Learning preferences with irrelevant alternatives

Joint work with Alex Peysakhovich, Stephen Ragain, and Arjun Seshadri

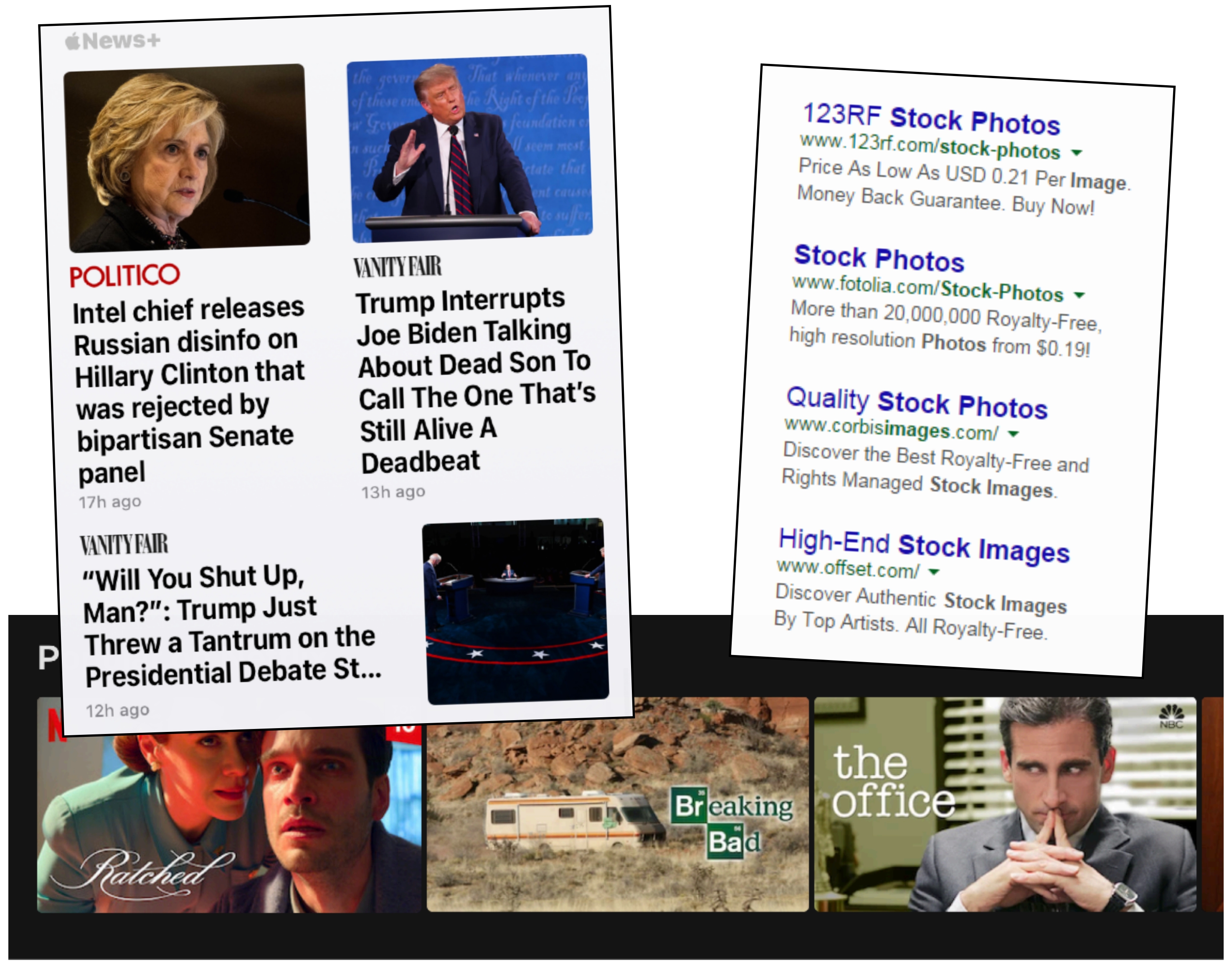
Johan Ugander, Stanford

Texas A&M Institute for Data Science, September 27, 2021

Preferences over sets

(McFadden, 1974)

(Manski, 1976)



- Given a universe set $\mathcal{X}$, consider a choice set $C \subseteq \mathcal{X}$. What do you choose?
- **Discrete choice**: learning distributions over items, for all sets $C \subseteq \mathcal{X}$.
- **Ranking**: distributions over permutations of $\mathcal{X}$.

Agenda

- **Choice systems** as mathematical objects.
- The **independence of irrelevant alternatives** (IIA) in discrete choice.

- Tractable **choice models** that forego IIA. (ICML 2019)
- Tractable **rankings models** that forego IIA. (NeurIPS 2020)

- When does data obey IIA? Lower bounds on **hypothesis testing.** (EC 2019)

Probabilistic discrete choice

- Focuses on a peculiar mathematical space, **choice systems**.

$\mathbb{R}^n$ $\{0,1\}^{n\times n}$ S_n

P_n

S^3

Δ^n C^k $\mathcal{T}_n$

Probabilistic discrete choice

- Focuses on a peculiar mathematical space, **choice systems**.
- Let $P_{x,C}$ denote the probability of choosing x from C.

- **Definition**: Conditional choice system (Falmagne, 1978):

$$\{P_{x,C}\}_{\forall C \subseteq \mathcal{X}, \forall x \in C}$$

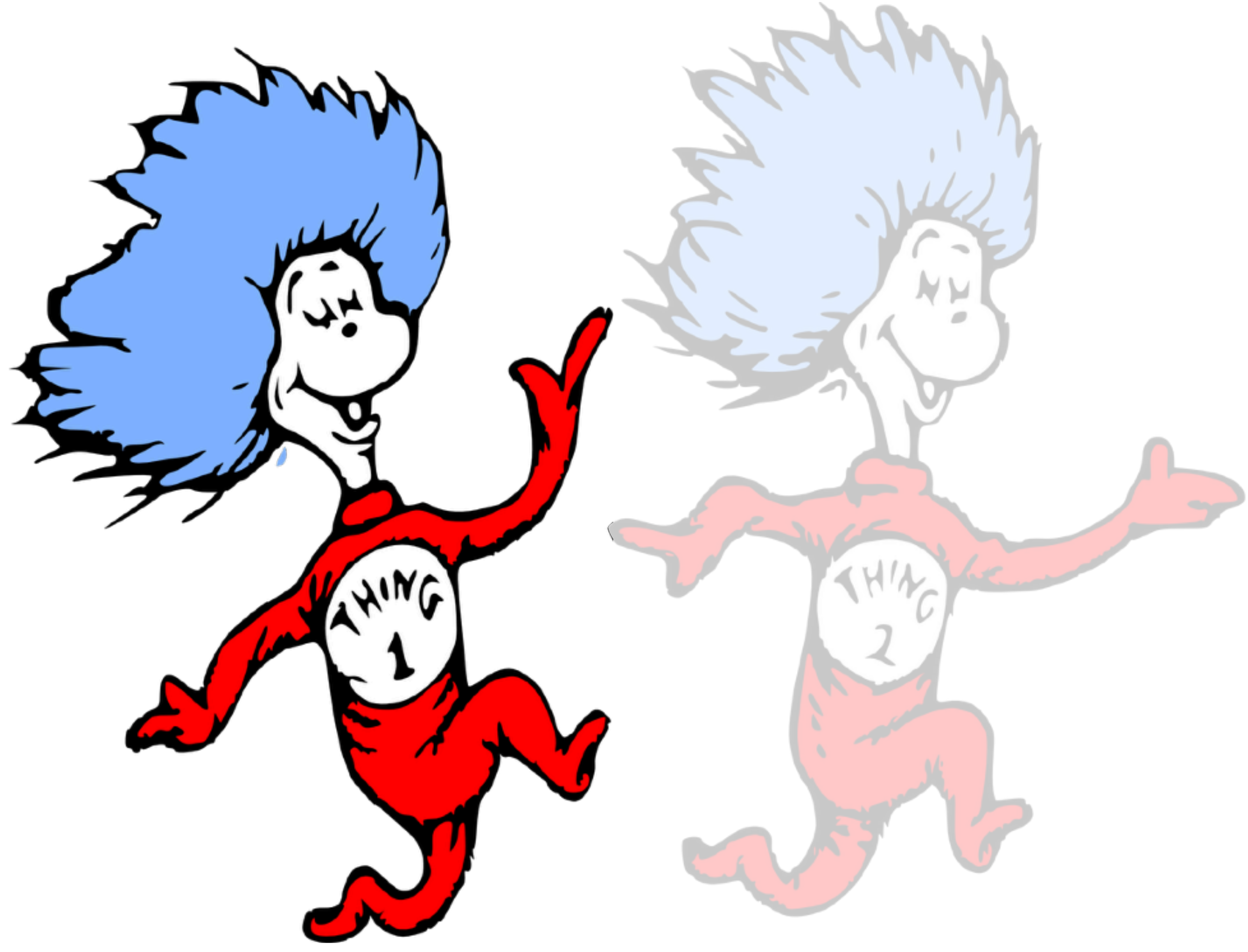
Probabilistic discrete choice

- Focuses on a peculiar mathematical space, **choice systems**.
- Let $P_{x,C}$ denote the probability of choosing x from C,
- **Definition**: Conditional choice system (Falmagne, 1978):

$$\{P_{x,C}\}_{\forall C \subseteq \mathcal{X}, \forall x \in C}$$

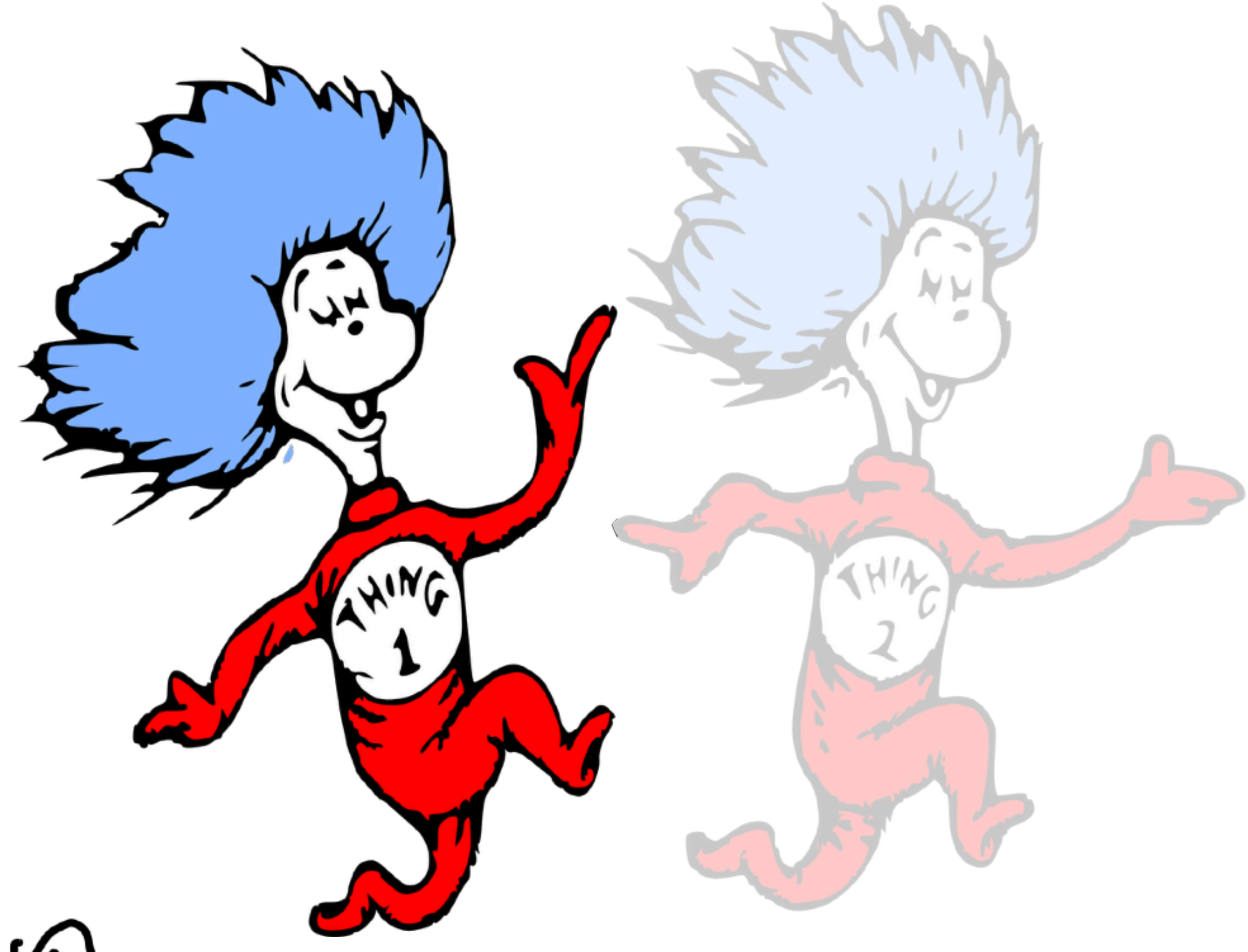
- Let $w(C)$ denote the probability of *choosing from* $C \subseteq \mathcal{X}$. Features in "unconditional choice system", not part of this talk.

