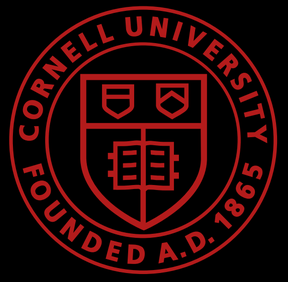
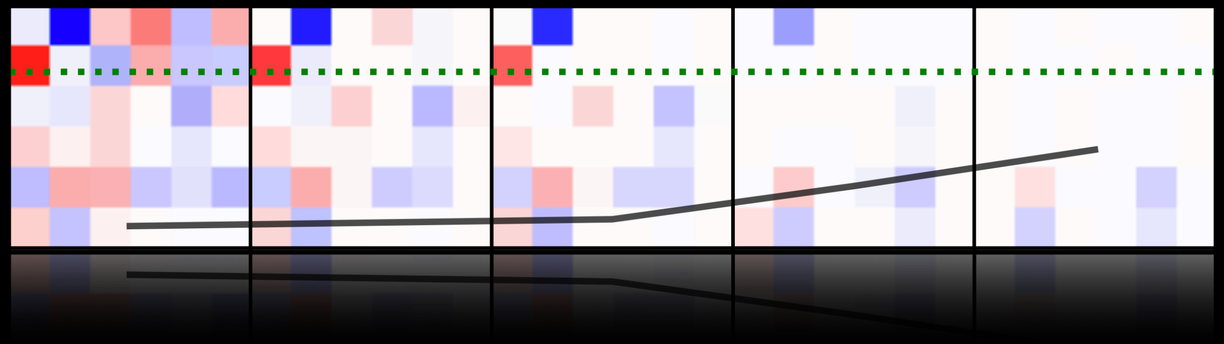


Slides: bit.ly/lcl-slides-2
Preprint: bit.ly/lcl-paper
Code: bit.ly/lcl-code
Data: bit.ly/lcl-data



Learning Interpretable Feature Context Effects in Discrete Choice

Kiran Tomlinson

@ Stanford SOAL

research with Austin R. Benson

1. Choices and context effects

Discrete choices are everywhere

amazon.com

Tervis Stanford Cardinal Logo Tumbler with Emblem and Black Lid 16oz Mug, Clear

★★★★★ 571

$16.99

✓prime FREE Delivery **Sun, Nov 22**

+4

Champion NCAA Men's Eco Powerblend Pocketless 1/4 Zip

★★★★☆ 18

$53.00

FREE Delivery for Prime members

Champion NCAA Men's Shirt Short Sleeve Officially Licensed Team Color Tee

★★★★½ 1,486

$18.80

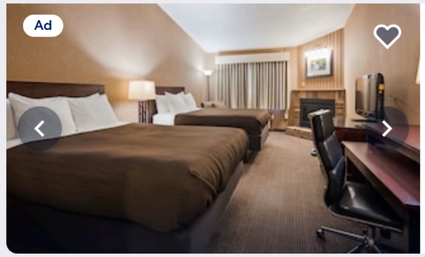
✓prime

Top of the World NCAA Mens Adjustable Vintage Team Icon hat

★★★★★ 7

$19.99

✓prime



Best Western University Inn

Ithaca

Expedia

Black Friday / Cyber Monday Deals Now

Free Shuttle Transportation, Grab & Go Breakfast, WiFi & Parking. Pet friendly, Outdoor Pool, Fitness Center. Sanitizing Daily

Breakfast included

3.9/5 Good (999 reviews)

$63

per night

$71 total

Includes taxes & fees

Quality Inn Ithaca - University Area

Ithaca

Black Friday / Cyber Monday Deals Now

Complimentary Breakfast. Free Airport Shuttle, WiFi & parking. Close to Ithaca College & Cornell University. Pets welcome.

Breakfast included

3.6/5 Good (694 reviews)

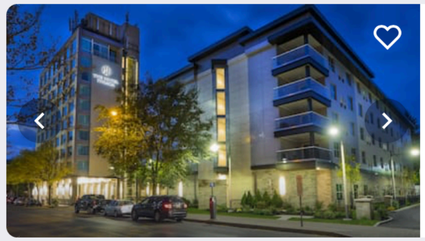
Member Price available

$59

per night

$66 total

Includes taxes & fees



Hotel Ithaca

Ithaca

Member Price available

$94

per night

$106 total

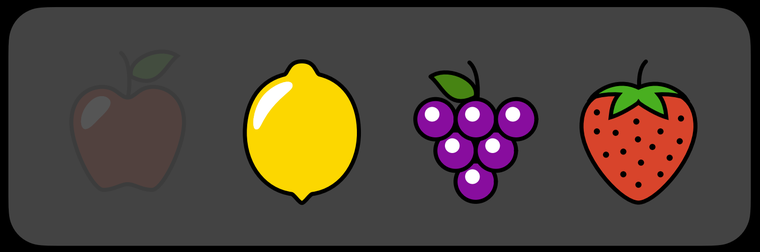
Includes taxes & fees

4.0/5 Very Good (842 reviews)

“The fundamental problem of discrete choice”

“The fundamental problem of discrete choice”

