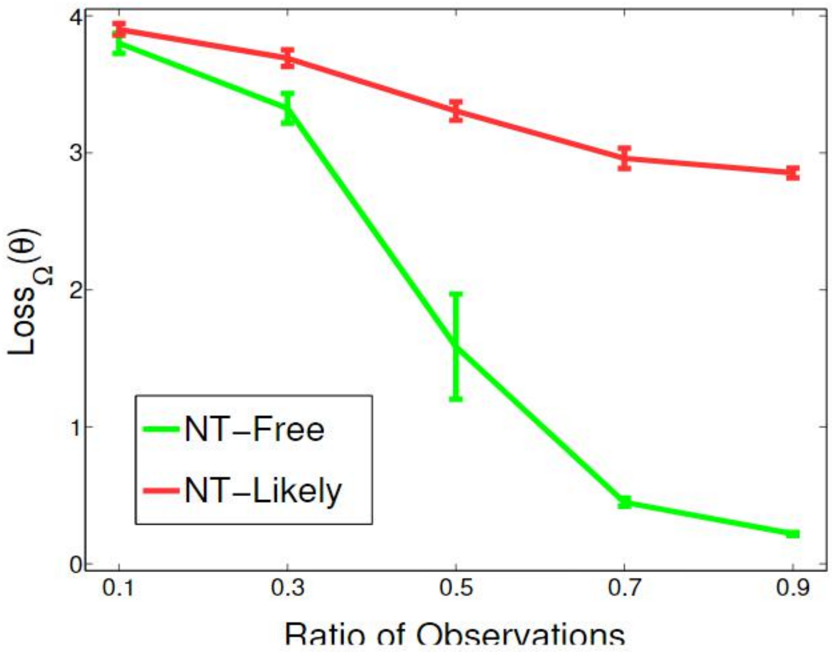
Lemma 7. *Let* $T(\vec{M}_{\vec{\omega}})$ *be any side information. Then*

$$I(V; \vec{M}_{\vec{\omega}}) \leq I(V; \vec{M}_{\vec{\omega}} | T(\vec{M}_{\vec{\omega}})). \tag{28}$$

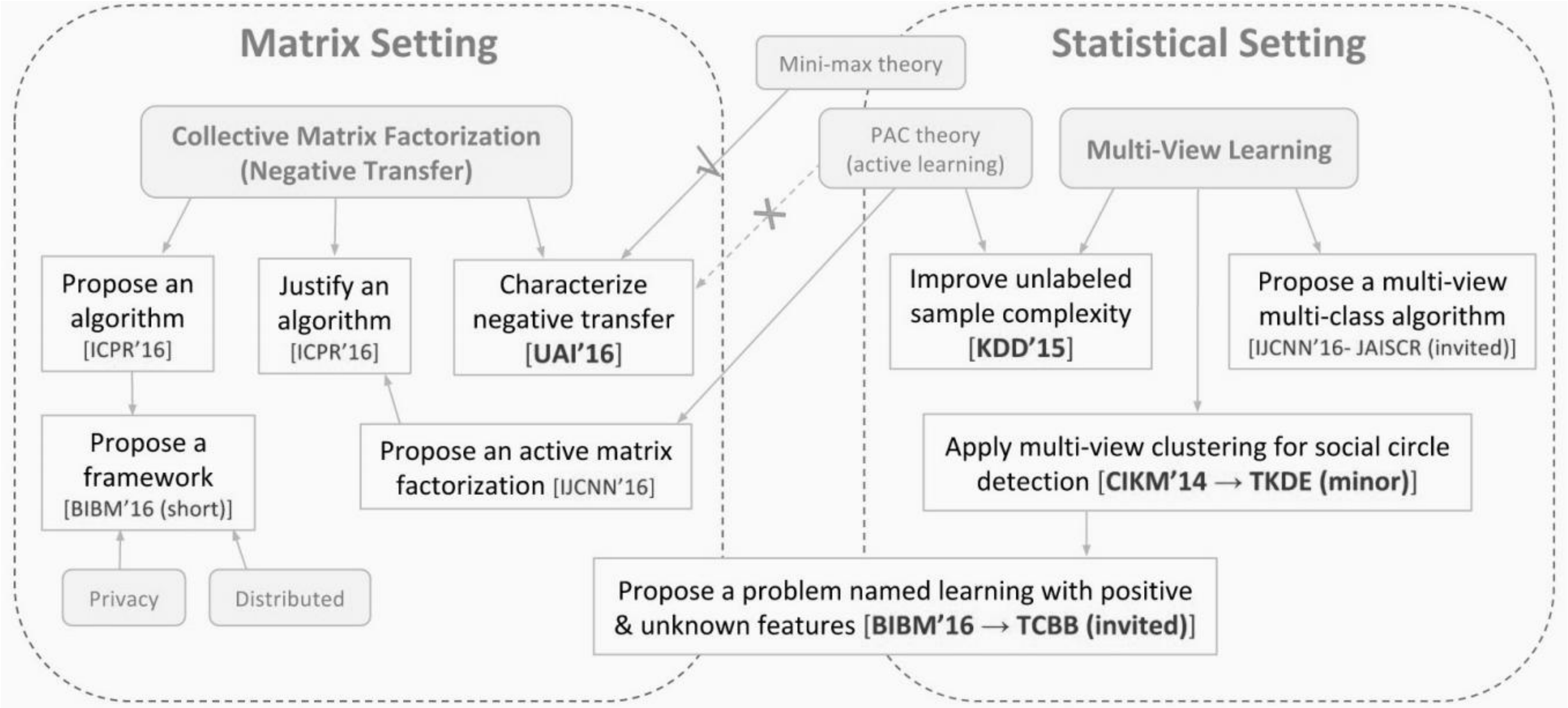
Lemma 8. *For any* $\bar{M}, \bar{M}' \in \mathbb{M}_k^{n,\vec{p}}$ *and* $\vec{\omega}$,

