Probabilistic Models for Permutations

Guy Lebanon
Georgia Institute of Technology

Outline

- Basic facts
- Models on permutations
- Models on with-ties and incomplete preferences
- Non-parametric approaches
- Important challenges and open problems

Basic Facts 1

- Permutations are bijections $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$
- Set of permutations forms the symmetric group $\mathfrak{S}_n$
- Rankings correspond to permutations mapping items to ranks
- With-ties ranking e.g., $1 \prec 2, 3 \prec 4$ correspond to cosets of the symmetric group $\mathfrak{S}_n \pi_0 \subset \mathfrak{S}_n$.
- Incomplete rankings e.g., $1 \prec 4$ correspond to a disjoint union of cosets $A \subset \mathfrak{S}_n$.

Basic Facts 1

- Permutations are bijections $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$
- Set of permutations forms the symmetric group $\mathfrak{S}_n$
- Rankings correspond to permutations mapping items to ranks
- With-ties ranking e.g., $1 \prec 2, 3 \prec 4$ correspond to cosets of the symmetric group $\mathfrak{S}_n \pi_0 \subset \mathfrak{S}_n$.
- Incomplete rankings e.g., $1 \prec 4$ correspond to a disjoint union of cosets $A \subset \mathfrak{S}_n$.

Basic Facts 1

- Permutations are bijections $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$
- Set of permutations forms the symmetric group $\mathfrak{S}_n$
- Rankings correspond to permutations mapping items to ranks
- With-ties ranking e.g., $1 \prec 2, 3 \prec 4$ correspond to cosets of the symmetric group $\mathfrak{S}_n \pi_0 \subset \mathfrak{S}_n$.
- Incomplete rankings e.g., $1 \prec 4$ correspond to a disjoint union of cosets $A \subset \mathfrak{S}_n$.

Basic Facts 1

- Permutations are bijections $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$
- Set of permutations forms the symmetric group $\mathfrak{S}_n$
- Rankings correspond to permutations mapping items to ranks
- With-ties ranking e.g., $1 \prec 2, 3 \prec 4$ correspond to cosets of the symmetric group $\mathfrak{S}_n \pi_0 \subset \mathfrak{S}_n$.
- Incomplete rankings e.g., $1 \prec 4$ correspond to a disjoint union of cosets $A \subset \mathfrak{S}_n$.

Basic Facts 1

- Permutations are bijections $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$
- Set of permutations forms the symmetric group $\mathfrak{S}_n$
- Rankings correspond to permutations mapping items to ranks
- With-ties ranking e.g., $1 \prec 2, 3 \prec 4$ correspond to cosets of the symmetric group $\mathfrak{S}_n \pi_0 \subset \mathfrak{S}_n$.
- Incomplete rankings e.g., $1 \prec 4$ correspond to a disjoint union of cosets $A \subset \mathfrak{S}_n$.

Basic Facts 2

- We have m permutations $D = \{\pi_1, \dots, \pi_m\}$ corresponding to preferences drawn from a population (people, computer programs, etc.)
- The population defines a distribution p_0 on permutations that is the main object of interest
- Censoring effect replaces permutations by with-ties or incomplete ratings $\pi_i \mapsto A_i \subset \mathfrak{S}_n$

$$q(A_i|\pi_i) \propto 1_{\{\pi_i \in A_1\}} q(\pi_1|A_1) q(A_1).$$

- Collaborative filtering example: users drawn from a population submitting censored versions (with-ties and incomplete) of their true but unknown preferences.

Basic Facts 2

- We have m permutations $D = \{\pi_1, \dots, \pi_m\}$ corresponding to preferences drawn from a population (people, computer programs, etc.)
- The population defines a distribution p_0 on permutations that is the main object of interest
- Censoring effect replaces permutations by with-ties or incomplete ratings $\pi_i \mapsto A_i \subset \mathfrak{S}_n$

$$q(A_i|\pi_i) \propto 1_{\{\pi_i \in A_1\}} q(\pi_1|A_1) q(A_1).$$

- Collaborative filtering example: users drawn from a population submitting censored versions (with-ties and incomplete) of their true but unknown preferences.