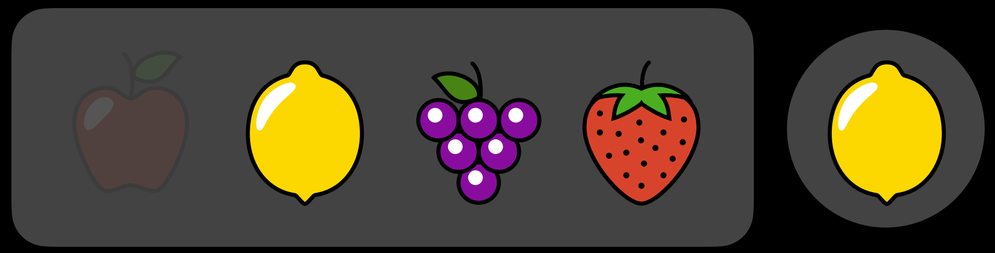
choice set



“The fundamental problem of discrete choice”

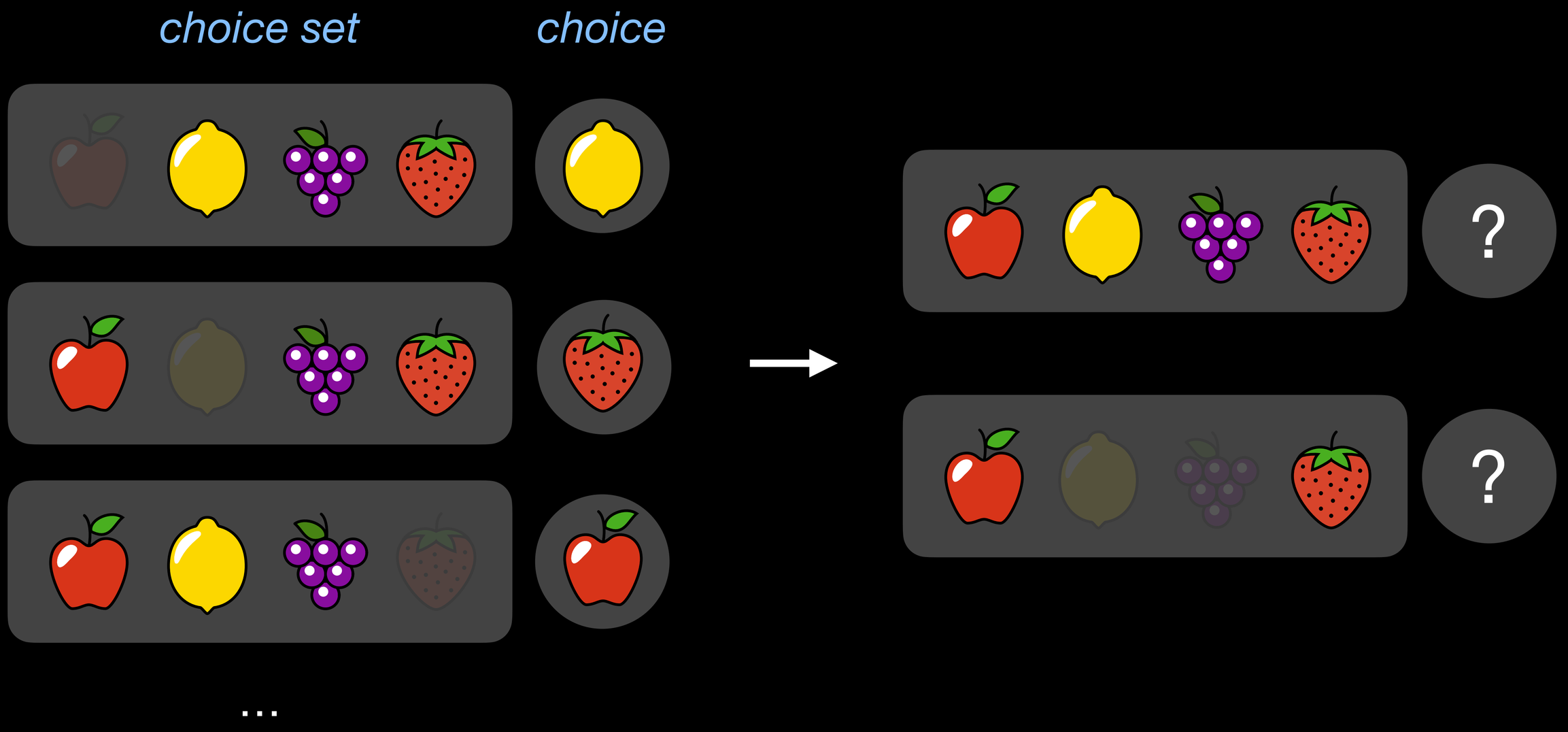
choice set

choice



“The fundamental problem of discrete choice”

“The fundamental problem of discrete choice”



The classic model: *multinomial logit (MNL)*

(McFadden, *Frontiers in Econometrics* 1973)

The classic model: *multinomial logit (MNL)*

(McFadden, *Frontiers in Econometrics* 1973)