Probabilistic discrete choice



- Consider $\mathcal{X} = \{a, b, c\}$. What is $\{P_{x,C}\}_{\forall C \subseteq \mathcal{X}, \forall x \in C}$?

Probabilistic discrete choice



- Consider $\mathcal{X} = \{a, b, c\}$. What is $\{P_{x,C}\}_{\forall C \subseteq \mathcal{X}, \forall x \in C}$?

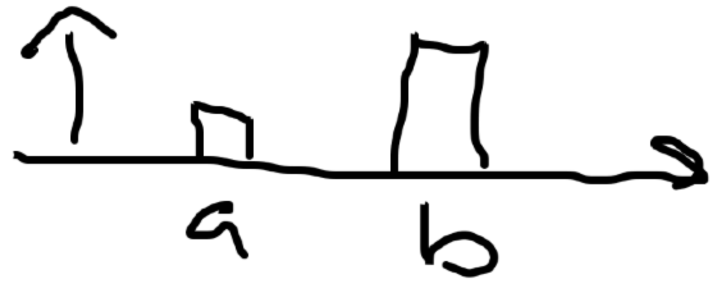
$C_1 = \{a, b\}$

$C_2 = \{b, c\}$

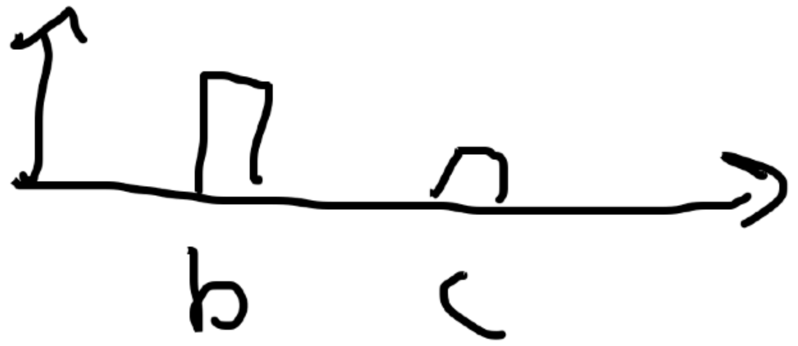
$C_3 = \{a, c\}$

$C_4 = \{a, b, c\}$

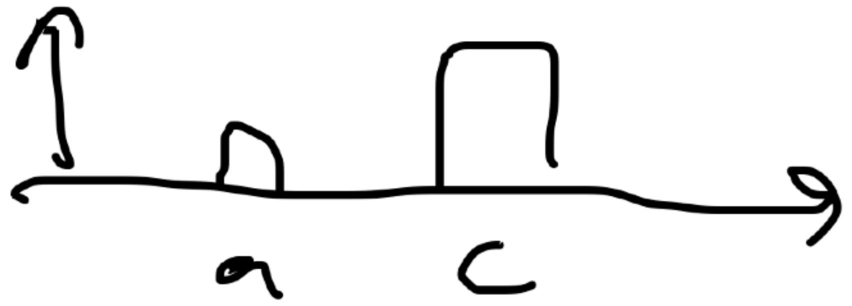
P_{x,C_1}:



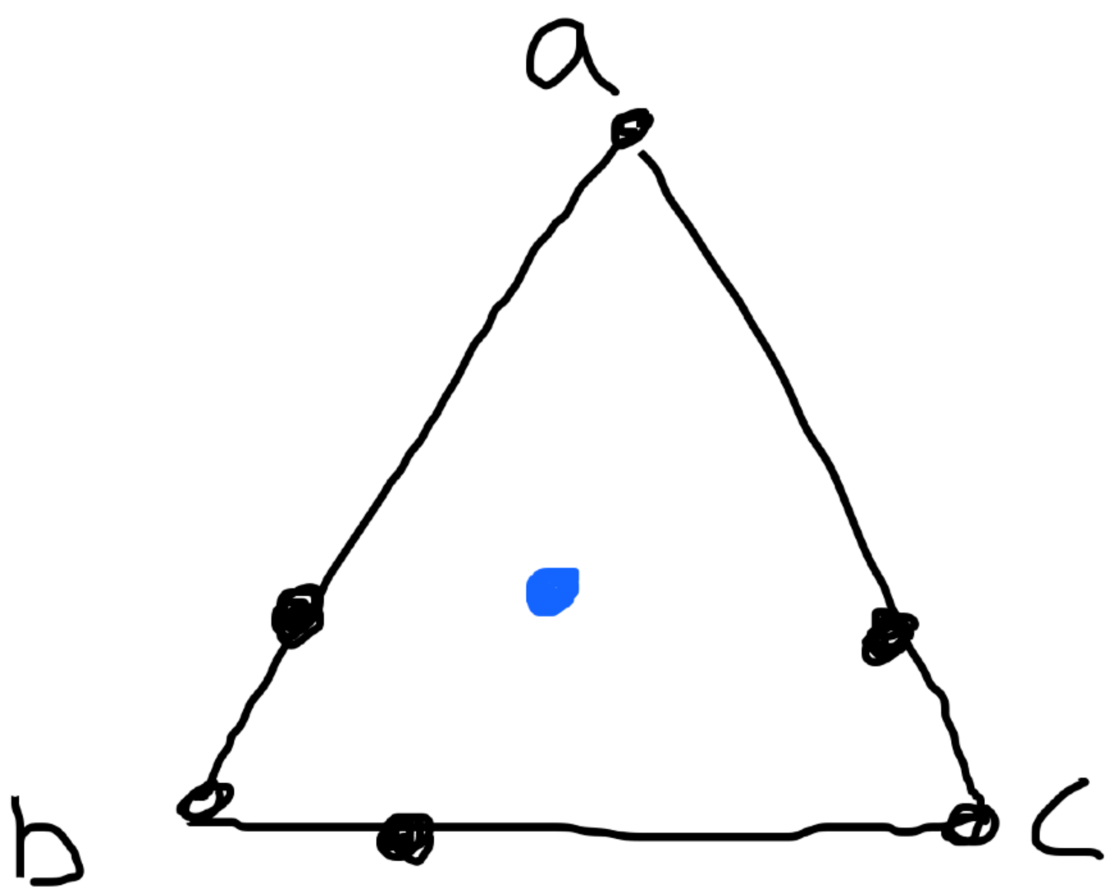
P_{x,C_2}:



P_{x,C_3}:



P_{x,C_4}:



Independence of Irrelevant Alternatives (IIA)

- Arbitrary choice systems (i.e., McFadden's *universal logit*) make no assumptions about the relationship between distributions on different sets.

- IIA (Luce, 1959): For every $x \in \mathcal{X}$, $C \subseteq \mathcal{X}$:

$$\frac{P_{x,\{x,y\}}}{P_{y,\{x,y\}}} = \frac{P_{x,\{x,y\}\cup C}}{P_{y,\{x,y\}\cup C}}.$$

- Consequence: the ratio between x and y stays the same, no matter what "**irrelevant alternatives**" you add to the choice set.

Independence of Irrelevant Alternatives (IIA)

- Arbitrary choice systems (i.e., McFadden's *universal logit*) make no assumptions about the relationship between distributions on different sets.

- IIA (Luce, 1959): For every $x \in \mathcal{X}$, $C \subseteq \mathcal{X}$:

$$\frac{P_{x,\{x,y\}}}{P_{y,\{x,y\}}} = \frac{P_{x,\{x,y\}\cup C}}{P_{y,\{x,y\}\cup C}}.$$

- Consequence: the ratio between x and y stays the same, no matter what "**irrelevant alternatives**" you add to the choice set.

- Models obeying IIA admit a **ratio representation:**

$$P_{x,C} = \frac{\gamma_x}{\sum_{z\in C}\gamma_z}, \forall C \subseteq \mathcal{X}, \forall x \in C.$$

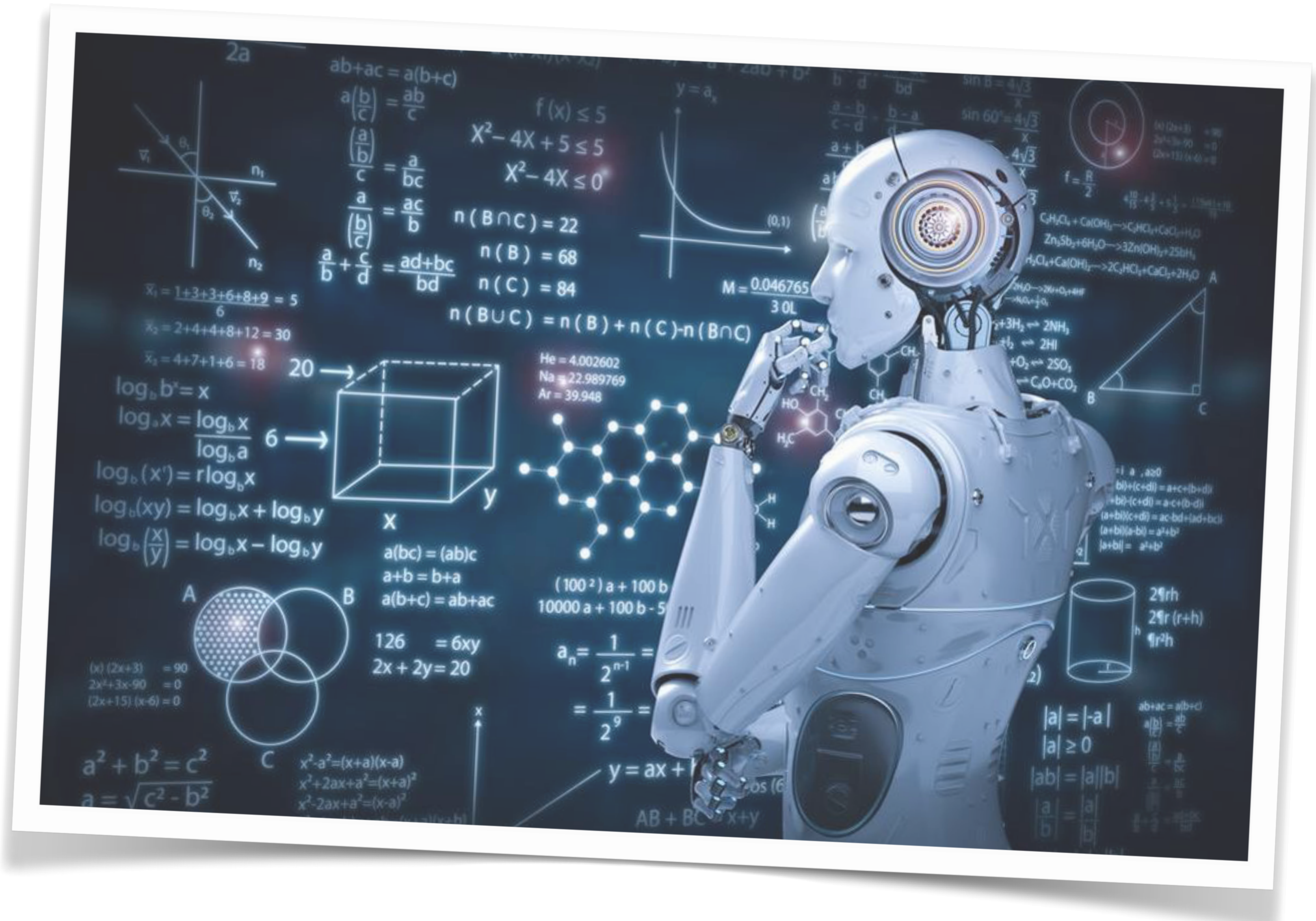
Independence of Irrelevant Alternatives (IIA)