Basic Facts 2

- We have m permutations $D = \{\pi_1, \dots, \pi_m\}$ corresponding to preferences drawn from a population (people, computer programs, etc.)
- The population defines a distribution p_0 on permutations that is the main object of interest
- Censoring effect replaces permutations by with-ties or incomplete ratings $\pi_i \mapsto A_i \subset \mathfrak{S}_n$

$$q(A_i|\pi_i) \propto 1_{\{\pi_i \in A_1\}} q(\pi_1|A_1) q(A_1).$$

- Collaborative filtering example: users drawn from a population submitting censored versions (with-ties and incomplete) of their true but unknown preferences.

Basic Facts 2

- We have m permutations $D = \{\pi_1, \ldots, \pi_m\}$ corresponding to preferences drawn from a population (people, computer programs, etc.)
- The population defines a distribution p_0 on permutations that is the main object of interest
- Censoring effect replaces permutations by with-ties or incomplete ratings $\pi_i \mapsto A_i \subset \mathfrak{S}_n$

$$q(A_i|\pi_i) \propto 1_{\{\pi_i \in A_1\}} q(\pi_1|A_1) q(A_1).$$

- Collaborative filtering example: users drawn from a population submitting censored versions (with-ties and incomplete) of their true but unknown preferences.

Basic Facts 3

$$A_i \sim q(\cdot|\pi_i) \sim p_0$$

- The observed censored data $A_i \subset \mathfrak{S}_n$ typically increses in size as n increases.
- Estimate p_0 given the censored observations $A_1, \ldots, A_m$
- Some assumptions need to be made on censoring model q (censoring patterns are not systematic)

Basic Facts 3

$$A_i \sim q(\cdot|\pi_i) \sim p_0$$

- The observed censored data $A_i \subset \mathfrak{S}_n$ typically increses in size as n increases.
- Estimate p_0 given the censored observations $A_1, \ldots, A_m$
- Some assumptions need to be made on censoring model q (censoring patterns are not systematic)

Basic Facts 3

$$A_i \sim q(\cdot|\pi_i) \sim p_0$$

- The observed censored data $A_i \subset \mathfrak{S}_n$ typically increses in size as n increases.
- Estimate p_0 given the censored observations $A_1, \ldots, A_m$
- Some assumptions need to be made on censoring model q (censoring patterns are not systematic)

The Mallows Model for Permutations

- parametric location-spread model on fully ranked data

$$p_{\mu,c}(\pi) = \psi^{-1}(c)\, \exp\left(-c\, d(\pi,\mu)\right) \qquad \pi,\mu \in \mathfrak{S}_n \quad c \in \mathbb{R}_+$$

$d(\pi,\sigma)$ Kendall's tau

- Analogous to the normal distribution but lacks many of its nice properties
- Normalization term ψ has closed form and does not depend on the location parameter

The Mallows Model for Permutations

- parametric location-spread model on fully ranked data

$$p_{\mu,c}(\pi) = \psi^{-1}(c)\,\exp\left(-c\,d(\pi,\mu)\right) \qquad \pi,\mu \in \mathfrak{S}_n \quad c \in \mathbb{R}_+$$

$d(\pi,\sigma)$ Kendall's tau

- Analogous to the normal distribution but lacks many of its nice properties
- Normalization term ψ has closed form and does not depend on the location parameter

The Mallows Model for Permutations

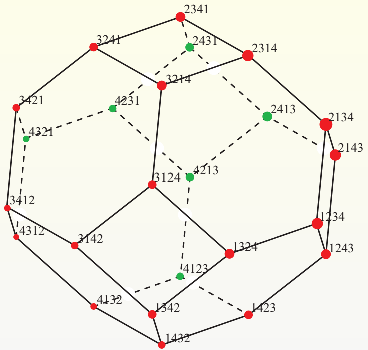
- parametric location-spread model on fully ranked data

$$p_{\mu,c}(\pi) = \psi^{-1}(c)\,\exp\left(-c\,d(\pi,\mu)\right) \qquad \pi,\mu \in \mathfrak{S}_n \quad c \in \mathbb{R}_+$$

$d(\pi,\sigma)$ Kendall's tau

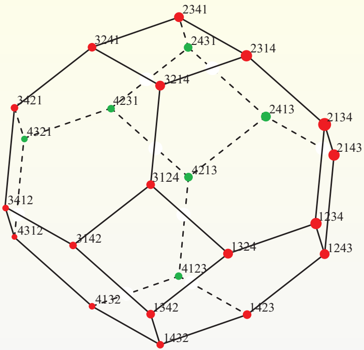
- Analogous to the normal distribution but lacks many of its nice properties
- Normalization term ψ has closed form and does not depend on the location parameter