Assume *item* i has *utility* u_i

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

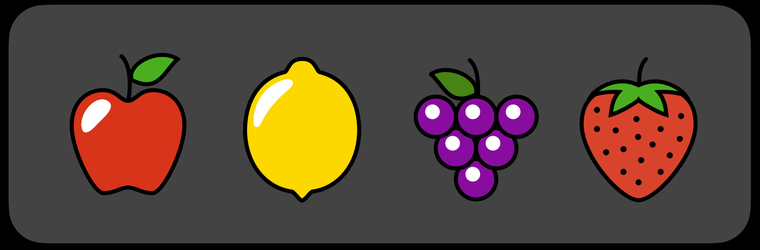
The classic model: *multinomial logit (MNL)*

(McFadden, *Frontiers in Econometrics* 1973)

Assume *item* i has *utility* u_i

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

C



u_i	1	-1	0	2
$\Pr(i \mid C)$	.24	.03	.09	.64

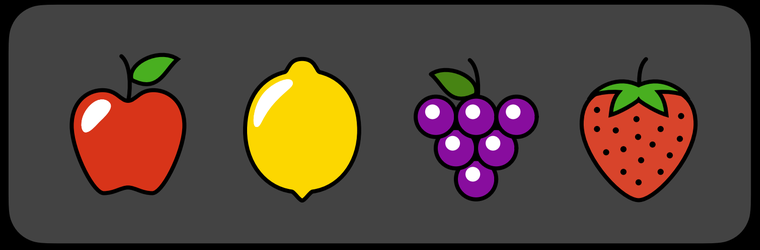
The classic model: *multinomial logit (MNL)*

(McFadden, *Frontiers in Econometrics* 1973)

Assume *item* i has *utility* u_i

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

C



u_i	1	-1	0	2
$\Pr(i \mid C)$	.24	.03	.09	.64

Econometric derivation:

- Draw random utilities $u_i + \epsilon_i$ for each item (ϵ_i i.i.d. Gumbel)
- Rational agent picks $\text{argmax}_{i \in C}\, u_i + \epsilon_i$

→ MNL

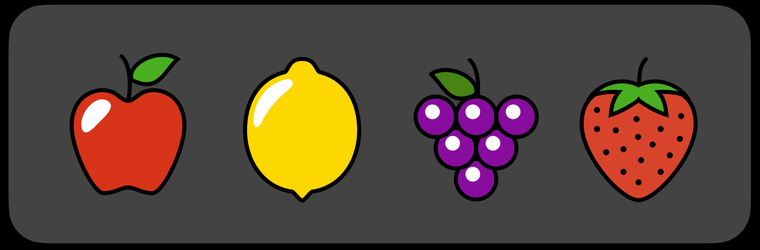
The classic model: *multinomial logit (MNL)*

(McFadden, *Frontiers in Econometrics* 1973)

Assume *item* i has *utility* u_i

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

C



u_i	1	-1	0	2
$\Pr(i \mid C)$	.24	.03	.09	.64

Econometric derivation:

- Draw random utilities $u_i + \epsilon_i$ for each item (ϵ_i i.i.d. Gumbel)
- Rational agent picks $\text{argmax}_{i \in C}\, u_i + \epsilon_i$

→ MNL

Unique choice model satisfying
independence of irrelevant alternatives (IIA):

(Luce, *Individual Choice Behavior* 1959)

$$\frac{\Pr(i \mid C)}{\Pr(j \mid C)} = \frac{\Pr(i \mid C')}{\Pr(j \mid C')}$$

Learning an MNL from choice data

Learning an MNL from choice data

Likelihood of utilities u given dataset $\mathscr{D}$: $\mathscr{L}(u;\mathscr{D}) = \prod_{(i,C)\in\mathscr{D}} \frac{\exp(u_i)}{\sum_{j\in C}\exp(u_j)}$

Learning an MNL from choice data

Likelihood of utilities u given dataset $\mathscr{D}$: $$\mathscr{L}(u;\mathscr{D}) = \prod_{(i,C)\in\mathscr{D}} \frac{\exp(u_i)}{\sum_{j\in C}\exp(u_j)}$$

Log-likelihood is concave: $$\ell(u;\mathscr{D}) = \sum_{(i,C)\in\mathscr{D}} u_i - \log \sum_{j\in C}\exp(u_j)$$

Learning an MNL from choice data

Likelihood of utilities u given dataset $\mathscr{D}$: $\mathscr{L}(u;\mathscr{D}) = \prod_{(i,C)\in\mathscr{D}} \frac{\exp(u_i)}{\sum_{j\in C}\exp(u_j)}$

Log-likelihood is concave: $\ell(u;\mathscr{D}) = \sum_{(i,C)\in\mathscr{D}} u_i - \log \sum_{j\in C}\exp(u_j)$

→ gradient descent to learn u
(SGD, Adam, …)

Problem for MNL: *context effects*

Problem for MNL: *context effects*

The choice set influences preferences.

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise
(Simonson, 1989)

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise
(Simonson, 1989)

Problem for MNL: *context effects*

The choice set influences preferences.