- Assuming IIA $\Rightarrow$ **Multinomial Logit (MNL)** model of discrete choice:

$$P_{x,C} = \frac{\exp(u_x)}{\sum_{z \in C} \exp(u_z)}.$$

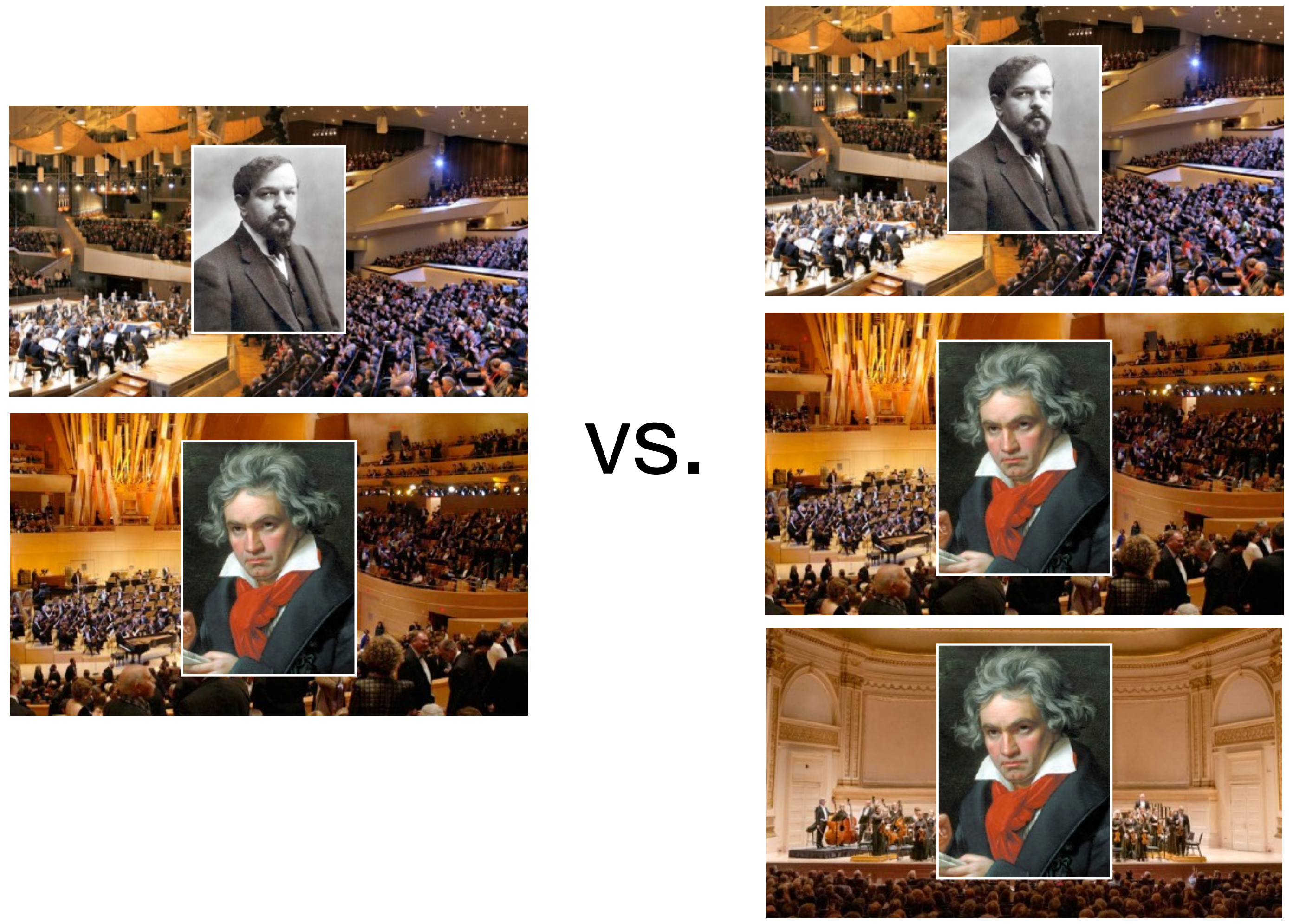
- Major workhorse of modern machine learning
- If $u_x = \beta^T f_x$, linear model

#IJALM

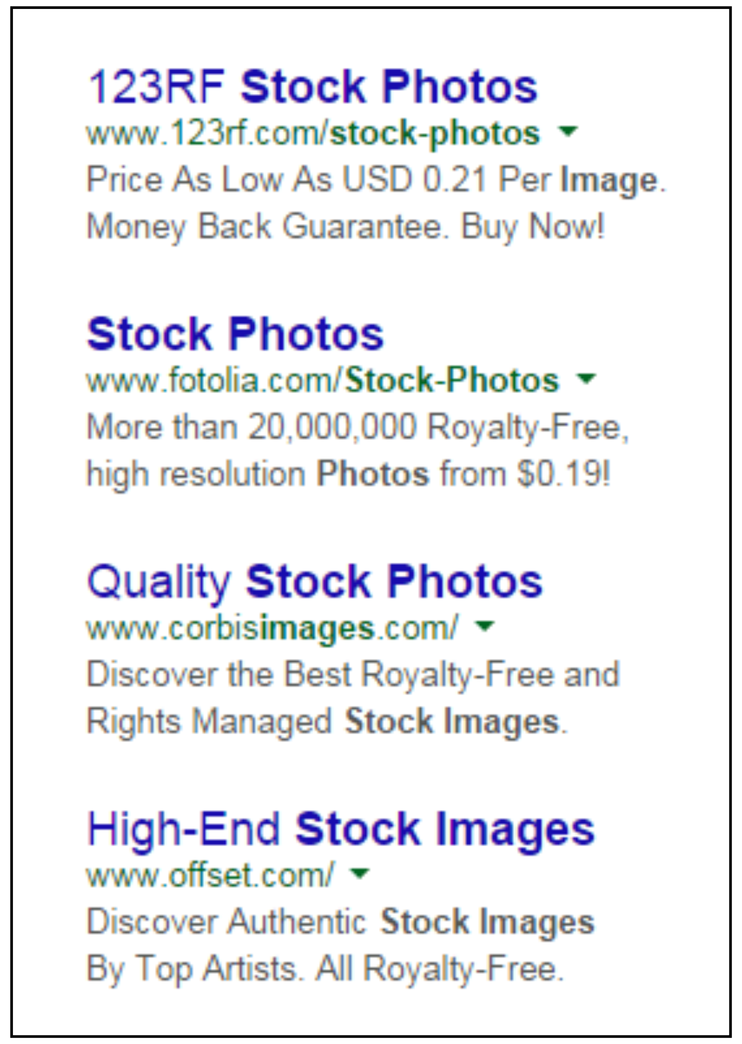


Independence of Irrelevant Alternatives (IIA)

- Examples where it (arguably) doesn't hold:



Music
(Debreu, 1960)

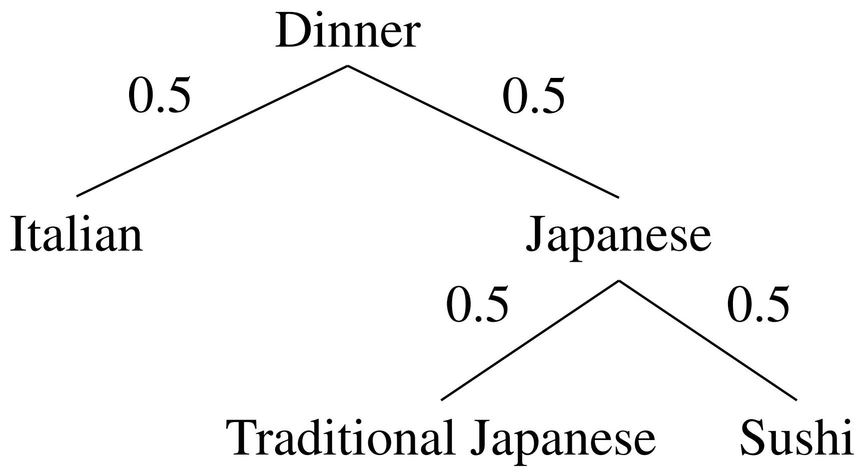

123RF Stock Photos
www.123rf.com/stock-photos
Price As Low As USD 0.21 Per Image.
Money Back Guarantee. Buy Now!

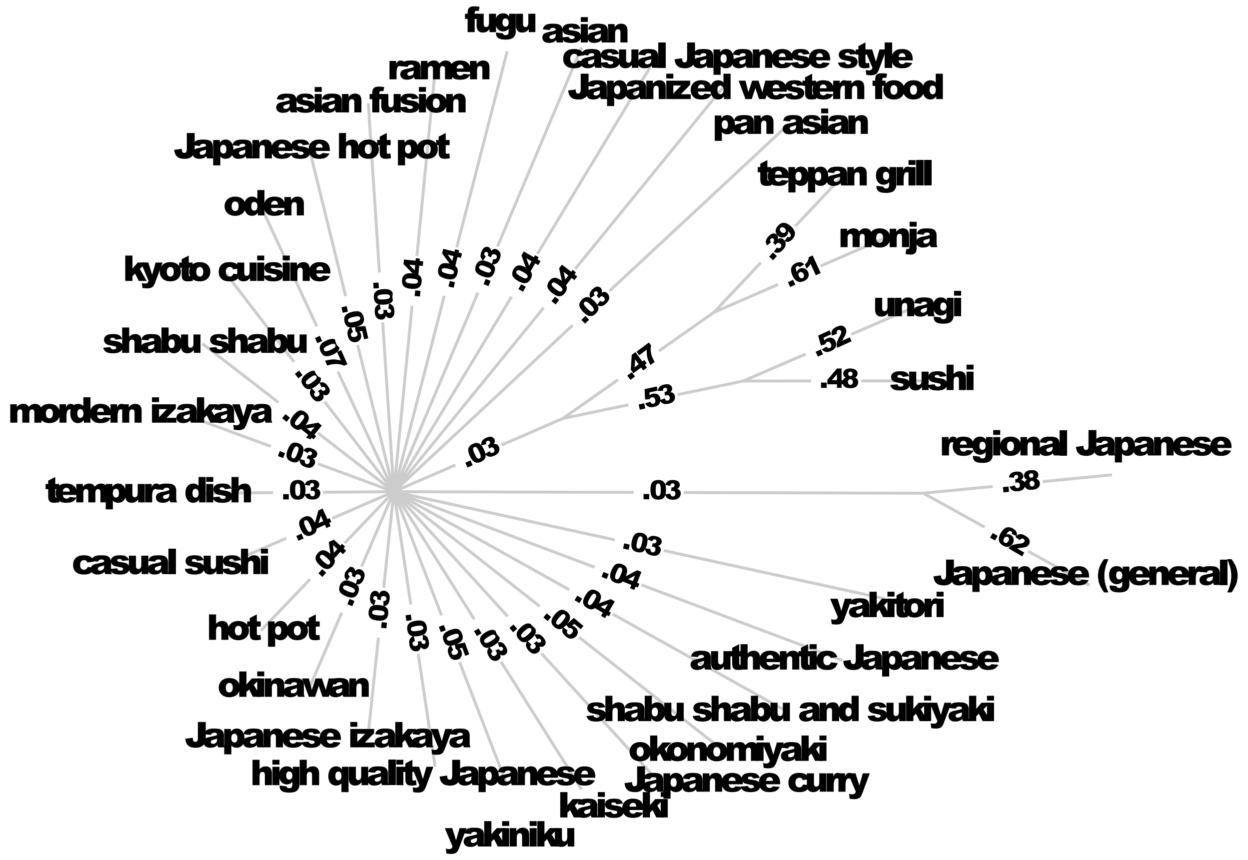
Stock Photos
www.fotolia.com/Stock-Photos
More than 20,000,000 Royalty-Free,
high resolution Photos from $0.19!

Quality Stock Photos
www.corbisimages.com/
Discover the Best Royalty-Free and
Rights Managed Stock Images.

High-End Stock Images
www.offset.com/
Discover Authentic Stock Images
By Top Artists. All Royalty-Free.