Mallows Model and the Permutation Polytope



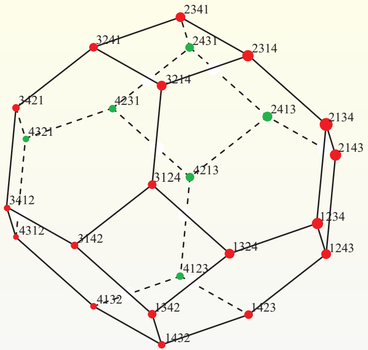
- The Mallows model is often unrealistic and impractical
- Unimodal parametric shape is too restricted
- MLE involves impossible discrete search (for large n)
- Many extensions have been proposed by exploring other exponential forms (Babington Smith, Bradley Terry, etc.).

Mallows Model and the Permutation Polytope



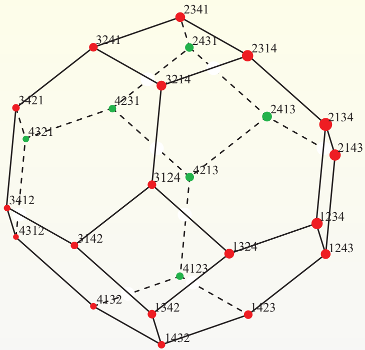
- The Mallows model is often unrealistic and impractical
- Unimodal parametric shape is too restricted
- MLE involves impossible discrete search (for large n)
- Many extensions have been proposed by exploring other exponential forms (Babington Smith, Bradley Terry, etc.).

Mallows Model and the Permutation Polytope



- The Mallows model is often unrealistic and impractical
- Unimodal parametric shape is too restricted
- MLE involves impossible discrete search (for large n)
- Many extensions have been proposed by exploring other exponential forms (Babington Smith, Bradley Terry, etc.).

Mallows Model and the Permutation Polytope



- The Mallows model is often unrealistic and impractical
- Unimodal parametric shape is too restricted
- MLE involves impossible discrete search (for large n)
- Many extensions have been proposed by exploring other exponential forms (Babington Smith, Bradley Terry, etc.).

Luce-Plackett/Thurstone Models

A different approach: multi-stage ranking models

- top ranked item is sampled

$$p(\pi(i) = 1) = \nu_i / \sum_{j=1}^{n} \nu_j$$

- given identity of items ranked $1, \ldots, j-1$, the next ranked item is sampled from the conditional distribution

$$p(\pi(k) = j | \pi(i_1) = 1, \ldots, \pi(i_{j-1}) = j-1) = \nu_k / \sum_{l : j \notin \{i_1, \ldots, i_{j-1}\}} \nu_l$$

Each item is assigned a parameter controlling its popularity.

Handling With-Ties and Incomplete Rankings

Approach 1: define new distance or dissimilarity functions on with-ties and incomplete rankings and proceed with distance based models e.g., Mallows model

- Expected distance with respect to distribution r

$$d^*(A\,;B) = \frac{1}{|A|\cdot|B|}\sum_{\pi\in A}\sum_{\sigma\in B} r(\pi)r(\sigma)d(\pi,\sigma)$$

- Hausdorff distance

$$d^*(A,B) = \max\left\{\max_{\pi\in A}\min_{\sigma\in B} d(\pi,\sigma), \max_{\sigma\in B}\min_{\pi\in A} d(\pi,\sigma)\right\}$$

In both cases d^* can be efficiently computed but resulting models lack interpretation

Handling With-Ties and Incomplete Rankings 2

Assume observed data $A_i \subset \mathfrak{S}_n$ is a censored form of $\pi_i \in A_i$. Proceed with standard estimation techniques for missing data (observed likelihood, etc.).