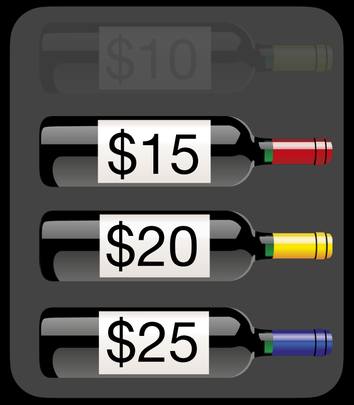
Compromise
(Simonson, 1989)

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise
(Simonson, 1989)

★

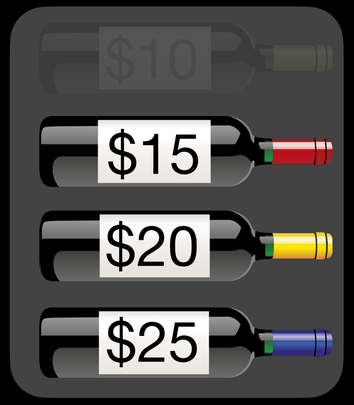


Problem for MNL: *context effects*

The choice set influences preferences.

Compromise
(Simonson, 1989)

★



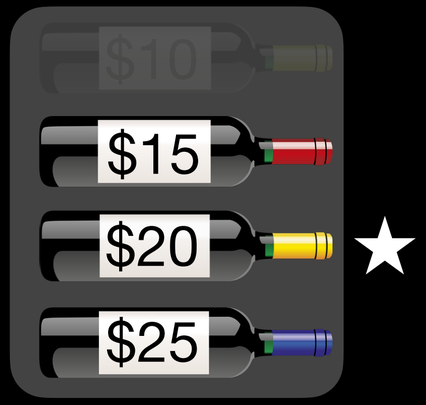
★

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



Asymmetric dominance

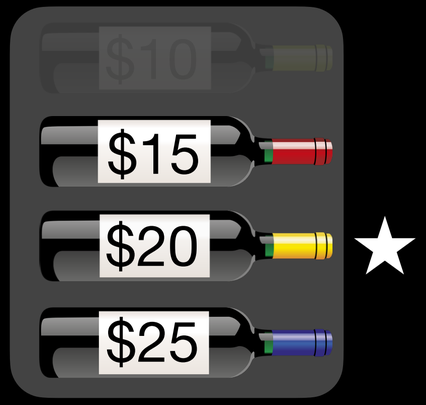
(Huber et al., 1982)

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



Asymmetric dominance

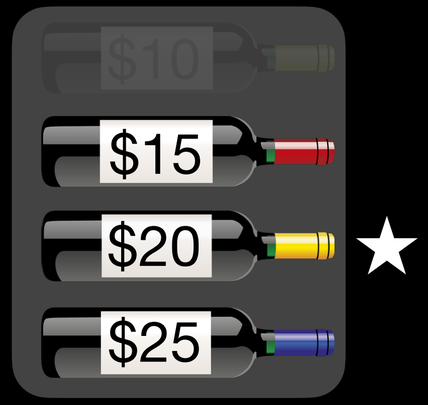
(Huber et al., 1982)

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



Asymmetric dominance

(Huber et al., 1982)

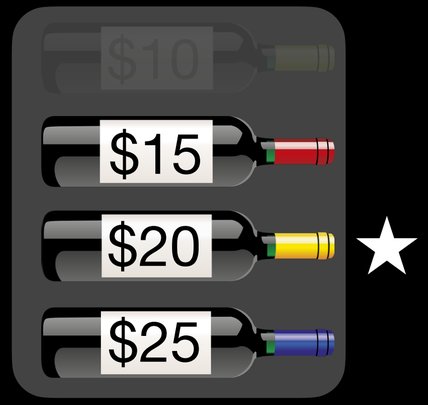
(Ariely, 2008)

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



Asymmetric dominance

(Huber et al., 1982)

(Ariely, 2008)

Similarity

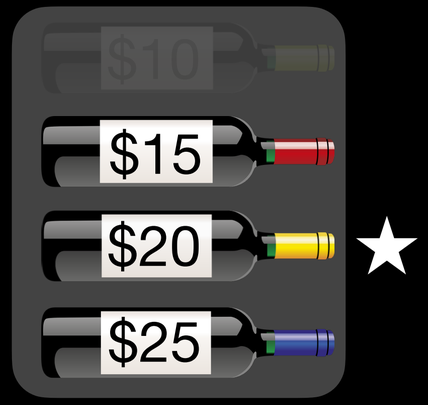
(Tversky, 1972)

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



Asymmetric dominance

(Huber et al., 1982)

(Ariely, 2008)

Similarity

(Tversky, 1972)

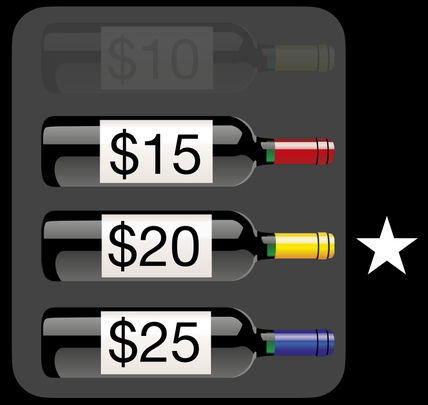
David Perdue* ✔	Rep.	2,462,617	**49.7%**
Jon Ossoff ✔	Dem.	2,374,519	**47.9**

Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



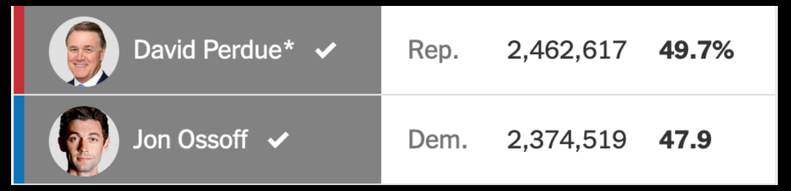
Asymmetric dominance

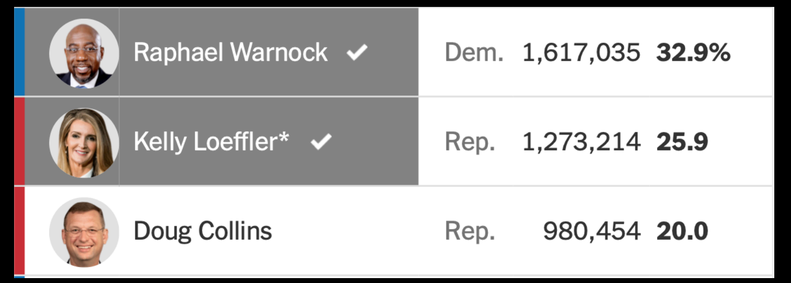
(Huber et al., 1982)

(Ariely, 2008)

Similarity

(Tversky, 1972)



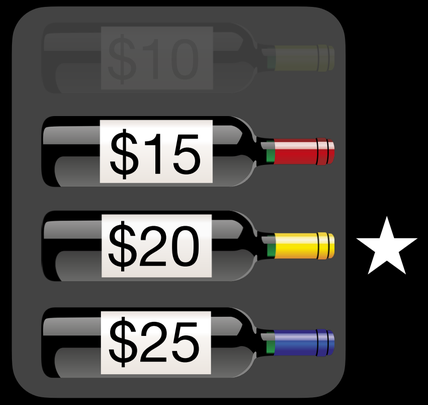


Problem for MNL: *context effects*

The choice set influences preferences.

Compromise

(Simonson, 1989)



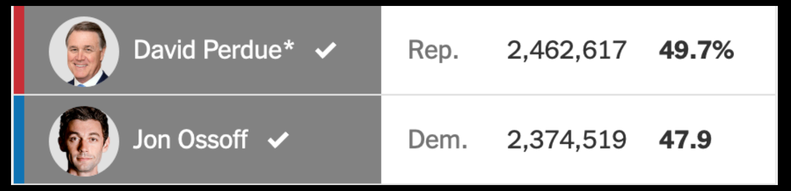
Asymmetric dominance

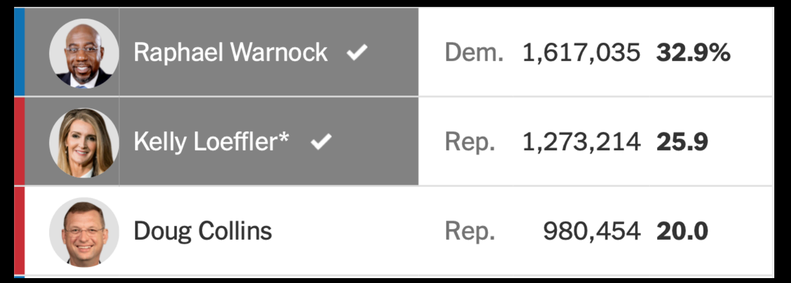
(Huber et al., 1982)

(Ariely, 2008)

Similarity

(Tversky, 1972)





IIA violations:

$$\frac{\Pr(i \mid C)}{\Pr(j \mid C)} \neq \frac{\Pr(i \mid C')}{\Pr(j \mid C')}$$

Natural context effect model: *CDM*

(Seshadri, Peysakhovich, & Ugander, ICML 2019)

Natural context effect model: *CDM*

(Seshadri, Peysakhovich, & Ugander, ICML 2019)

Item j exerts *pull* u_{ij} on item i, item utility is sum of pulls:

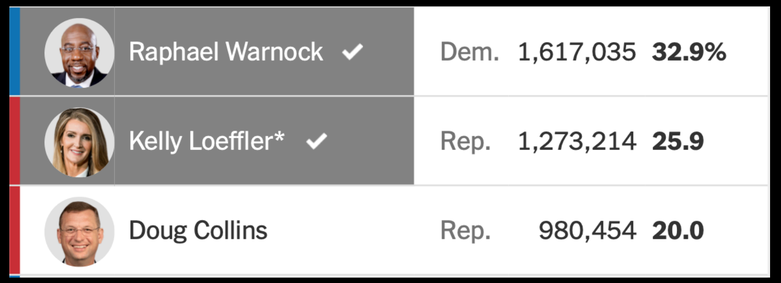
$$\Pr(i \mid C) = \frac{\exp\left(\sum_{k \in C \setminus i} u_{ik}\right)}{\sum_{j \in C} \exp\left(\sum_{k \in C \setminus i} u_{jk}\right)}$$