$$D_{k\ell}(\mathbb{P}_{\bar{M}|\vec{\omega}} || \mathbb{P}_{\bar{M}'|\vec{\omega}}) = \frac{1}{2\sigma^4}||W_{\vec{\omega}} \circ (\bar{M} - \bar{M}')||^2. \tag{30}$$

Simulation Result



Connecting Statistical Learning and Matrix Recovery



Outline

Introduction to multi-view learning

My contributions to multi-view learning

- statistical learning

- matrix recovery

A brief introduction to fairness problem in machine learning

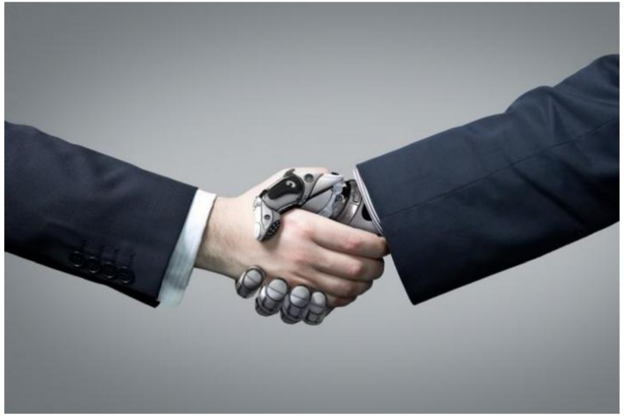
Our current research initiatives

Background

Machine learning is increasingly applied to assist consequential decision making.

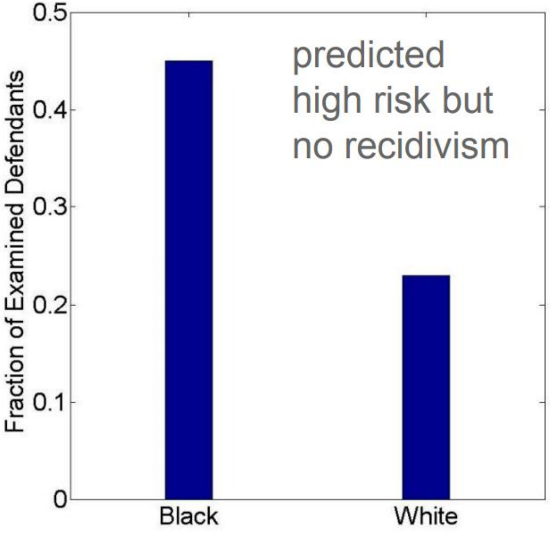
Background

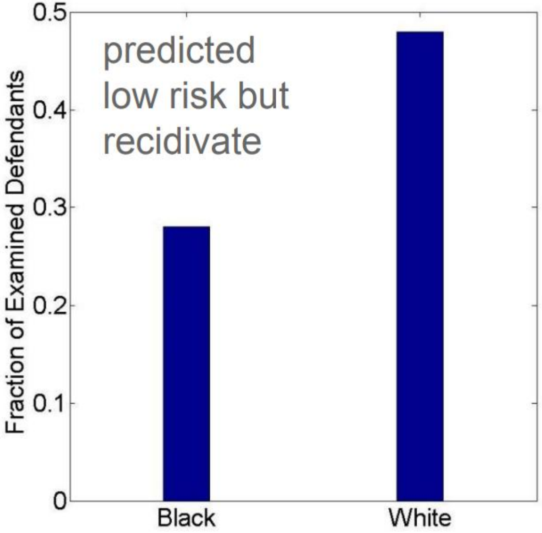
Algorithms are increasingly accountable for decision fairness.