- Strong interpretation
- What are appropriate assumptions on censoring model?
- Estimation and inference often intractable

What Does the Data Look Like

- Using

$$T^*(A, B) = \frac{1}{|A| \cdot |B|} \sum_{\pi \in A} \sum_{\sigma \in B} T(\pi, \sigma)$$

embed $\mathcal{D} = \{A_1, \ldots, A_m\}$ by multidimensional scaling $h : (\mathfrak{S}_n, T^*) \to (\mathbb{R}^2, \| \cdot \|_2)$ in order to minimize distortion

$$R(h) = \sum_{i,j} (T^*(A_i, A_j) - \|h(A_i) - h(A_j)\|)^2.$$

- Estimate density of embedded points $\{h(A_1), \ldots, h(A_m)\}$ using kernel density estimation in $(\mathbb{R}^2, \| \cdot \|_2)$.

What Does the Data Look Like

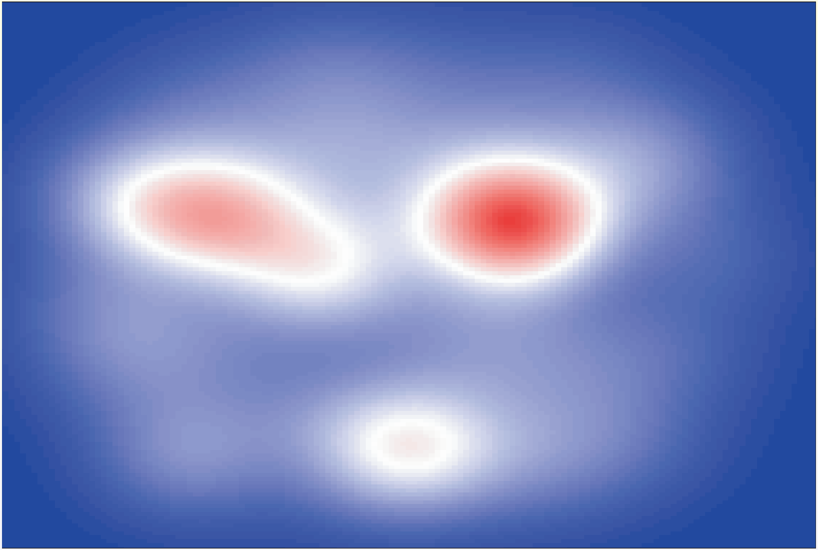
- Using

$$T^*(A,B) = \frac{1}{|A|\cdot|B|}\sum_{\pi\in A}\sum_{\sigma\in B} T(\pi,\sigma)$$

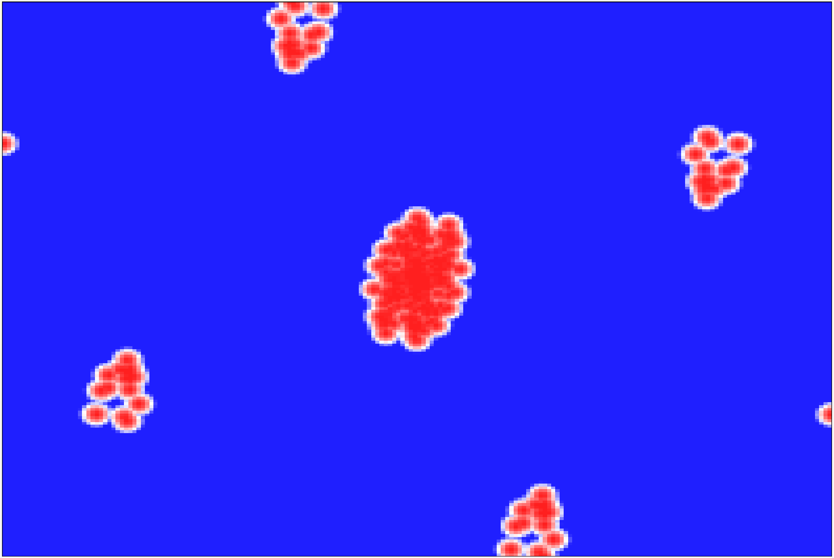
embed $\mathcal{D} = \{A_1,\ldots,A_m\}$ by multidimensional scaling $h : (\mathfrak{S}_n, T^*) \to (\mathbb{R}^2, \|\cdot\|_2)$ in order to minimize distortion