Natural context effect model: *CDM*

(Seshadri, Peysakhovich, & Ugander, ICML 2019)

Item j exerts *pull* u_{ij} on item i, item utility is sum of pulls:

$$\Pr(i \mid C) = \frac{\exp\left(\sum_{k \in C \setminus i} u_{ik}\right)}{\sum_{j \in C} \exp\left(\sum_{k \in C \setminus i} u_{jk}\right)}$$



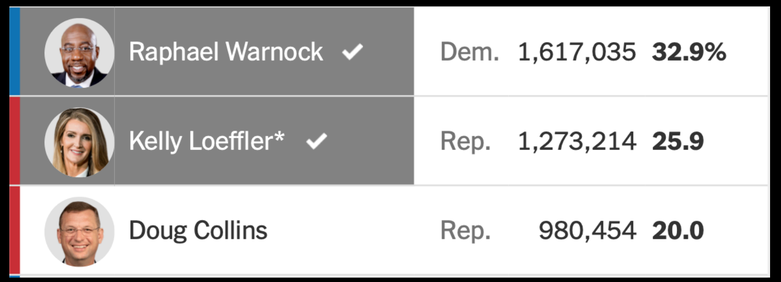
$$u_{\text{Loeffler, Collins}} < 0$$
$$u_{\text{Collins, Loeffler}} < 0$$

Natural context effect model: *CDM*

(Seshadri, Peysakhovich, & Ugander, ICML 2019)

Item j exerts *pull* u_{ij} on item i, item utility is sum of pulls:

$$\Pr(i \mid C) = \frac{\exp\left(\sum_{k \in C \setminus i} u_{ik}\right)}{\sum_{j \in C} \exp\left(\sum_{k \in C \setminus i} u_{jk}\right)}$$



$u_{\text{Loeffler, Collins}} < 0$

$u_{\text{Collins, Loeffler}} < 0$

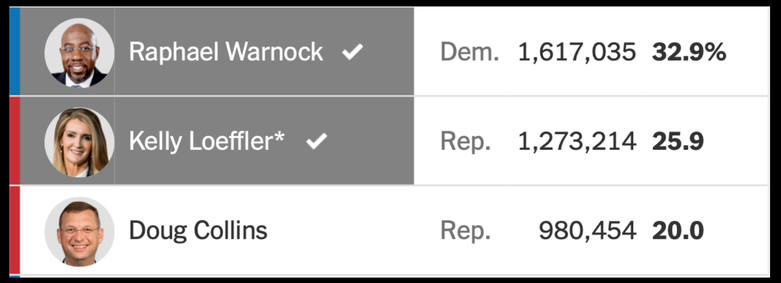
$u_{\text{Print\&Web, Print}} > 0$

Natural context effect model: *CDM*

(Seshadri, Peysakhovich, & Ugander, ICML 2019)

Item j exerts *pull* u_{ij} on item i, item utility is sum of pulls:

$$\Pr(i \mid C) = \frac{\exp\left(\sum_{k \in C \setminus i} u_{ik}\right)}{\sum_{j \in C} \exp\left(\sum_{k \in C \setminus i} u_{jk}\right)}$$



$u_{\text{Loeffler, Collins}} < 0$

$u_{\text{Collins, Loeffler}} < 0$

$u_{\text{Print\&Web, Print}} > 0$

Assumes no higher-order effects
(i.e., context effects decompose additively into effects of items)

2. Item features and the LCL

Choice models with *item features*

Choice models with *item features*

So far, models have per-item parameters

Choice models with *item features*

So far, models have per-item parameters

→ can’t generalize to new items not in training set

Choice models with *item features*

So far, models have per-item parameters

→ can’t generalize to new items not in training set

→ hard to learn utilities for rare items

Choice models with *item features*

So far, models have per-item parameters

→ can’t generalize to new items not in training set

→ hard to learn utilities for rare items

→ too many parameters with many items

Choice models with *item features*

So far, models have per-item parameters

→ can’t generalize to new items not in training set

→ hard to learn utilities for rare items

→ too many parameters with many items

Use item features:

genre: drama,
in_top_10: True,
has_new_episodes: True,
producer: Netflix

genre: comedy,
in_top_10: False,
has_new_episodes: False,
producer: NBC

genre: drama,
in_top_10: True,
has_new_episodes: False,
producer: Netflix

genre: reality,
in_top_10: True,
has_new_episodes: False,
producer: Banijay

MNL with item features: *conditional logit*

MNL with item features: *conditional logit*

Feature vector $x_i \in \mathbb{R}^d$ for each item i

Preference vector $\theta \in \mathbb{R}^d$

MNL with item features: *conditional logit*

Feature vector $x_i \in \mathbb{R}^d$ for each item i
Preference vector $\theta \in \mathbb{R}^d$

MNL:

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

MNL with item features: *conditional logit*

Feature vector $x_i \in \mathbb{R}^d$ for each item i
Preference vector $\theta \in \mathbb{R}^d$

MNL:

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

Conditional logit:

$$\Pr(i \mid C) = \frac{\exp(\theta^T x_i)}{\sum_{j \in C} \exp(\theta^T x_j)}$$

MNL with item features: *conditional logit*

Feature vector $x_i \in \mathbb{R}^d$ for each item i
Preference vector $\theta \in \mathbb{R}^d$

MNL:

$$\Pr(i \mid C) = \frac{\exp(u_i)}{\sum_{j \in C} \exp(u_j)}$$

Conditional logit:

$$\Pr(i \mid C) = \frac{\exp(\theta^T x_i)}{\sum_{j \in C} \exp(\theta^T x_j)}$$

Preference coefficient θ_k is easy to interpret: importance of the k^{th} feature

Incorporating *feature context effects* into conditional logit

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$ → *Contextual utility*: $u_{i,C} = [\theta + F(C)]^T x_i$

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$ → *Contextual utility*: $u_{i,C} = [\theta + F(C)]^T x_i$

Simplifying assumptions on $F(C)$:

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$ → *Contextual utility*: $u_{i,C} = [\theta + F(C)]^T x_i$

Simplifying assumptions on $F(C)$:

1. *Additivity*: $F(C) \propto \sum_{j \in C} f(x_j)$ for some function f

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$ → *Contextual utility*: $u_{i,C} = [\theta + F(C)]^T x_i$

Simplifying assumptions on $F(C)$:

1. *Additivity*: $F(C) \propto \sum_{j \in C} f(x_j)$ for some function f

2. *Dilution*: $F(C) = \frac{1}{|C|} \sum_{j \in C} f(x_j)$

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$ → *Contextual utility*: $u_{i,C} = [\theta + F(C)]^T x_i$

Simplifying assumptions on $F(C)$:

1. *Additivity*: $F(C) \propto \sum_{j \in C} f(x_j)$ for some function f
2. *Dilution*: $F(C) = \frac{1}{|C|} \sum_{j \in C} f(x_j)$
3. *Linearity*: $f(x_j) = Ax_j$ for some matrix $A \in \mathbb{R}^{d \times d}$

Incorporating *feature context effects* into conditional logit

Conditional logit utility: $u_i = \theta^T x_i$ → *Contextual utility*: $u_{i,C} = [\theta + F(C)]^T x_i$

Simplifying assumptions on $F(C)$:

1. *Additivity*: $F(C) \propto \sum_{j \in C} f(x_j)$ for some function f
2. *Dilution*: $F(C) = \frac{1}{|C|} \sum_{j \in C} f(x_j)$
3. *Linearity*: $f(x_j) = Ax_j$ for some matrix $A \in \mathbb{R}^{d \times d}$

→ $u_{i,C} = (\theta + Ax_C)^T x_i$ ($x_C = \frac{1}{|C|} \sum_{j \in C} x_j$ is the *mean feature vector*)

The *Linear Context Logit (LCL)*

The *Linear Context Logit (LCL)*

$$\Pr(i \mid C) = \frac{\exp([\theta + Ax_C]^T x_i)}{\sum_{j \in C} \exp([\theta + Ax_C]^T x_j)}$$

The *Linear Context Logit (LCL)*

$$\Pr(i \mid C) = \frac{\exp([\theta + Ax_C]^T x_i)}{\sum_{j \in C} \exp([\theta + Ax_C]^T x_j)}$$

→ convex negative log-likelihood

The *Linear Context Logit (LCL)*

$$\Pr(i \mid C) = \frac{\exp([\theta + Ax_C]^T x_i)}{\sum_{j \in C} \exp([\theta + Ax_C]^T x_j)}$$

→ convex negative log-likelihood

→ θ: *base preference coefficients*

The *Linear Context Logit (LCL)*

$$\Pr(i \mid C) = \frac{\exp([\theta + Ax_C]^T x_i)}{\sum_{j \in C} \exp([\theta + Ax_C]^T x_j)}$$

→ convex negative log-likelihood

→ θ: *base preference coefficients*

→ $A_{pq} > 0$: when q is *high* in the choice set, p is *more* preferred

The *Linear Context Logit (LCL)*

$$\Pr(i \mid C) = \frac{\exp([\theta + Ax_C]^T x_i)}{\sum_{j \in C} \exp([\theta + Ax_C]^T x_j)}$$

