

# Preference Modeling with Context-Dependent Salient Features

Amanda Bower

Joint work with Laura Balzano

University of Michigan

37th International Conference on Machine Learning  
July 2020

# Preference Learning Problem Description

Suppose we have  $n$  items and we'd like to obtain one universal ranking of these items.



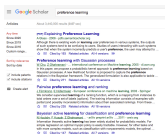
One cannot ask someone to rank them all for large  $n$ .

Instead, we ask for  $k$ -wise comparisons,  $k \ll n$ , and try to *learn the complete preference ordering*.

Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

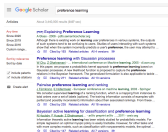
# Preference Learning Motivation

- Job candidates or prospective graduate students
- Conference papers or science fair projects
- Sports teams
- Websites as related to a particular search
- Products that an industry would like you to buy
- Anomalies that need to be investigated (in a computer network, in a medical situation)



# Preference Learning Challenges

- Noisy  $k$ -wise comparisons
- Limited feedback ( $k$  is small, say 2, not all pairs are compared, or only click data is available)
- **Often the comparisons have intransitivity (even from a single user)**



# Preference Learning – Intransitivity

Problem that is usually ignored: intransitivity.



Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

# Preference Learning – Intransitivity

Problem that is usually ignored: intransitivity.



But this is a prevalent characteristic of preference data.

Data Set	Valid Triplets	Strong Violations	Moderate Violations	Weak Violations
NBA	2654	1439 (54%)	1185 (45%)	272 (10%)
Tennis	4793	1092 (23%)	1080 (23%)	651 (14%)
Nascar	65003	26354 (41%)	17128 (26%)	4171 (6%)
Jester	161700	14560 (9%)	327 (.2%)	78 (.05%)
Sushi-A	120	28 (23%)	0 (0%)	0 (0%)
Sushi-B	139992	66013 (47%)	26366 (19%)	4939 (4%)
District	48	25 (52%)	8 (16%)	0 (0%)
Car	120	46 (38%)	7 (6%)	0 (0%)
Sonancia	874	175 (20%)	175 (20%)	108 (13%)
New Yorker Captions	3990	1823 (51%)	606 (17%)	199 (6%)

Table: [Dataset details](#) [Stochastic transitivity definitions](#)

Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

# Salient Features Motivation



Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

# Salient Features Motivation: Which is more comfortable?

Pair #1



Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>



# Salient Features Motivation: Which is more comfortable?

Pair #2



Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

## Violation of Rational Choice - Related Work

- Several social science papers have commented on the “salient feature” phenomenon  
[Bordalo et al., 2013, Rieskamp et al., 2006, Brown and Peterson, 2009, Tversky, 1972].
- [Seshadri et al., 2019, Chen and Joachims, 2016] model item utilities as context-dependent by learning features of the items, whereas  
[Rosenfeld et al., 2019, Pfannschmidt et al., 2019] learn item utilities as context-dependent with neural networks assuming known features.
- [Niranjan and Rajkumar, 2017, Kleinberg et al., 2017, Ragain and Ugander, 2016, Benson et al., 2016, Rajkumar et al., 2015, Yang and Wakin, 2015] ...

# Our Contributions

- Our model:
  - is a convex preference model of intransitivity where different comparisons use different features

# Our Contributions

- Our model:
  - is a convex preference model of intransitivity where different comparisons use different features,
  - is inspired by social choice theory

# Our Contributions

- Our model:
  - is a convex preference model of intransitivity where different comparisons use different features,
  - is inspired by social choice theory,
  - makes a direct connection between intransitive preferences and a ranking of all the items

# Our Contributions

- Our model:
  - is a convex preference model of intransitivity where different comparisons use different features,
  - is inspired by social choice theory,
  - makes a direct connection between intransitive preferences and a ranking of all the items, and
  - has sample complexity guarantees.

# Model Preliminaries

- $n$  items, each with a known feature vector  $U_j \in \mathbb{R}^d$ .  
 $U := [U_1 U_2 \cdots U_n] \in \mathbb{R}^{d \times n}$ .

# Model Preliminaries

- $n$  items, each with a known feature vector  $U_j \in \mathbb{R}^d$ .  
 $U := [U_1 U_2 \cdots U_n] \in \mathbb{R}^{d \times n}$ .
- We assume a universal ranking, and  $w^* \in \mathbb{R}^d$  are unknown *judgment weights*, which signify the importance of each feature for the universal ranking.