$$R(h) = \sum_{i,j}(T^*(A_i,A_j) - \|h(A_i) - h(A_j)\|)^2.$$

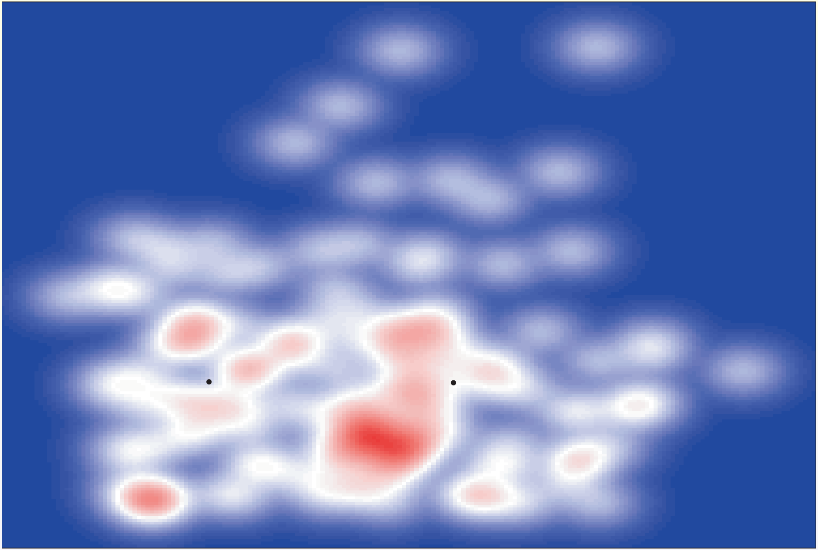
- Estimate density of embedded points $\{h(A_1),\ldots,h(A_m)\}$ using kernel density estimation in $(\mathbb{R}^2, \|\cdot\|_2)$.



APA votes



Jester



Movie Ranking

Non-Parametric Smoothing

- NP alternative that does not involve parametric optimization

$$\hat{p}(\pi) = \frac{1}{m}\sum_{j=1}^{m} K_h(\pi, \pi_i) = \frac{1}{m\,\psi(c)}\sum_{j=1}^{m} \exp\left(-cd(\pi, \pi_i)\right)$$

$$\hat{p}(\mathfrak{S}_\lambda \pi) = \frac{1}{m\,\psi(c)}\sum_{j=1}^{m}\sum_{\tau\in\mathfrak{S}_\lambda\pi} \exp\left(-cd(\tau, \pi_i)\right).$$

- Partially ranked training data $\{\mathfrak{S}_{\gamma_1}\pi_1, \ldots, \mathfrak{S}_{\gamma_m}\pi_m\}$ may be expressed as a latent variable (say MCAR uniformly)

$$\hat{p}(\mathfrak{S}_\lambda \pi) = \frac{1}{m\,\psi(c)}\sum_{i=1}^{m}\frac{1}{|\mathfrak{S}_{\gamma_i}|}\sum_{\mu\in\mathfrak{S}_\lambda\pi}\sum_{\tau\in\mathfrak{S}_{\gamma_i}\pi_i} \exp(-c\,d(\mu, \tau))$$

Non-Parametric Smoothing

- NP alternative that does not involve parametric optimization

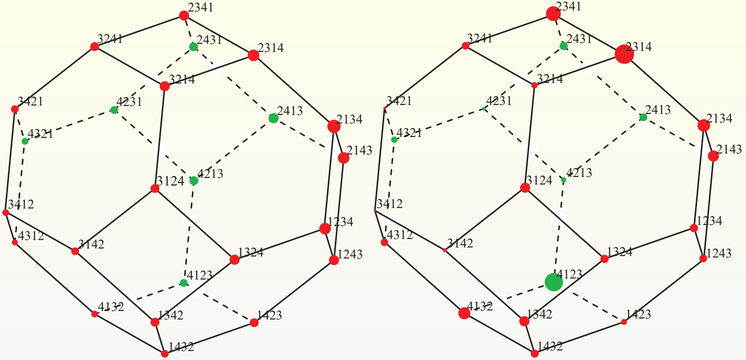
$$\hat{p}(\pi) = \frac{1}{m}\sum_{j=1}^{m} K_h(\pi,\pi_i) = \frac{1}{m\,\psi(c)}\sum_{j=1}^{m}\exp\left(-cd(\pi,\pi_i)\right)$$

$$\hat{p}(\mathfrak{S}_\lambda\pi) = \frac{1}{m\,\psi(c)}\sum_{j=1}^{m}\sum_{\tau\in\mathfrak{S}_\lambda\pi}\exp\left(-cd(\tau,\pi_i)\right).$$