Search ads
(Ieong-Mishra-Sheffet 2012, Yin et al. 2014)

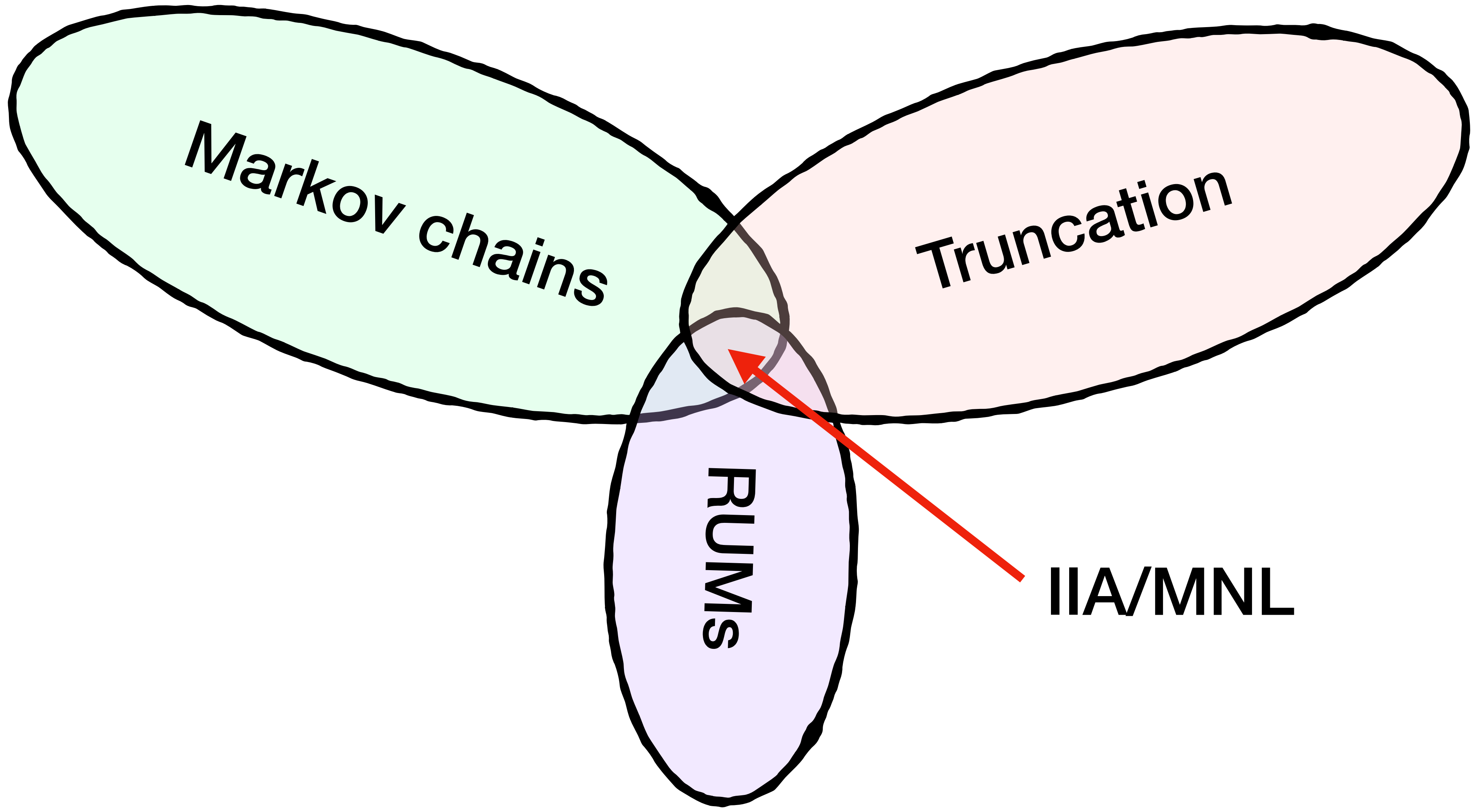




Web browsing
(Benson-Kumar-Tomkins, 2016)

Three perspectives on IIA, beyond IIA

1. **Random utility model** (RUM) with Gumbel noise (Yellot, 1977)
2. Stationary distribution of a **Markov chain** (Maystre & Grossglauser, 2015)
3. First-order truncation of a **Taylor-like expansion** of a choice system (Batsell & Polking,1985; Seshadri et al. 2019)

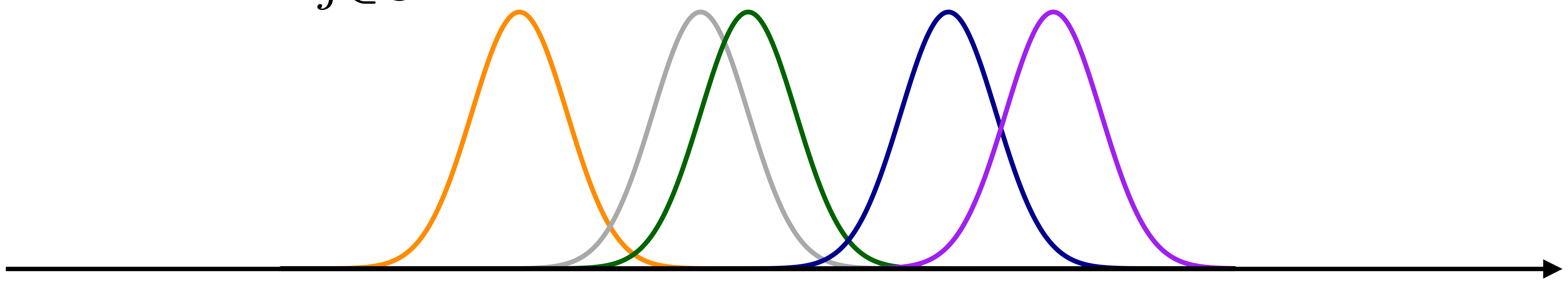


(Setting aside mixture/nested models today.)

Each derivation is its own path to a beyond-IIA model of choice.

(1) Random utility models and IIA

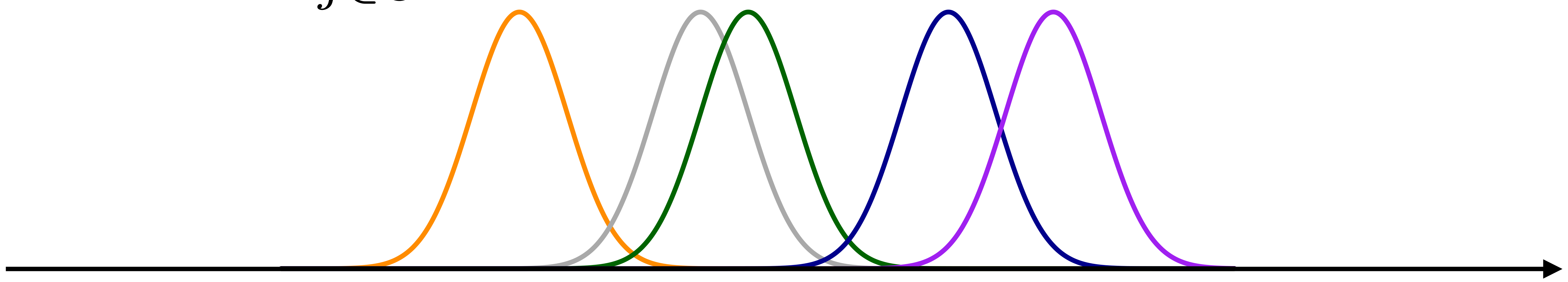
- For each $i \in \mathcal{X}$, associate a random variable $X_i = \mu_i + \epsilon_i$.
- Let $P_{i,C} = Pr(X_i = \max_{j \in C} X_j)$.



- Iff $\epsilon_1, \ldots, \epsilon_n$ are independent zero-mean Gumbel, $P_{i,C} = \dfrac{\exp(\mu_i)}{\sum_{j \in C} \exp(\mu_j)}$. (**MNL!**)

(1) Random utility models and IIA

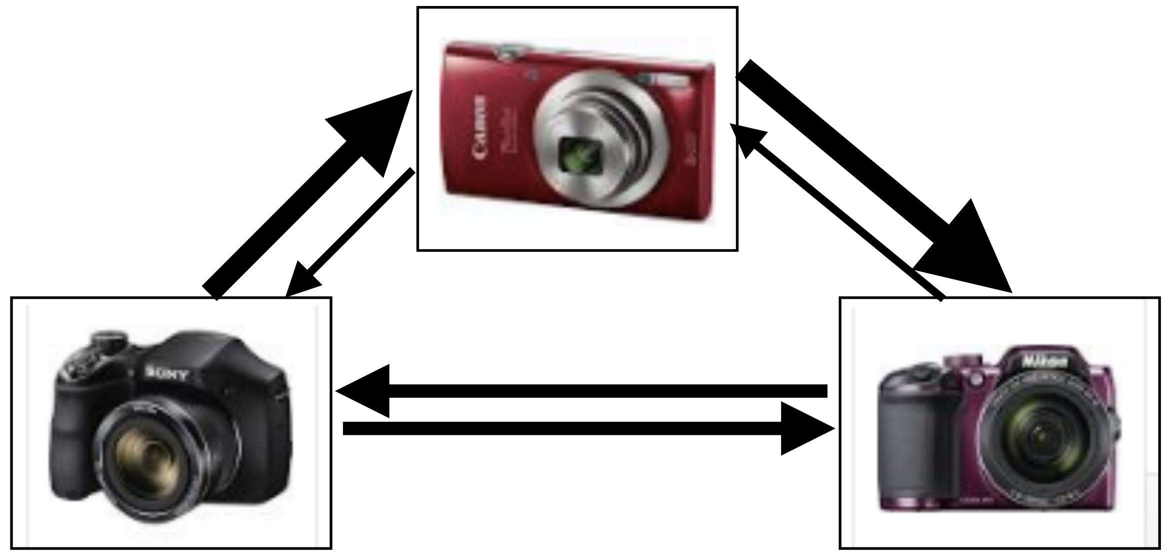
- For each $i \in \mathcal{X}$, associate a random variable $X_i = \mu_i + \epsilon_i$.
- Let $P_{i,C} = Pr(X_i = \max_{j \in C} X_j)$.



- Iff $\epsilon_1, \ldots, \epsilon_n$ are independent zero-mean Gumbel, $P_{i,C} = \dfrac{\exp(\mu_i)}{\sum_{j \in C} \exp(\mu_j)}$. (**MNL**!)
- See Falmagne (1978)'s characterization theorem of RUMs.
- RUMs need not be stochastically transitive! (Makhijani & U, 2019) connects transitivity to log-likelihood concavity of item-level parameterizations.

(2) Choice systems from Markov chains

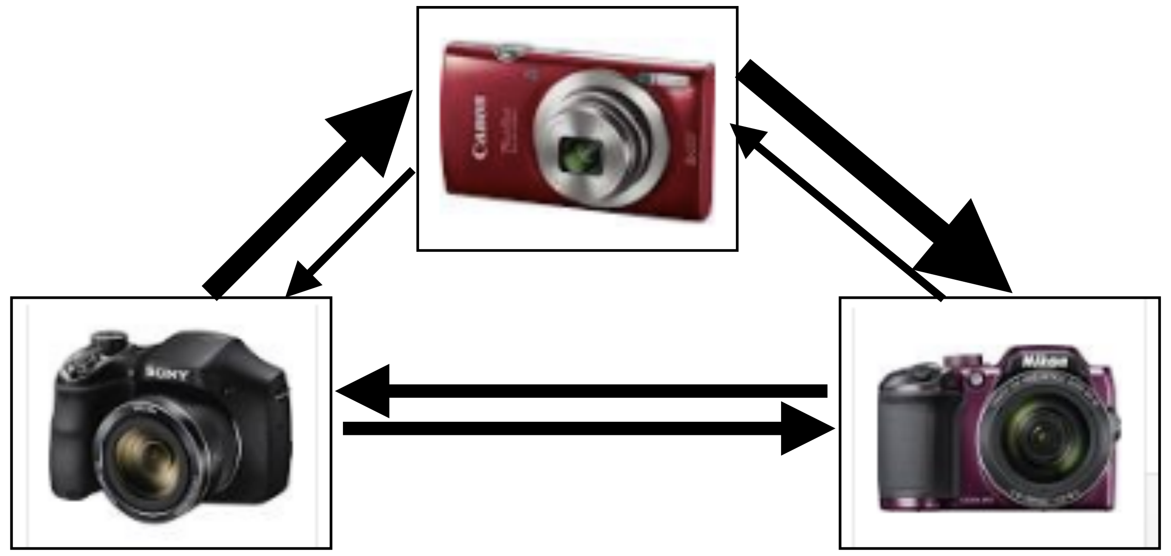
- Consider a continuous-time Markov chain defined on $\mathcal{X}$, parameterized by **Q.**



$$\pi^T \begin{bmatrix} -\sum_{i\neq 1} q_{1i} & q_{12} & q_{13} \\ q_{21} & -\sum_{i\neq 2} q_{2i} & q_{23} \\ q_{31} & q_{32} & -\sum_{i\neq 3} q_{3i} \end{bmatrix} = 0$$

(2) Choice systems from Markov chains

- Consider a continuous-time Markov chain defined on $\mathcal{X}$, parameterized by **Q**.



$$\pi^T \begin{bmatrix} -\sum_{i\neq 1} q_{1i} & q_{12} & q_{13} \\ q_{21} & -\sum_{i\neq 2} q_{2i} & q_{23} \\ q_{31} & q_{32} & -\sum_{i\neq 3} q_{3i} \end{bmatrix} = 0$$

- Define a chain for each subset $C \subseteq \mathcal{X}$ by restricting the rate matrix, e.g.:

$$\pi^T \begin{bmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{bmatrix} = 0$$

- These stationary distributions define a choice system (Ragain & U, 2016)
- See also: (Maystre & Grossglauser, 2015)

(3) Truncating choice systems

- Define item-set utilities $u(x|C), \forall x \in C,$ such that $\displaystyle\sum_{y \in C} u(y|C) = 0.$
- Arbitrary *universal logit model:*

$$P_{x,C} = \frac{\exp(u(x|C))}{\sum_{y \in C} \exp(u(y|C))}.$$

(3) Truncating choice systems

- Define item-set utilities $u(x|C), \forall x \in C$, such that $\sum_{y \in C} u(y|C) = 0.$
- Arbitrary *universal logit model:*

$$P_{x,C} = \frac{\exp(u(x|C))}{\sum_{y \in C} \exp(u(y|C))}.$$

- Item-set utilities can be uniquely* expanded as (Batsell & Polking, 1985):

$$u(x|C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{\{y\} \in C \setminus x} v(x|\{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\} \subseteq C \setminus x} v(x|\{y,z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x|C \setminus \{x\})}_{|C|\text{th order}}$$

*with constraints, not shown.