# Model Preliminaries

- $n$  items, each with a known feature vector  $U_j \in \mathbb{R}^d$ .  
 $U := [U_1 U_2 \cdots U_n] \in \mathbb{R}^{d \times n}$ .
- We assume a universal ranking, and  $w^* \in \mathbb{R}^d$  are unknown *judgment weights*, which signify the importance of each feature for the universal ranking.
- $P := \{(i, j) \in [n] \times [n] : i < j\}$  is the set of all pairs.

# Model Preliminaries

- $n$  items, each with a known feature vector  $U_j \in \mathbb{R}^d$ .  
 $U := [U_1 U_2 \cdots U_n] \in \mathbb{R}^{d \times n}$ .
- We assume a universal ranking, and  $w^* \in \mathbb{R}^d$  are unknown *judgment weights*, which signify the importance of each feature for the universal ranking.
- $P := \{(i, j) \in [n] \times [n] : i < j\}$  is the set of all pairs.
- $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$  are  $m$  independent pairwise comparisons, where a pair  $(i_\ell, j_\ell) \in P$  has outcome  $y_\ell \in \{0, 1\}$  (1 means  $i_\ell$  beats  $j_\ell$ ).

# Salient Feature Preference Model

- $\tau : [n] \times [n] \rightarrow \text{Powerset}([d])$  is the known *selection function* that determines which features are used in each pairwise comparison.

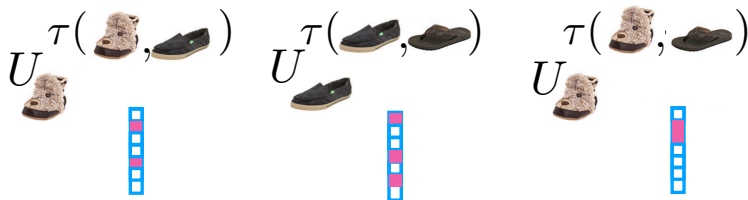


Image source: <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

## Salient Feature Preference Model

- $\tau : [n] \times [n] \rightarrow \text{Powerset}([d])$  is the known *selection function* that determines which features are used in each pairwise comparison.
- $y_\ell \sim \text{Bern}(\mathbb{P}(i_\ell >_B j_\ell))$  where  $i >_B j$  means '*i* beats *j*' and

$$\mathbb{P}(i >_B j) = \frac{\exp(\langle U_i^{\tau(i,j)}, w^* \rangle)}{\exp(\langle U_j^{\tau(i,j)}, w^* \rangle) + \exp(\langle U_i^{\tau(i,j)}, w^* \rangle)}. \quad (1)$$

## Salient Feature Preference Model

- $\tau : [n] \times [n] \rightarrow \text{Powerset}([d])$  is the known *selection function* that determines which features are used in each pairwise comparison.
- $y_\ell \sim \text{Bern}(\mathbb{P}(i_\ell >_B j_\ell))$  where  $i >_B j$  means ‘ $i$  beats  $j$ ’ and

$$\mathbb{P}(i >_B j) = \frac{\exp(\langle U_i^{\tau(i,j)}, w^* \rangle)}{\exp(\langle U_j^{\tau(i,j)}, w^* \rangle) + \exp(\langle U_i^{\tau(i,j)}, w^* \rangle)}. \quad (1)$$

- The utility of the item is dependent on the context of that particular comparison.

## Salient Feature Preference Model

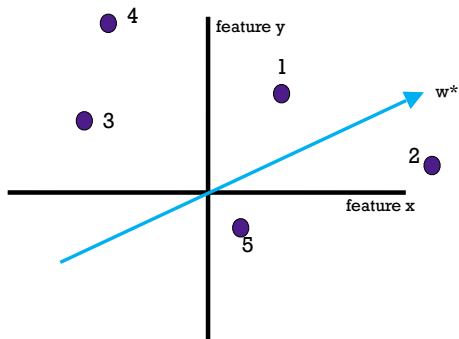
- $\tau : [n] \times [n] \rightarrow \text{Powerset}([d])$  is the known *selection function* that determines which features are used in each pairwise comparison.
- $y_\ell \sim \text{Bern}(\mathbb{P}(i_\ell >_B j_\ell))$  where  $i >_B j$  means ‘ $i$  beats  $j$ ’ and

$$\mathbb{P}(i >_B j) = \frac{\exp(\langle U_i^{\tau(i,j)}, w^* \rangle)}{\exp(\langle U_j^{\tau(i,j)}, w^* \rangle) + \exp(\langle U_i^{\tau(i,j)}, w^* \rangle)}. \quad (1)$$