- Partially ranked training data $\{\mathfrak{S}_{\gamma_1}\pi_1,\ldots,\mathfrak{S}_{\gamma_m}\pi_m\}$ may be expressed as a latent variable (say MCAR uniformly)

$$\hat{p}(\mathfrak{S}_\lambda\pi) = \frac{1}{m\,\psi(c)}\sum_{i=1}^{m}\frac{1}{|\mathfrak{S}_{\gamma_i}|}\sum_{\mu\in\mathfrak{S}_\lambda\pi}\sum_{\tau\in\mathfrak{S}_{\gamma_i}\pi_i}\exp(-c\,d(\mu,\tau))$$

Mallows Model vs. NP Smoothing



Visualizing estimated probabilities for EachMovie data by permutation polytopes: Mallows model (left) and non-parametric model for $c = 2$ (right). The Mallows model locates a single mode at $2|1|3|4$ while the non-parametric estimator locates the global mode at $2|3|1|4$ and a second local mode at $4|1|2|3$.

Open Problems 1

Censoring model $\pi \mapsto A_i$ is typically unknown

$$q(A|\pi) \propto 1_{\{\pi \in A\}} q(\pi|A) q(A).$$

- Estimate $q(\pi|A)$ and $q(A)$ from data.
- What is the relationship between q and estimation accuracy (asymptotic variance, conditions on consistency)
- In survey design q is determined by the survey policy. When designing a survey, what tie or incomplete structures should be chosen?

Open Problems 1

Censoring model $\pi \mapsto A_i$ is typically unknown

$$q(A|\pi) \propto 1_{\{\pi \in A\}} q(\pi|A) q(A).$$

- Estimate $q(\pi|A)$ and $q(A)$ from data.
- What is the relationship between q and estimation accuracy (asymptotic variance, conditions on consistency)
- In survey design q is determined by the survey policy. When designing a survey, what tie or incomplete structures should be chosen?

Open Problems 1

Censoring model $\pi \mapsto A_i$ is typically unknown

$$q(A|\pi) \propto 1_{\{\pi \in A\}} q(\pi|A) q(A).$$

- Estimate $q(\pi|A)$ and $q(A)$ from data.
- What is the relationship between q and estimation accuracy (asymptotic variance, conditions on consistency)
- In survey design q is determined by the survey policy. When designing a survey, what tie or incomplete structures should be chosen?

Open Problems 2

Models based on Kendall's tau or similar distances are rank-symmetric

$$d(1|2|3|4, 2|1|3|4) = d(1|2|3|4, 1|2|4|3).$$

- Items at top ranks may be more important to match than at bottom ranks.
- Items at top and bottom ranks may be more important to match than middle ranks.
- Develop models with non symmetric distances such that when learned from data will be more accurate in the correct ranks e.g., shortest path on polytope with certain weight structure.

Open Problems 2

Models based on Kendall's tau or similar distances are rank-symmetric

$$d(1|2|3|4, 2|1|3|4) = d(1|2|3|4, 1|2|4|3).$$

- Items at top ranks may be more important to match than at bottom ranks.
- Items at top and bottom ranks may be more important to match than middle ranks.
- Develop models with non symmetric distances such that when learned from data will be more accurate in the correct ranks e.g., shortest path on polytope with certain weight structure.

Open Problems 2

Models based on Kendall's tau or similar distances are rank-symmetric

$$d(1|2|3|4, 2|1|3|4) = d(1|2|3|4, 1|2|4|3).$$

- Items at top ranks may be more important to match than at bottom ranks.
- Items at top and bottom ranks may be more important to match than middle ranks.
- Develop models with non symmetric distances such that when learned from data will be more accurate in the correct ranks e.g., shortest path on polytope with certain weight structure.

Open Problems 3

- Permutation models have mostly ignored covariate information.
- For example in movie recommendation
  - rater covariates: age and gender of rater in movie recommendations
  - item covariates: genre, director, year, etc. in movie recommendations

  Develop models that take one or both forms of covariates into account.

Thank You!

Collaborators: Bill Cleveland, Josh Dillon, Paul Kidwell, Yi Mao

- *IEEE Trans on Visualization and Computer Graphics* 14(6) 2008
- *Journal of Machine Learning Research* 9, 2008
- *Advances in Neural Information Processing Systems* 20, 2008