(3) Truncating choice systems

$$u(x|C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{\{y\}\in C\setminus x} v(x|\{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\}\subseteq C\setminus x} v(x|\{y,z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x|C\setminus\{x\})}_{|C|\text{th order}}$$

- Call p^{th} order model $\mathscr{M}_p$. Notice that $\mathscr{M}_1 \subset \mathscr{M}_2 \subset \ldots \subset \mathscr{M}_{n-1}$.

(3) Truncating choice systems

$$u(x|C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{\{y\}\in C\setminus x} v(x|\{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\}\subseteq C\setminus x} v(x|\{y,z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x|C \setminus \{x\})}_{|C|\text{th order}}$$

- Call p^{th} order model $\mathscr{M}_p$. Notice that $\mathscr{M}_1 \subset \mathscr{M}_2 \subset \ldots \subset \mathscr{M}_{n-1}$.

MNL/IIA

Universal logit

Context-dependent utility model

- For $\mathscr{M}_2$, after manipulations, choice probabilities can be written as:

$$P_{x,C} = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{z \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}.$$

- Assumes “Pairwise Linear Dependence of Alternatives”
- Negative log likelihood is **convex** in parameters U!

Context-dependent utility model

- For $\mathscr{M}_2$, after manipulations, choice probabilities can be written as:

$$P_{x,C} = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{z \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}.$$

- Assumes "Pairwise Linear Dependence of Alternatives"
- Negative log likelihood is **convex** in parameters U!
- Can be made **low-rank** (non-convex), essentially a **matrix factorization loss**:

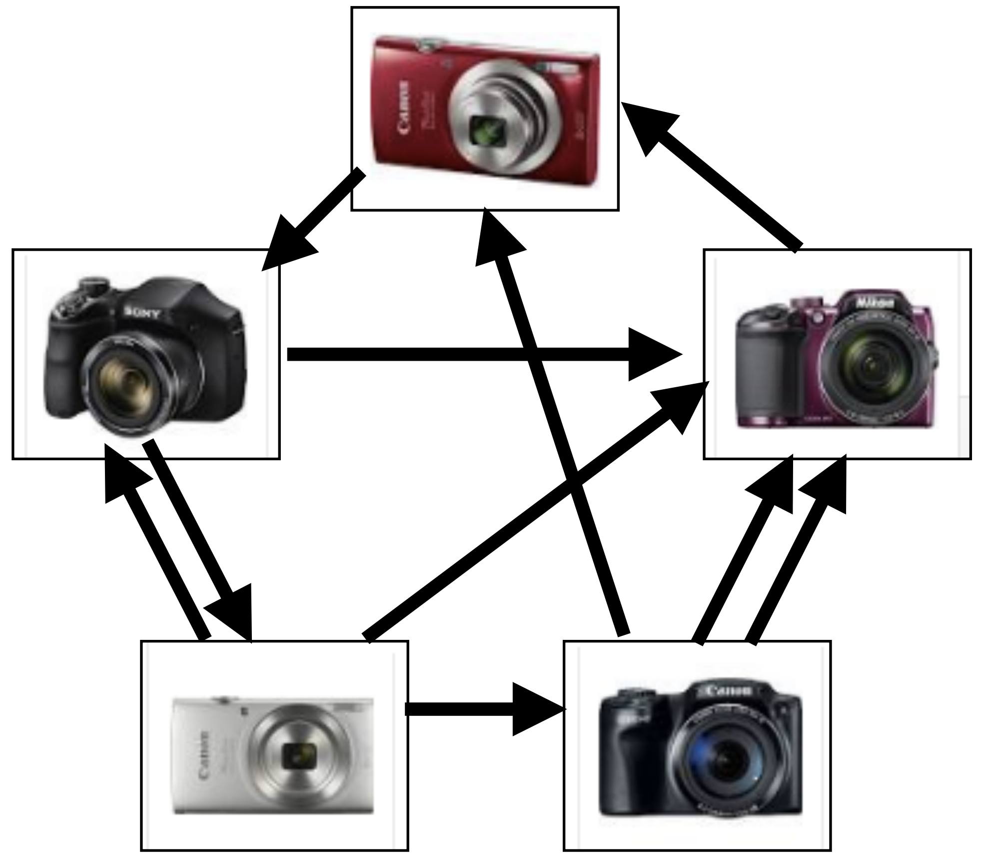
$$P_{x,C} = \frac{\exp(\sum_{z \in C \setminus x} c_z^T t_x)}{\sum_{z \in C} \exp(\sum_{z \in C \setminus y} c_z^T t_y)}.$$

Structure-dependent convergence rate

- **Identifiability conditions** in choice models are combinatorial (Ford 1957).
- Batsell & Polking used least squares (cleverly!), not MLE.
- Under mild regularity conditions, we show

$$\mathbb{E}[||\hat{u}_{MLE}(\mathcal{D}) - u^*||_2^2] \leq \frac{c}{\lambda_2(L(\mathcal{D}))} \frac{n(n-1)}{m}.$$

  where n is the number of items, m the size of the data, and $\mathcal{D}$ a random dataset generated under the model.



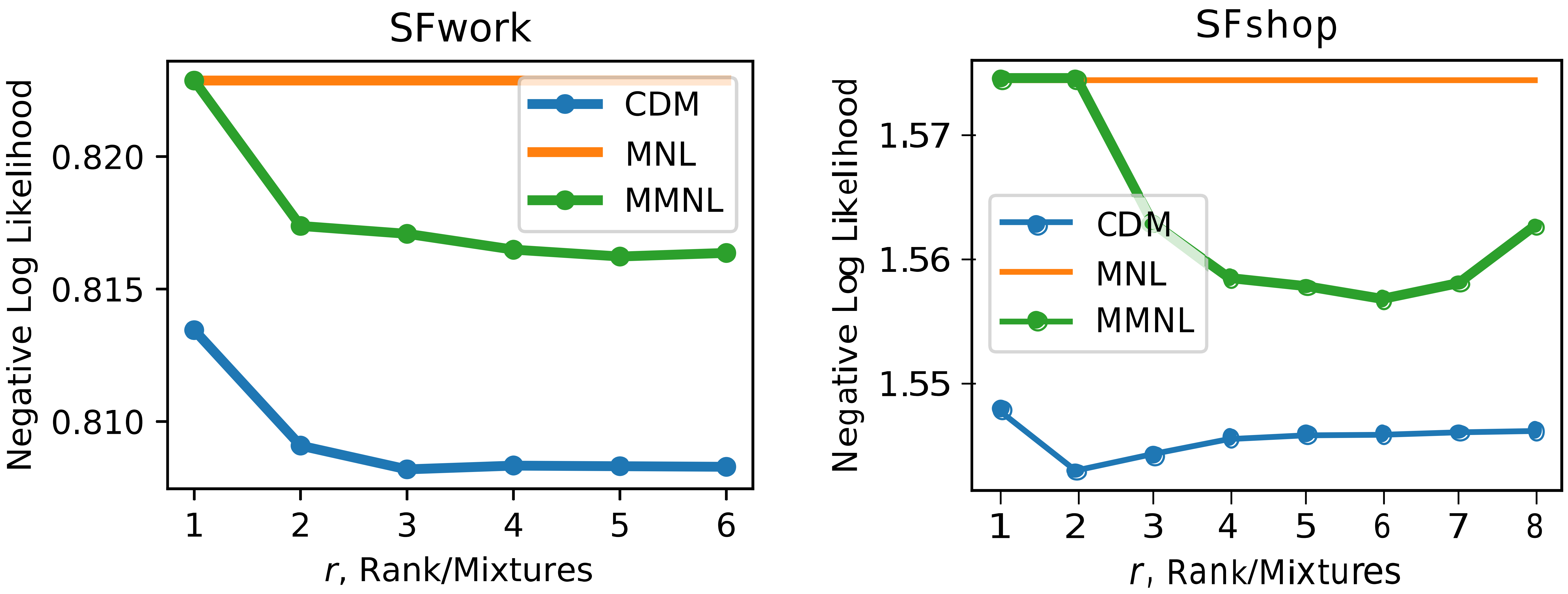
- Here $\lambda_2(L(\mathcal{D}))$ is the second smallest eigenvalue of a **Laplacian-like matrix.** For pairwise comparisons: Laplacian of comparison graph (Shah et al. 2016).

Broader implications

- Convergence result is for full-rank case; bound still applies when low-rank.
- Analysis also applies to **Blade-Chest** model (Chen & Joachims, 2016a,b) and many **word2vec**-type models (Mikolov et al., 2013).
  - For word2vec, the likelihood objective is typically approximated by "negative sampling" the choice set, also changes the objective.
- Recent related work**:**
  - Extension to "salient" features (Bower & Balzano, 2020).
  - Promoting a particular choice (Tomlinson & Benson, 2020).

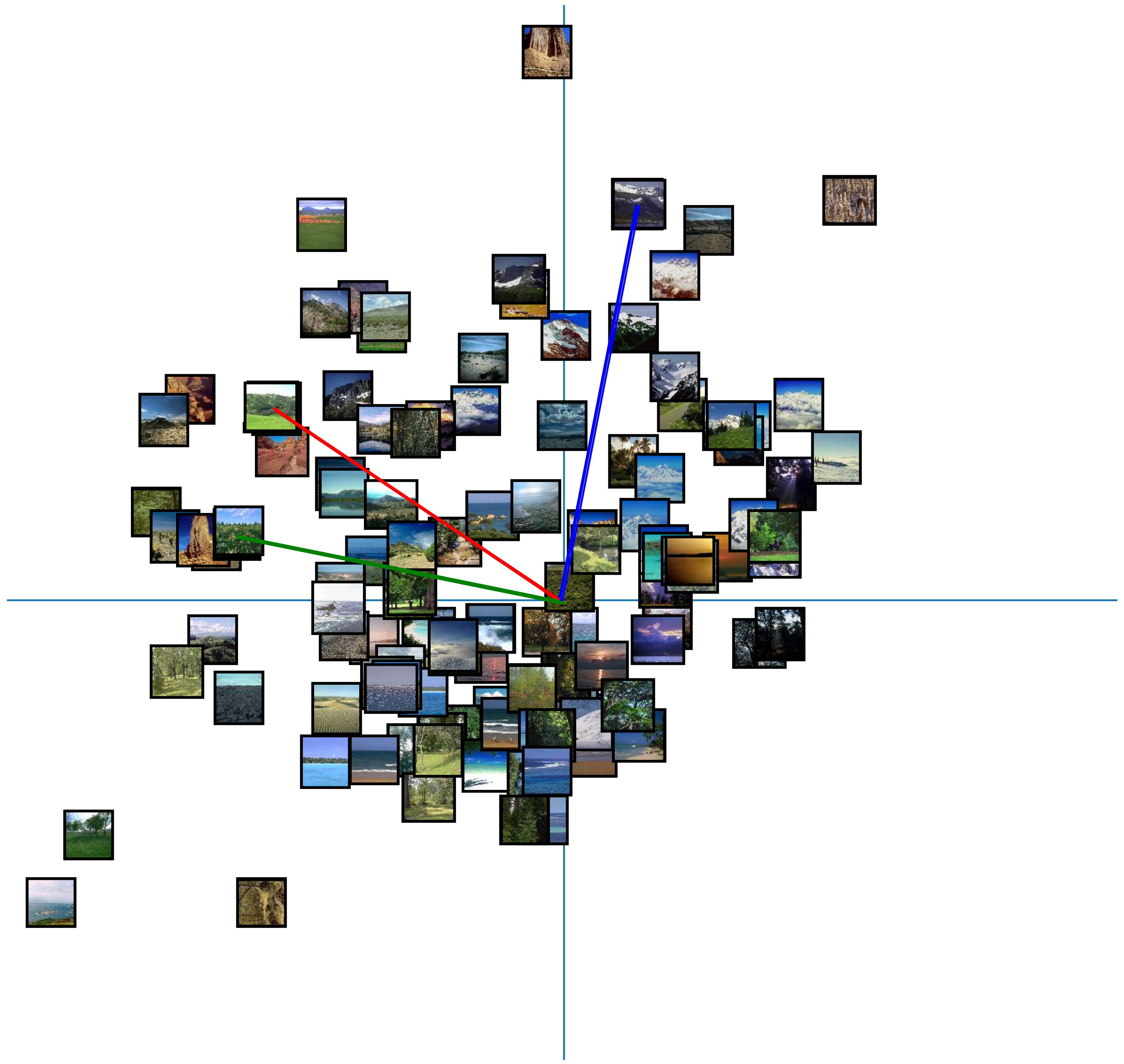
CDM empirical results

- Predicting transportation choices (Koppelman & Bhat, 2006) with the CDM:

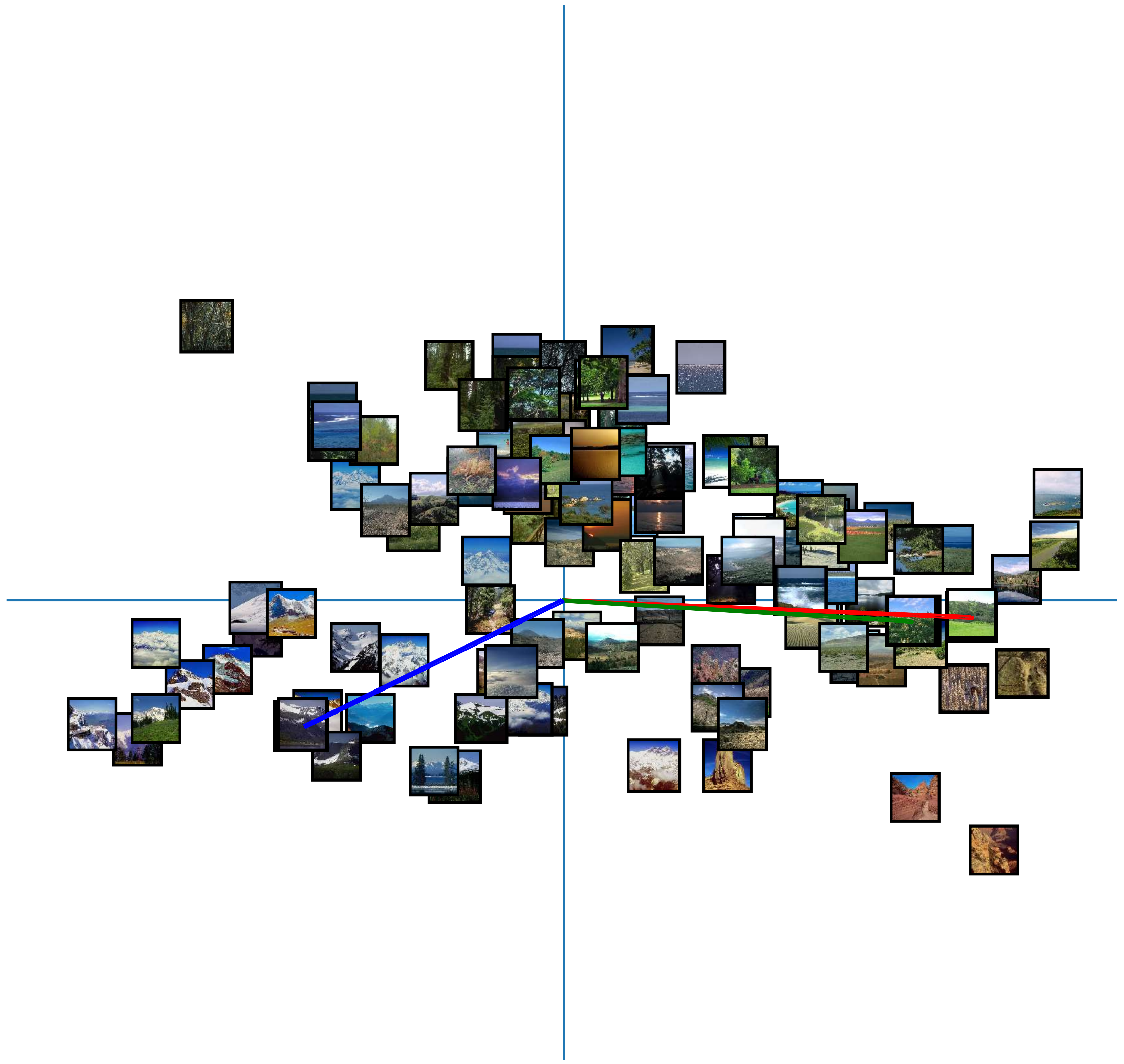


Low-rank factorization of U: embeddings

- “One of these things is not like the other…” triplets (Heikinheimo & Ukkonen, 2013)



CDM Target Vectors, $r = 2$

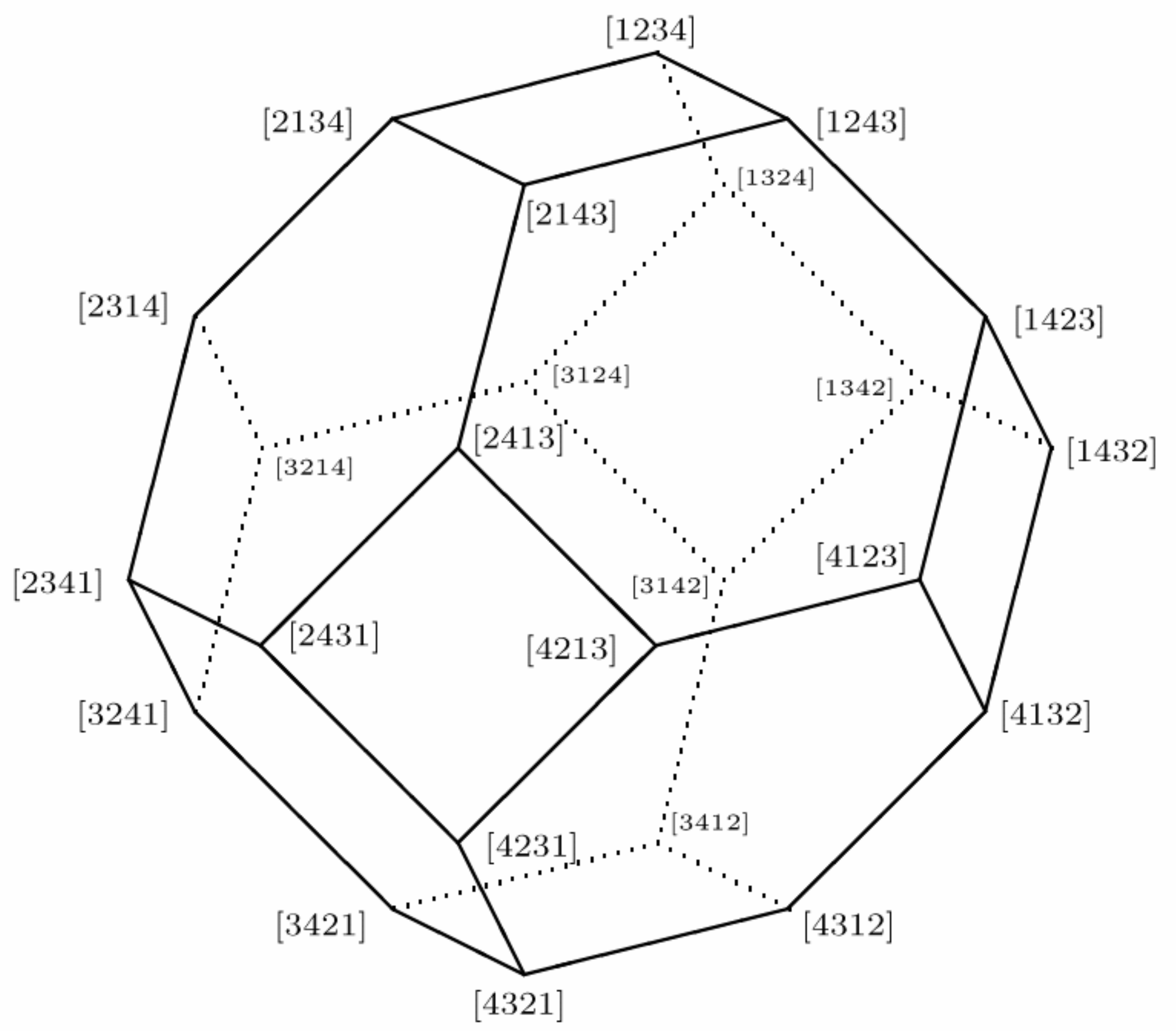


CDM Context Vectors, $r = 2$

Ranking as choice

- Plackett-Luce: distributions over S_n as "repeated MNL choice":

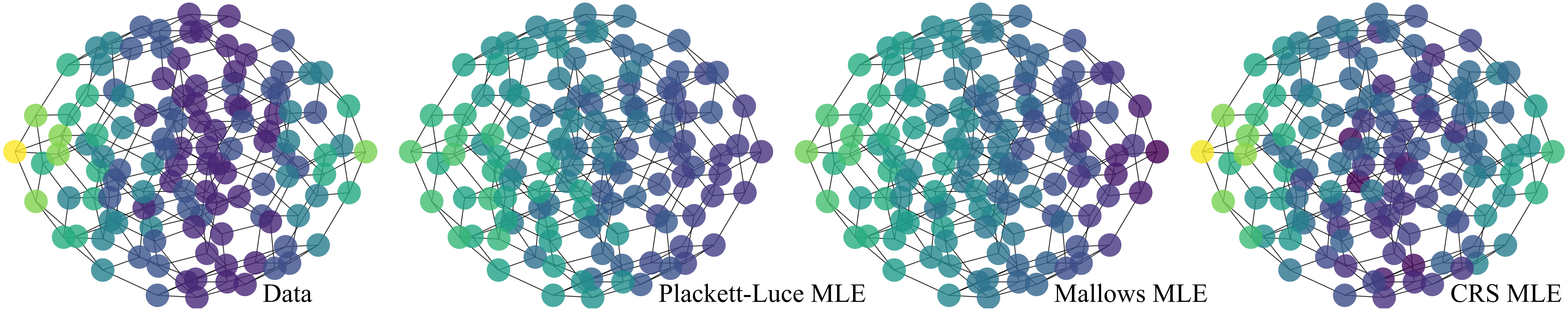
$$\Pr[\pi = 123\cdots n] = \prod_{i=1}^{n} \frac{\exp(u_i)}{\sum_{j=i}^{n} \exp(u_j)}$$



- See also: Mallows, mixtures of Mallows/PL.
- What happens if we replace MNL with CDM?

Ranking distributions

- Contextual repeated selection (CRS) can represent rich, multi-modal distributions with the same learning efficiency/guarantees as CDM choice.



Ranking MLE from data

- Similar to choice result, expected risk bound, with ℓ rankings of length n:

$$\mathbb{E}\left[\,\|\hat{u}_{MLE}(\mathcal{R}) - u^\star\|_2^2\,\right] \le \mathbb{E}\left[\min\left\{\frac{c'_B n^3}{\ell\lambda_2(L)}, 4B^2 n\right\}\right] \le c_B \frac{n^7}{\ell}$$

- Notice second eigenvalue can be bounded absolutely.

Ranking MLE from data

- Similar to choice result, expected risk bound, with ℓ rankings of length n:

$$\mathbb{E}\left[\,\|\hat{u}_{MLE}(\mathcal{R}) - u^\star\|_2^2\,\right] \leq \mathbb{E}\left[\min\left\{\frac{c_B' n^3}{\ell\lambda_2(L)}, 4B^2 n\right\}\right] \leq c_B \frac{n^7}{\ell}$$

- Notice second eigenvalue can be bounded absolutely.

- Paper also has **tail bounds** (not just expected risk).
- Paper also sharpens convergence analysis of vanilla **MNL, Plackett-Luce** (!)