- The utility of the item is dependent on the context of that particular comparison.
- The items are ranked by sorting the full feature utilities:  $\langle U_i, w^* \rangle$ . FBTL is a special case where  $\tau = [d]$ .

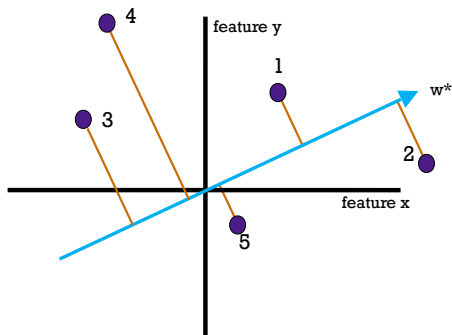
# Salient Features

Bradley-Terry-Luce with features (FBTL)



# Salient Features

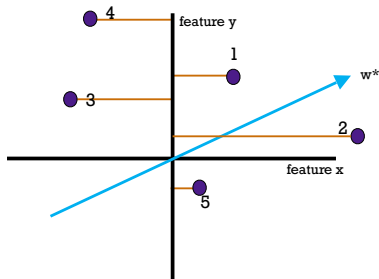
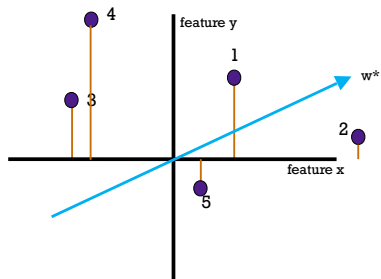
Bradley-Terry-Luce with features (FBTL)





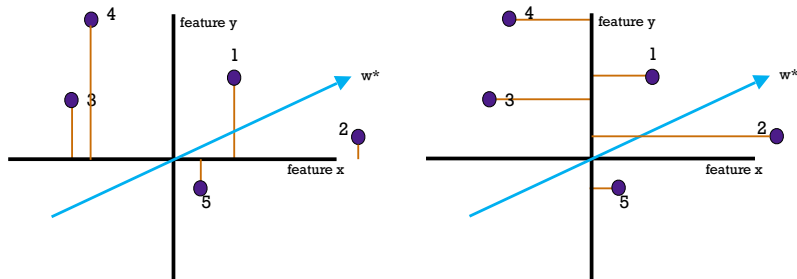
# Salient Features

Bradley-Terry-Luce with *salient* features



# Salient Features

Bradley-Terry-Luce with *salient* features



If the pair 4,2 uses only feature  $y$ ,  $4 > 2$ . If pair 2,5 uses only feature  $x$ ,  $2 > 5$ . But if 4,5 uses feature  $x$  or both features, then  $5 > 4$ , giving us intransitivity.

# Maximum Likelihood Estimation

Given observations  $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$ , item features  $U \in \mathbb{R}^{d \times n}$ , and a selection function  $\tau$ , the negative log-likelihood of  $w \in \mathbb{R}^d$  is

$$\mathcal{L}_m(w; U, S_m, \tau) = \sum_{\ell=1}^m \log(1 + \exp(u_{i_\ell, j_\ell})) - y_\ell u_{i_\ell, j_\ell}, \quad (2)$$

where  $u_{i_\ell, j_\ell} = \langle w, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle$ .

## Sample Complexity of MLE - Preliminaries

- Let

$$b^* := \max_{(i,j) \in P} |\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle|,$$

i.e. the maximum absolute difference between a pair of context dependent utilities

## Sample Complexity of MLE - Preliminaries

- Let

$$b^* := \max_{(i,j) \in P} |\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle|,$$

i.e. the maximum absolute difference between a pair of context dependent utilities

- For  $(i,j) \in P$ , let

$$Z_{(i,j)} := (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T$$
$$\lambda := \lambda_{\min}(\mathbb{E}Z_{(i,j)}),$$

where expectation is taken with respect to a uniformly chosen random pair of items from  $P$ .