Example 1: Recidivism Prediction

COMPAS recidivism prediction shows racial bias.

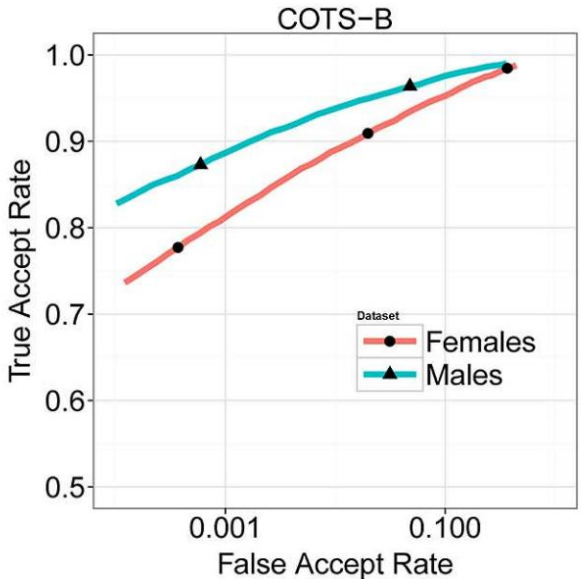


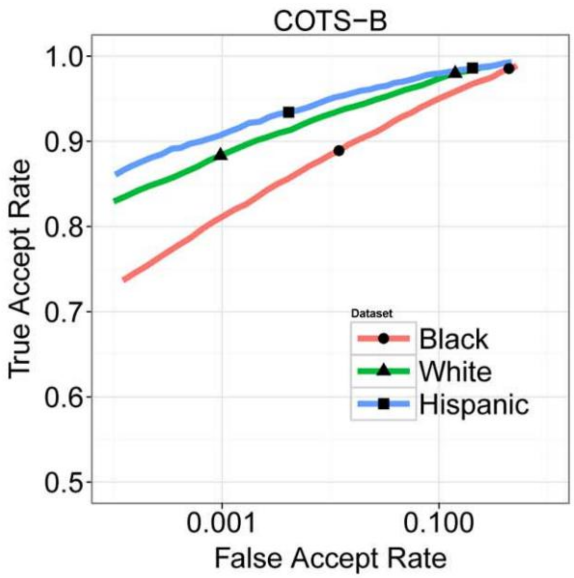


Larson et al. How we analyzed the COMPASS recidivism algorithm. ProPublica, 2016.

Example 2: Face Recognition

Face recognition systems show gender bias and racial bias.

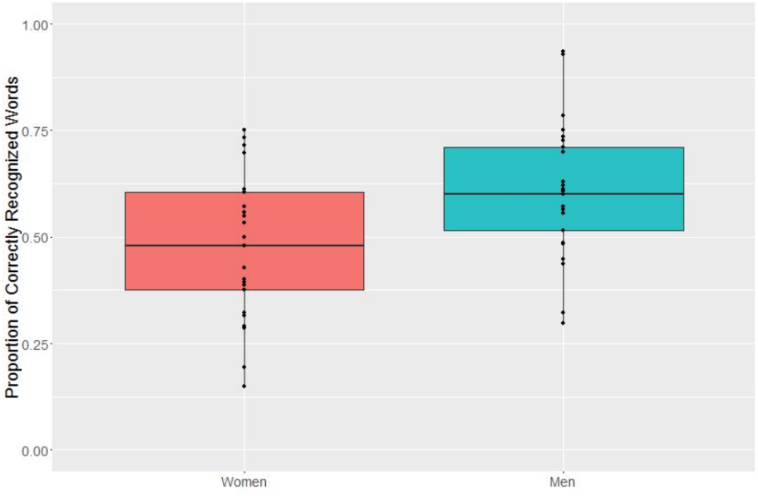




Klare et al. Face recognition performance: Role of demographic information. IEEE Trans. Information Forensics and Security, 2012.