Testing IIA

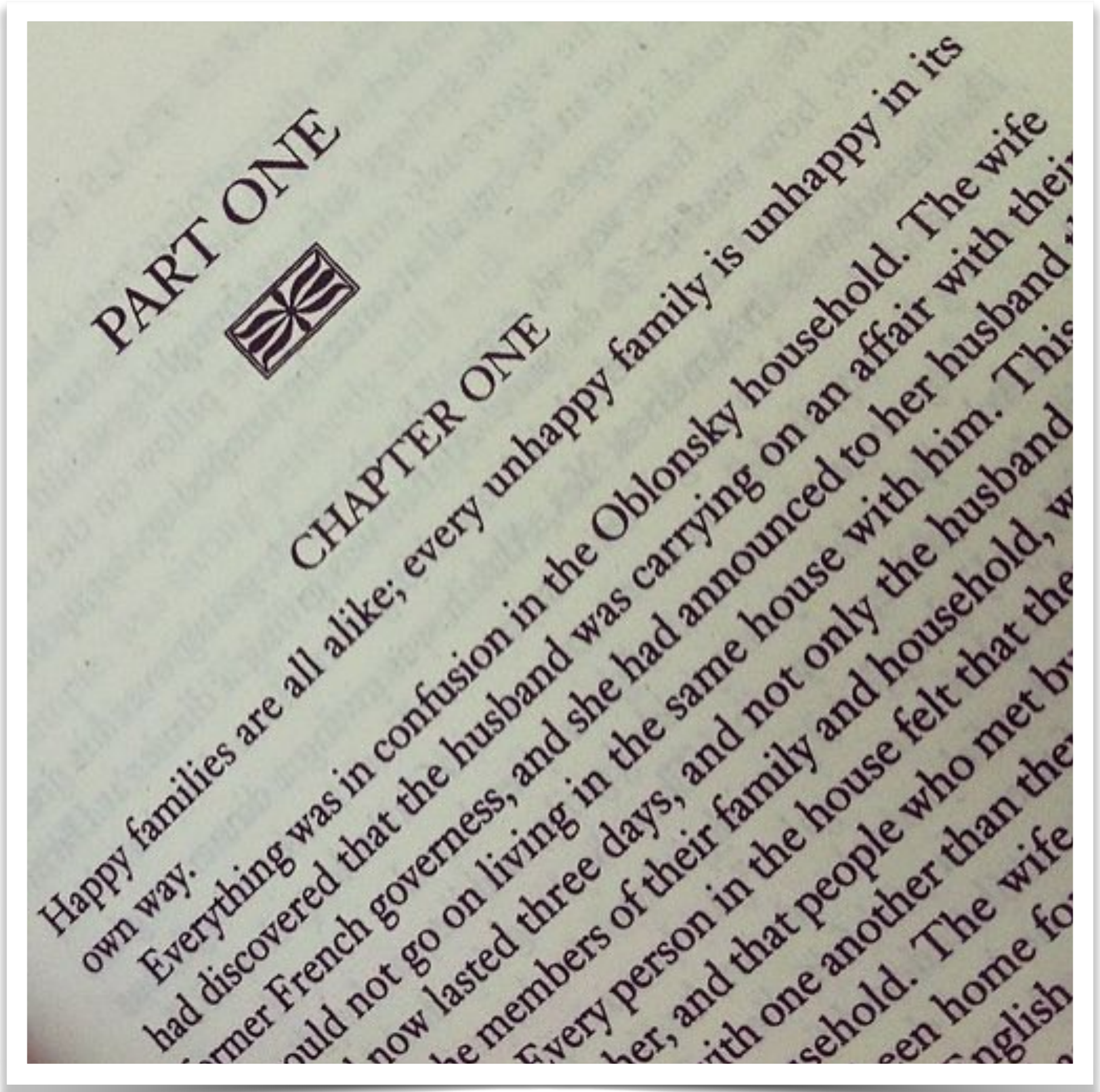
Why is testing IIA hard?

- **Anna Karenina Principle** of high-dimensional hypothesis testing: “all nulls are alike; deviations from the null all deviate in their own way.”

- **Applied to IIA:** there are only a few ways to be “rational,” there are a many unique ways that people can be “irrational.”

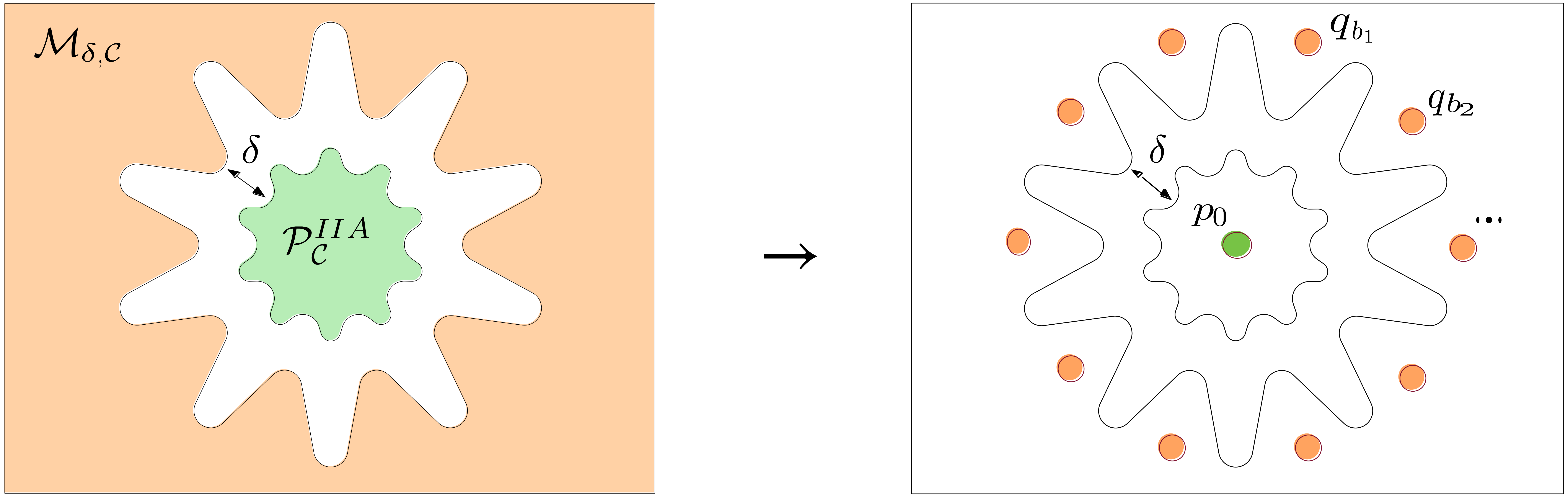
- Follows the burst of work on finite-sample lower bounds on testing:

  (Paninski 2008; Wei & Wainwright 2016; Valiant & Valiant 2017; Daskalakis, Kamath, Wright 2018; Balakrishnan & Wasserman 2018).



Separation and “orthogonal” perturbations

- Begin with the basic formula for lower bounds on minimax risk (and testing):
  - Define separation (TV distance).
  - Simplify to testing uniform choice system p_0 vs. composite of other distributions perturbed out of the space of IIA.



Structure-dependent lower bounds

- In a strict sense, if data doesn't contain choices from every subset, the full implications of IIA can't be tested.
- Instead: let $\mathcal{C}$ be the set of subsets being compared.
- **Example:** $\mathcal{X} = \{1, 2, 3, 4\}$

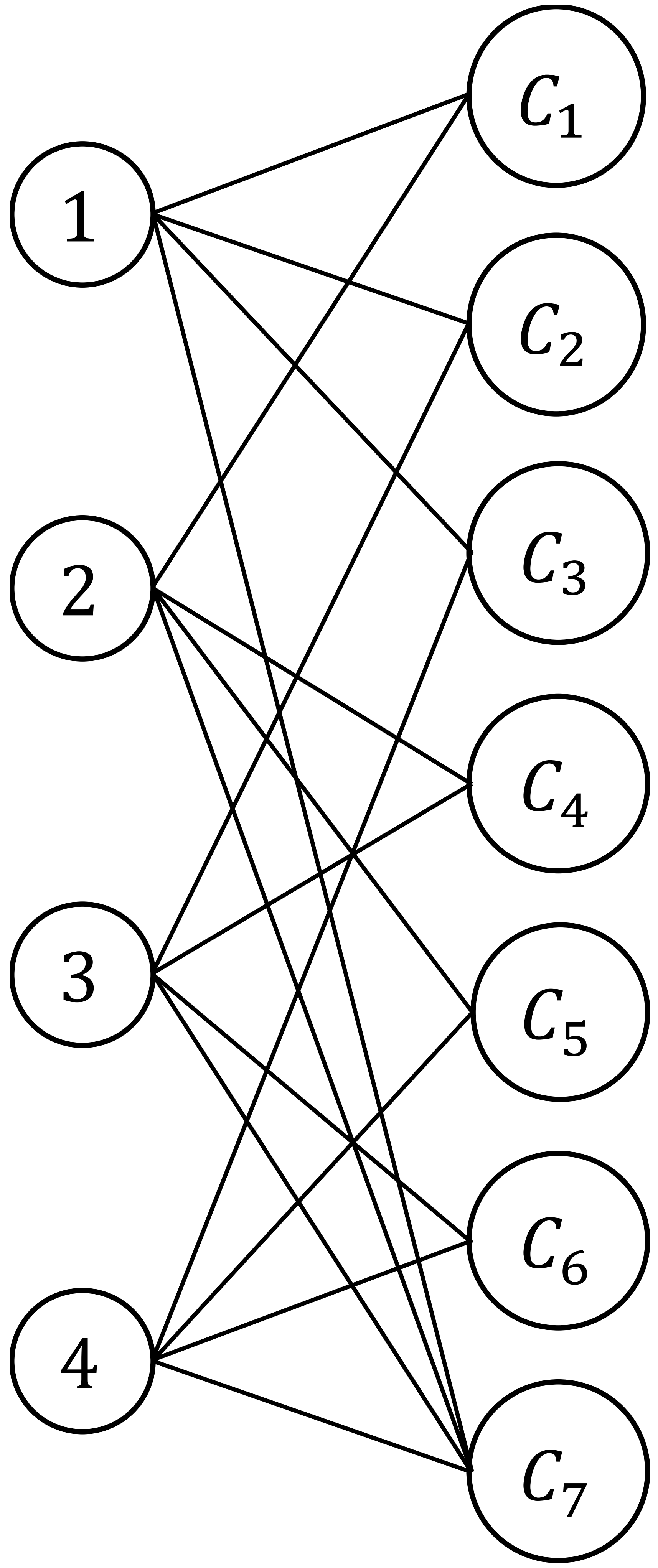
$$\mathcal{C} = \{\underbrace{\{1,2\}}_{C_1}, \underbrace{\{1,3\}}_{C_2}, \underbrace{\{1,4\}}_{C_3}, \underbrace{\{2,3\}}_{C_4}, \underbrace{\{2,4\}}_{C_5}, \underbrace{\{3,4\}}_{C_6}, \underbrace{\{1,2,3,4\}}_{C_7}\}$$

Structure-dependent lower bounds

- In a strict sense, if data doesn't contain choices from every subset, the full implications of IIA can't be tested.
- Instead: let $\mathcal{C}$ be the set of subsets being compared.
- **Example:** $\mathcal{X} = \{1, 2, 3, 4\}$

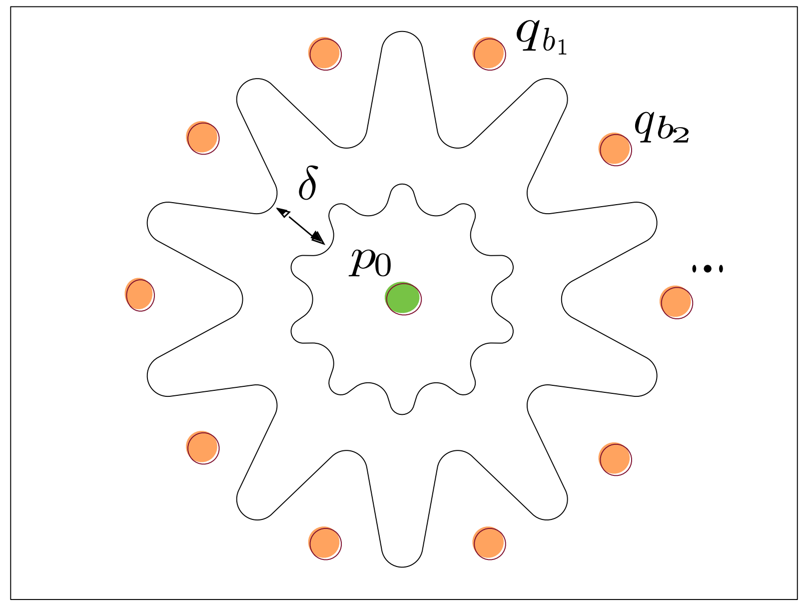
$$\mathcal{C} = \{\underbrace{\{1,2\}}_{C_1}, \underbrace{\{1,3\}}_{C_2}, \underbrace{\{1,4\}}_{C_3}, \underbrace{\{2,3\}}_{C_4}, \underbrace{\{2,4\}}_{C_5}, \underbrace{\{3,4\}}_{C_6}, \underbrace{\{1,2,3,4\}}_{C_7}\}$$