## Sample Complexity of MLE - Preliminaries

- Let

$$b^* := \max_{(i,j) \in P} |\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle|,$$

i.e. the maximum absolute difference between a pair of context dependent utilities

- For  $(i,j) \in P$ , let

$$Z_{(i,j)} := (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T$$

$$\lambda := \lambda_{\min}(\mathbb{E}Z_{(i,j)}),$$

where expectation is taken with respect to a uniformly chosen random pair of items from  $P$ .

- Let

$$\beta := \max_{(i,j) \in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_{\infty}. \quad (3)$$

# Sample Complexity of MLE - Theorem

## Theorem 1

Let  $\hat{w}$  be the maximum likelihood estimator. Let  $\delta > 0$ . If  $\lambda > 0$  and

$$m \geq \max \left\{ C_1(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta), C_2 \frac{\log(2d/\delta)}{\lambda} \right\},$$

then with probability at least  $1 - \delta$ ,

$$\|w^* - \hat{w}\|_2 = O \left( \frac{\exp(b^*)}{\lambda} \sqrt{\frac{(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta)}{m}} \right)$$

where  $C_1, C_2$  are constants, probability is w.r.t both the random pairs and random outcomes of pairwise comparisons.

# Sample Complexity of MLE - Discussion

## Theorem 1

Let  $\delta > 0$ . If  $\lambda > 0$  and  $m \geq \max \left\{ C_1(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta), C_2 \frac{\log(2d/\delta)}{\lambda} \right\}$ , then with probability at least  $1 - \delta$ ,

$$\|w^* - \hat{w}\|_2 = O \left( \frac{\exp(b^*)}{\lambda} \sqrt{\frac{(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta)}{m}} \right)$$

- With roughly constant  $\beta, \lambda, b^*$ , we need  $\Omega(d \log(d/\delta))$  measurements for the error to be  $O(1)$ .



# Sample Complexity of MLE - Discussion

## Theorem 1

Let  $\delta > 0$ . If  $\lambda > 0$  and  $m \geq \max \left\{ C_1(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta), C_2 \frac{\log(2d/\delta)}{\lambda} \right\}$ , then with probability at least  $1 - \delta$ ,

$$\|w^* - \hat{w}\|_2 = O \left( \frac{\exp(b^*)}{\lambda} \sqrt{\frac{(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta)}{m}} \right)$$

- With roughly constant  $\beta, \lambda, b^*$ , we need  $\Omega(d \log(d/\delta))$  measurements for the error to be  $O(1)$ .
- Corollaries: (1)  $\tau$  selects all features, (2)  $\tau$  selects one feature per pair, and (3) learning the ranking

# Congressional Districts in the United States

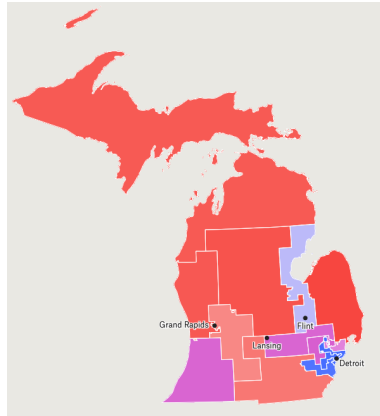


Image source: <https://projects.fivethirtyeight.com/redistricting-maps/michigan/>

# Intrascitivity



Image source: [Kaufman et al., 2017]

# Intransitivity

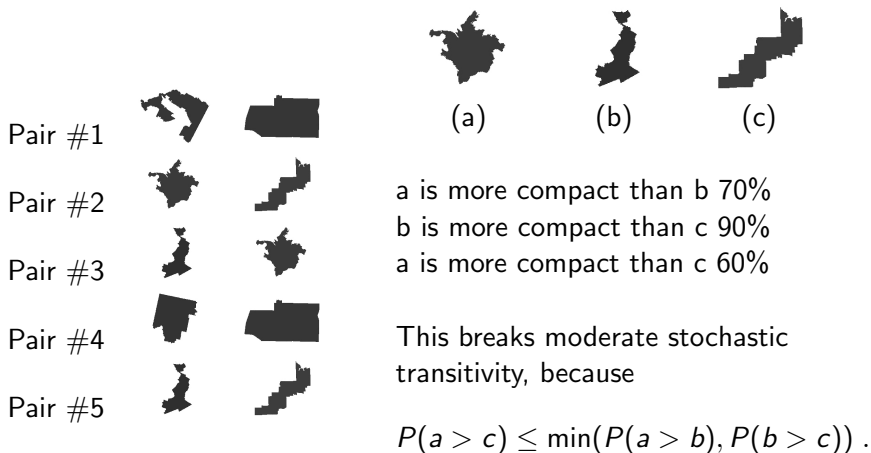


Image source: [Kaufman et al., 2017]