Example 3: Speech Recognition

Google's speech recognition technology shows gender bias.



On average, for each female speaker less than half (47%) her words were captioned correctly. The average male speaker, on the other hand, was captioned correctly 60% of the time.

R. Tatman. Google's speech recognition has a gender bias, 2016.

Fairness-aware machine learning is an emerging field which aims to mitigate unfairness in model prediction.

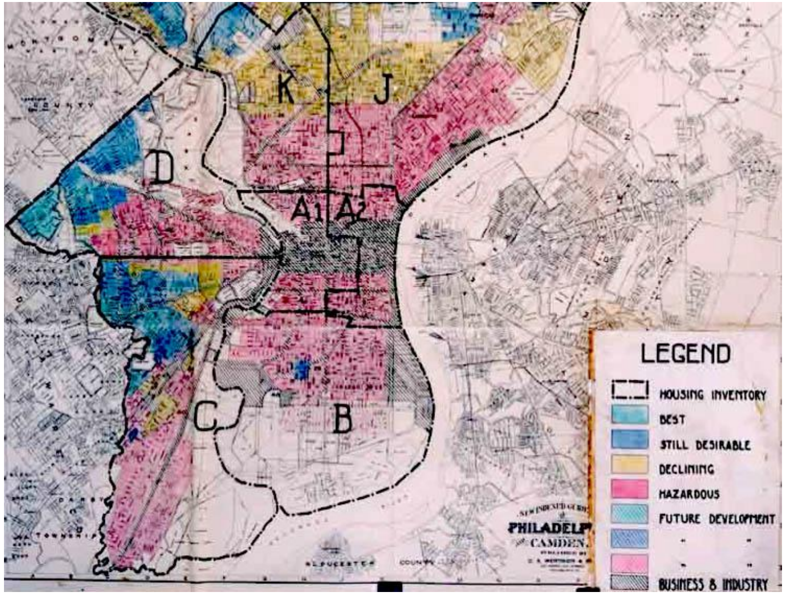
But combating unfairness in machine learning is non-trivial...

Suppose we want to mitigate gender bias...

What about hiding gender during resume filtering?

Redlining Effect

Redlining refers to the discriminatory practice of fencing off areas where [banks] would avoid investments based on community demographics.



W. Norton. Cultural geography: Environments, landscapes, identities, inequalities. Oxford University Press, 2013.

Outline

Introduction to multi-view learning

My contributions to multi-view learning

- statistical learning
- matrix recovery

A brief introduction to fairness problem in machine learning

Our current research initiatives

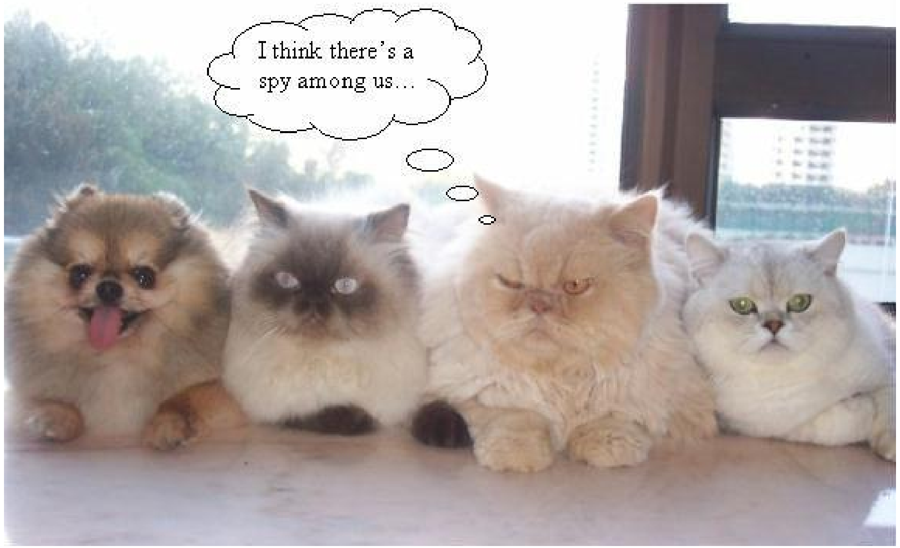
My Background

B.Eng. 2008 & M.Eng., 2011

Ph.D., 2017

AP, 2017-

Current Research Directions in Our Lab



Current Lab Members