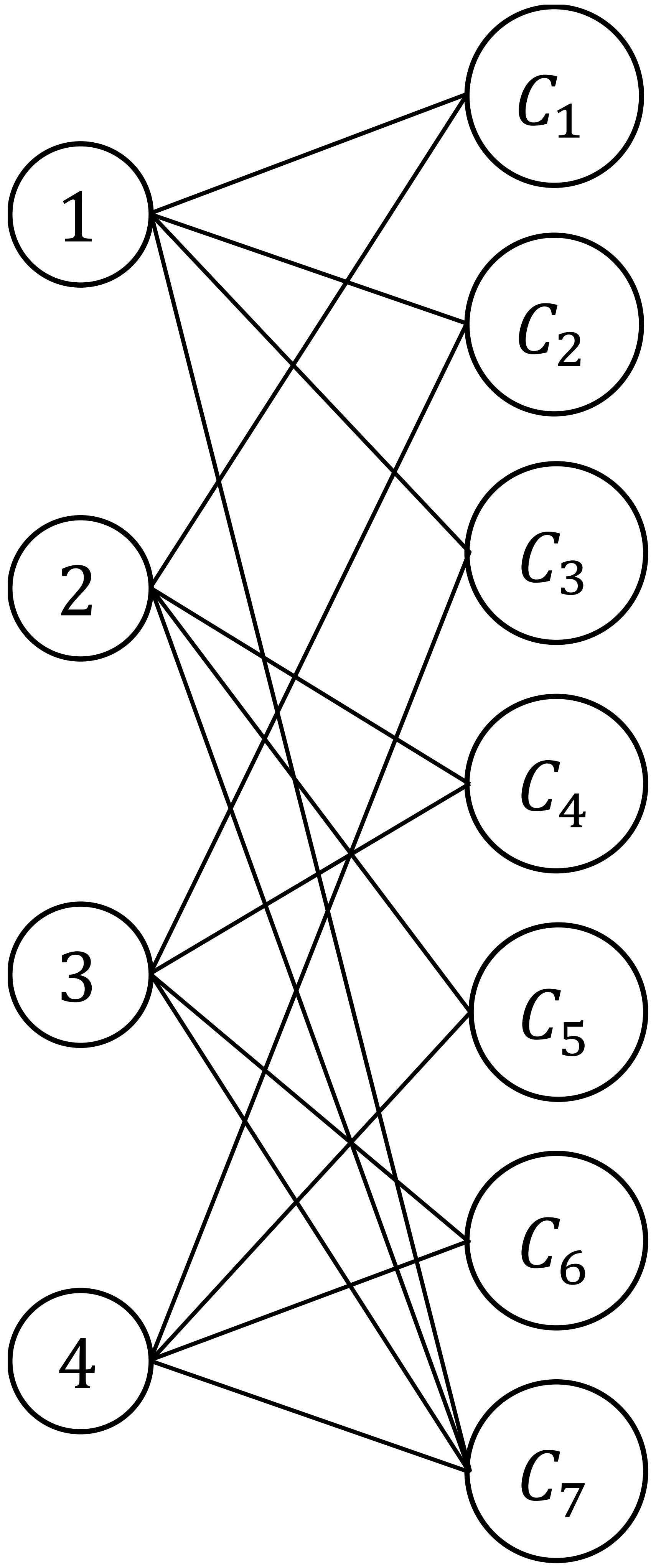
- Consider: bipartite comparison incidence graph $G_{\mathcal{C}} = (\mathcal{X}, \mathcal{C}, E)$:



Constructing perturbations

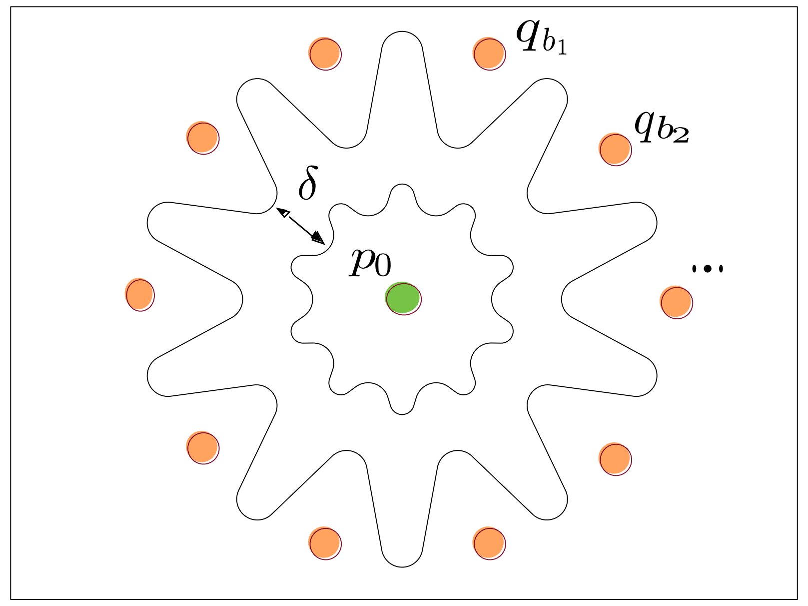
- Starting at uniform, want perturbations out of IIA space that all still **project back** onto uniform.
- Want as **many perturbations** as possible.





Constructing perturbations

- Starting at uniform, want perturbations out of IIA space that all still **project back** onto uniform.



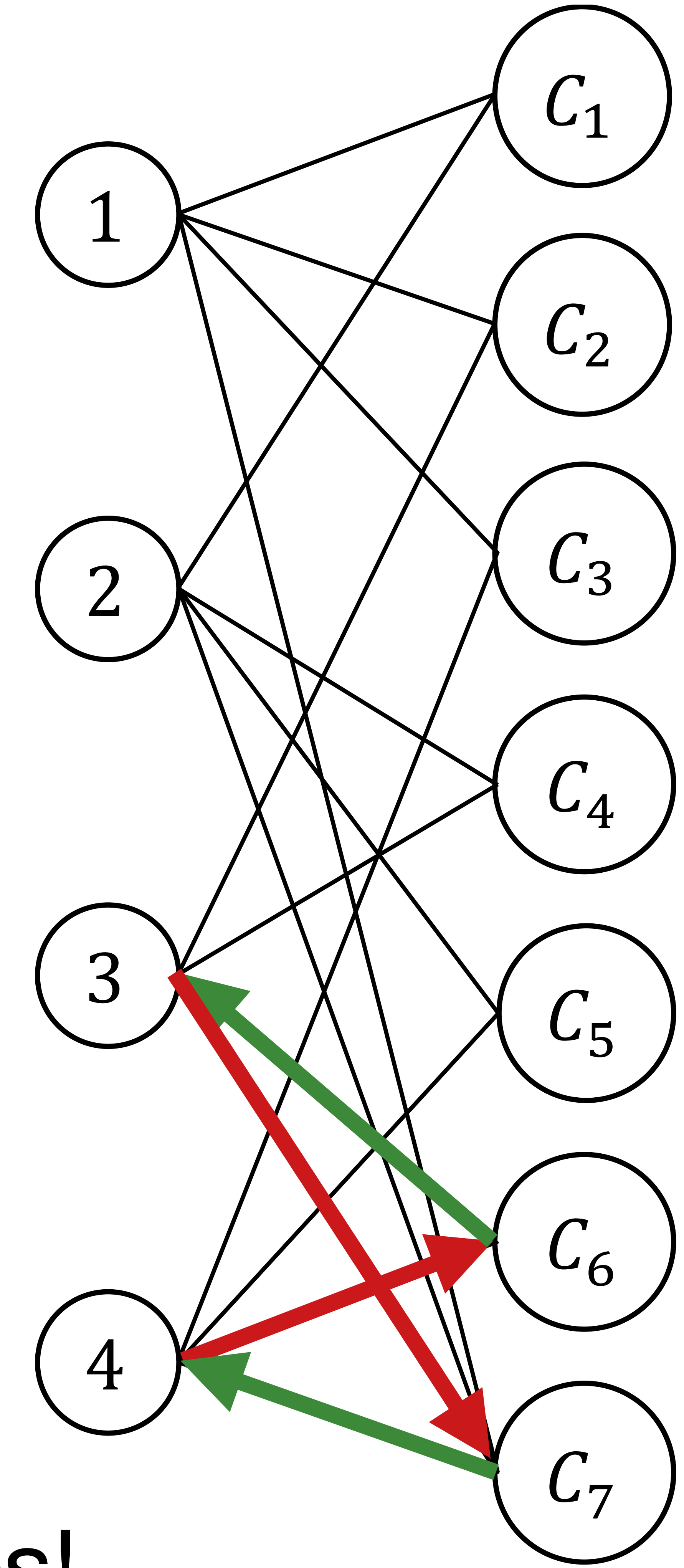
- Want as **many perturbations** as possible.

- Sketch of construction:

  - Need **sets** to maintain their frequency, **items** to maintain their choice frequency.

  - Seek perturbations of parameters that keep overall item probabilities fixed, set probabilities fixed.

  - Seek a **cycle decomposition** of $G_{\mathcal{C}} = (\mathcal{X}, \mathcal{C}, E)$ into many cycles!



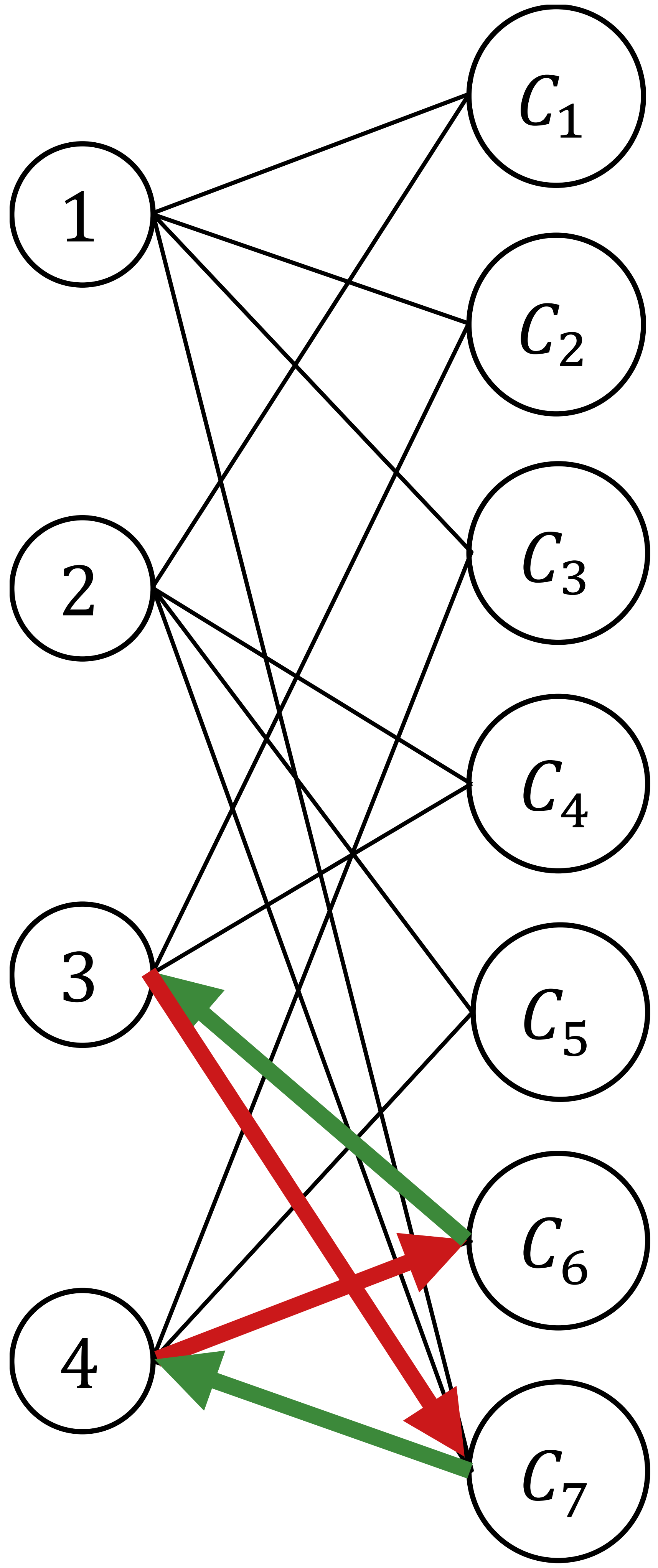
Structure-dependent lower bounds

- Let $\mu(\sigma)$ and $\alpha(\sigma)$ be properties of some cycle decomposition σ of $G_{\mathcal{C}} = (\mathcal{X}, \mathcal{C}, E)$. Then for N choices:

Structure of C	$R_{N,\delta}(\mathcal{P}_C^{\mathrm{IIA}})$
General	$\geq \frac{1}{2} - \frac{1}{4}\left(\exp\left(\frac{8\mu(\sigma)^4\alpha(\sigma)N^2\delta^4}{d}\right) - 1\right)^{\frac{1}{2}}$
All subsets, $d = n2^{n-1}$	$\geq \frac{1}{2} - \frac{1}{4}\left(\exp\left(\frac{c\log(n)^5N^2\delta^4}{n2^{n-1}}\right) - 1\right)^{\frac{1}{2}}$
All pairs, $d = n(n-1)$	$\geq \frac{1}{2} - \frac{1}{4}\left(\exp\left(\frac{cN^2\delta^4}{n(n-1)}\right) - 1\right)^{\frac{1}{2}}$

- $R_{N,\delta} \geq 0$ means lower bound has fallen away.

- No upper bounds, no tests analyzed.



Thank you!

- **Choice systems** are beautiful things.

- Doors have recently opened to introduce and analyze tractable models beyond IIA based on **Markov chains**, based on **truncations**.

- **Testing IIA**: we replace ambiguity with rigorous pessimism.

- **Papers:**
  <u>PCMC</u>: Ragain & Ugander, NeurIPS 2016
  <u>CDM</u>: Seshadri, Peysakhovich, Ugander, ICML 2019
  <u>Testing</u>: Seshadri & Ugander, EC 2019
  <u>Choice models of networks</u>: Overgoor et al. WWW 2019, KDD 2020
  <u>Ranking:</u> Seshadri, Ragain, Ugander, NeurIPS 2020