# Intransitivity

Pair #1



Pair #2



Pair #4



vs

$k$ -wise rankings



Image source: [Kaufman et al., 2017]

# Intransitivity



A pairwise “approach enables respondents to make each paired comparison independently of the others allows, and **may even encourage, them to use different dimensions for different comparisons**” [Kaufman et al., 2017].

# Intransitivity



Image source: [Kaufman et al., 2017]

A pairwise “approach enables respondents to make each paired comparison independently of the others allows, and **may even encourage, them to use different dimensions for different comparisons**” [Kaufman et al., 2017].

Instead of abandoning pairwise comparisons, we model this behavior.

# Experimental results

**Table:** Average Kendall tau correlation over individual rankings on test sets for district compactness. The number in parenthesis is the standard deviation.

Model:	Shiny1	Shiny2	UG1-j1	UG1-j2	UG1-j3	UG1-j4	UG1-j5
Salient features	<b>0.14</b> (.26)	<b>0.26</b> (.2)	<b>0.48</b> (.21)	<b>0.41</b> (.09)	<b>0.6</b> (.1)	0.14 (.14)	<b>0.42</b> (.09)
FBTL	0.09 (.22)	0.18 (.17)	0.2 (.12)	0.26 (.07)	0.45 (.15)	0.2 (.13)	0.06 (.14)
Ranking SVM	0.09 (.22)	0.18 (.17)	0.22 (.12)	0.26 (.07)	0.45 (.15)	0.2 (.13)	0.06 (.14)
RankNet	0.12 (.24)	0.24 (.18)	0.28 (.14)	0.37 (.08)	0.53 (.11)	<b>0.28</b> (.08)	0.15 (.15)



# Future Work

- Learn the selection function or features.
- Apply and extend the salient feature framework to other scenarios that involve human comparisons.
- Understand the impact of intransitivity in preference data on the sample complexity.

# References I



Abbasnejad, E., Sanner, S., Bonilla, E. V., and Poupart, P. (2013).

Learning community-based preferences via dirichlet process mixtures of gaussian processes.

In *Twenty-Third International Joint Conference on Artificial Intelligence*.



Benson, A. R., Kumar, R., and Tomkins, A. (2016).

On the relevance of irrelevant alternatives.

In *Proceedings of the 25th International Conference on World Wide Web*, pages 963–973. International World Wide Web Conferences Steering Committee.

## References II



Bordalo, P., Gennaioli, N., and Shleifer, A. (2013).  
Salience and consumer choice.  
*Journal of Political Economy*, 121(5):803–843.



Brown, T. C. and Peterson, G. L. (2009).  
An enquiry into the method of paired comparison: reliability,  
scaling, and thurstone's law of comparative judgment.  
*Gen Tech. Rep. RMRS-GTR-216WWW. Fort Collins, CO: US  
Department of Agriculture, Forest Service, Rocky Mountain  
Research Station. 98 p., 216.*

## References III



Chen, S. and Joachims, T. (2016).

Predicting matchups and preferences in context.

In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 775–784. ACM.



Guiver, J. and Snelson, E. (2009).

Bayesian inference for plackett-luce ranking models.

In *proceedings of the 26th annual international conference on machine learning*, pages 377–384.



Kamishima, T. and Akaho, S. (2009).

Efficient clustering for orders.

In *Mining complex data*, pages 261–279. Springer.

## References IV



Kaufman, A., King, G., and Komisarchik, M. (2017).  
How to measure legislative district compactness if you only  
know it when you see it.  
*American Journal of Political Science*.



Kleinberg, J., Mullainathan, S., and Ugander, J. (2017).  
Comparison-based choices.  
In *Proceedings of the 2017 ACM Conference on Economics  
and Computation*, pages 127–144. ACM.



Lopes, P., Liapis, A., and Yannakakis, G. N. (2017).  
Modelling affect for horror soundscapes.  
*IEEE Transactions on Affective Computing*, 10(2):209–222.

# References V



Niranjan, U. and Rajkumar, A. (2017).

Inductive pairwise ranking: going beyond the  $n \log(n)$  barrier.  
*In Thirty-First AAAI Conference on Artificial Intelligence.*



Pfannschmidt, K., Gupta, P., and Hüllermeier, E. (2019).

Learning choice functions.  
*CoRR*, abs/1901.10860.



Ragain, S. and Ugander, J. (2016).

Pairwise choice markov chains.

*In Advances in Neural Information Processing Systems*, pages 3198–3206.

## References VI



Rajkumar, A., Ghoshal, S., Lim, L.-H., and Agarwal, S. (2015).

Ranking from stochastic pairwise preferences: Recovering condorcet winners and tournament solution sets at the top. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 665–673. JMLR.org.






Rieskamp, J., Busemeyer, J. R., and Mellers, B. A. (2006).

Extending the bounds of rationality: Evidence and theories of preferential choice.

*Journal of Economic Literature*, 44(3):631–661.

## References VII

-  Rosenfeld, N., Oshiba, K., and Singer, Y. (2019). Predicting choice with set-dependent aggregation. *arXiv preprint arXiv:1906.06365*.
-  Seshadri, A., Peysakhovich, A., and Ugander, J. (2019). Discovering context effects from raw choice data. *ICML*.
-  Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281.



## References VIII



Yang, D. and Wakin, M. B. (2015).

Modeling and recovering non-transitive pairwise comparison matrices.

*2015 International Conference on Sampling Theory and Applications, SampTA 2015*, pages 39–43.

## Additional Intransitivity Table

Data Set	Items	Match-ups	Unique Pairs	Valid Triplets
NBA 2015	30	1238	430 (99%)	2654 (65%)
Tennis 2014	288	2560	1080 (23%)	651 (14%)
Nascar [Guiver and Snelson, 2009]	83	64596	2875 (84%)	65003 (71%)
Jester	100	145354552	4950 (100%)	161700 (100%)
Sushi-A [Kamishima and Akaho, 2009]	10	225000	45 (100%)	120 (100%)
Sushi-B [Kamishima and Akaho, 2009]	100	225000	4809 (97%)	139992 (86%)
District [Kaufman et al., 2017]	122	5150	94 (1%)	48 (.02%)
Car [Abbasnejad et al., 2013]	10	2973	45 (100%)	120 (100%)
Sonancia [Lopes et al., 2017]	874	671	175 (20%)	108 (13%)
New Yorker Captions-508	29	31043	406 (100%)	3592 (98%)

**Table:** Match-ups is the number of pairwise comparison samples. Valid triplets are triplets of items  $(i, j, k)$  where data has been collected on  $i$  vs  $j$ ,  $j$  vs  $k$ , and  $k$  vs  $i$ . Back to [Main Intransitivity Table](#)

NBA: <https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018> Tennis: <https://www.kaggle.com/jordangoblet/atp-tour-20002016> Nascar: <http://personal.psu.edu/drh20/code/btmatlab/> Jester: <http://eigentaste.berkeley.edu/dataset/> Sushi: <http://www.kamishima.net/sushi/> Car: <http://users.cecs.anu.edu.au/~u4940058/CarPreferences.html> Sonancia: <http://plt.institutedigitalgames.com/datasets.php> New Yorker: <https://github.com/nextml/caption-contest-data/blob/master/contests/responses/508-round2-dueling-responses.csv.zip>

# Stochastic Transitivity Definitions

Let  $P_{ij} = \mathbb{P}(i >_B j)$  and  $T = \{(i, j, k) \in [n]^3 : P_{ij} > .5, P_{jk} > .5\}$ .  
Then  $(i, j, k) \in T$  satisfies strong stochastic transitivity if  $P_{ik} \geq \max\{P_{ij}, P_{jk}\}$ , moderate stochastic transitivity if  $P_{ik} \geq \min\{P_{ij}, P_{jk}\}$ , and weak stochastic transitivity if  $P_{ik} \geq .5$   
Back to [Main Intransitivity Table](#)