→ convex negative log-likelihood

→ θ: *base preference coefficients*

→ $A_{pq} > 0$: when q is *high* in the choice set, p is *more* preferred

→ $A_{pq} < 0$: when q is *high* in the choice set, p is *less* preferred

LCL example: restaurant selection

LCL example: restaurant selection

Icons by icons8.com

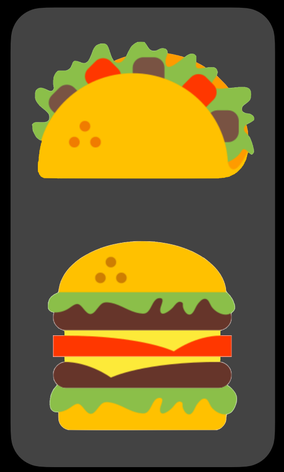
LCL example: restaurant selection

item features:
- price
- service speed
- wine selection

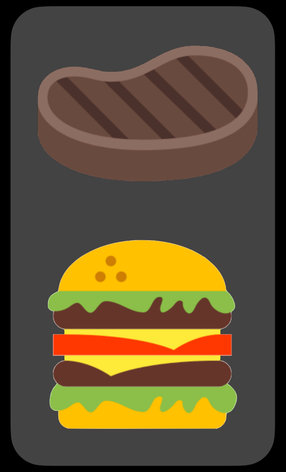
Icons by icons8.com

LCL example: restaurant selection

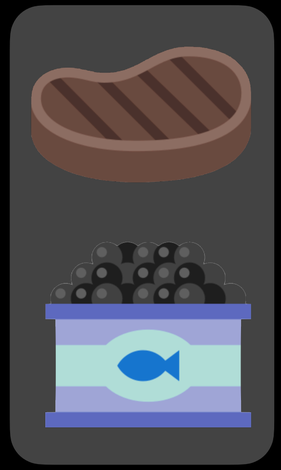
item features:
- price
- service speed
- wine selection



C_1



C_2



C_3

Icons by icons8.com

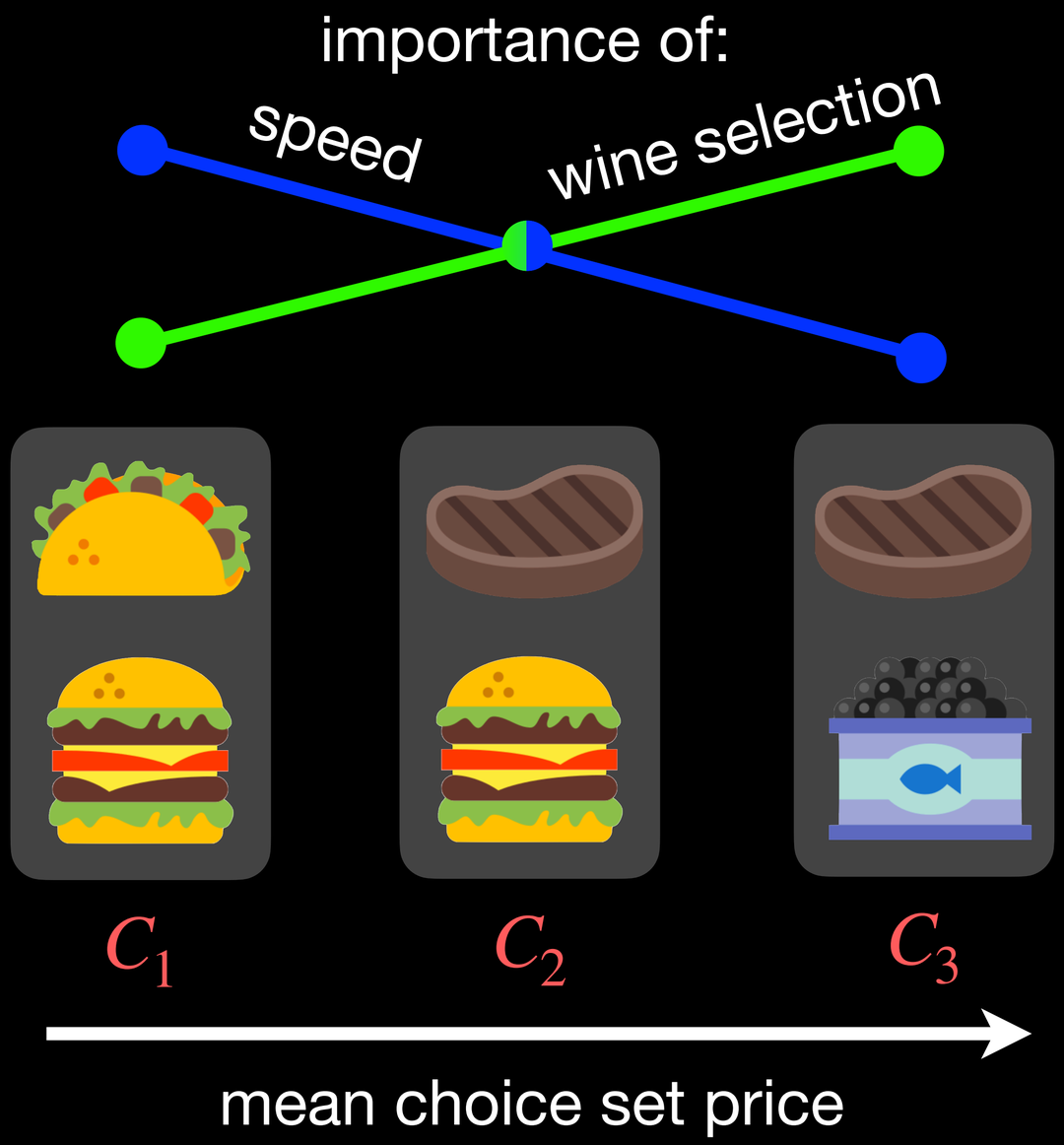
LCL example: restaurant selection

item features:
- price
- service speed
- wine selection

Icons by icons8.com

LCL example: restaurant selection

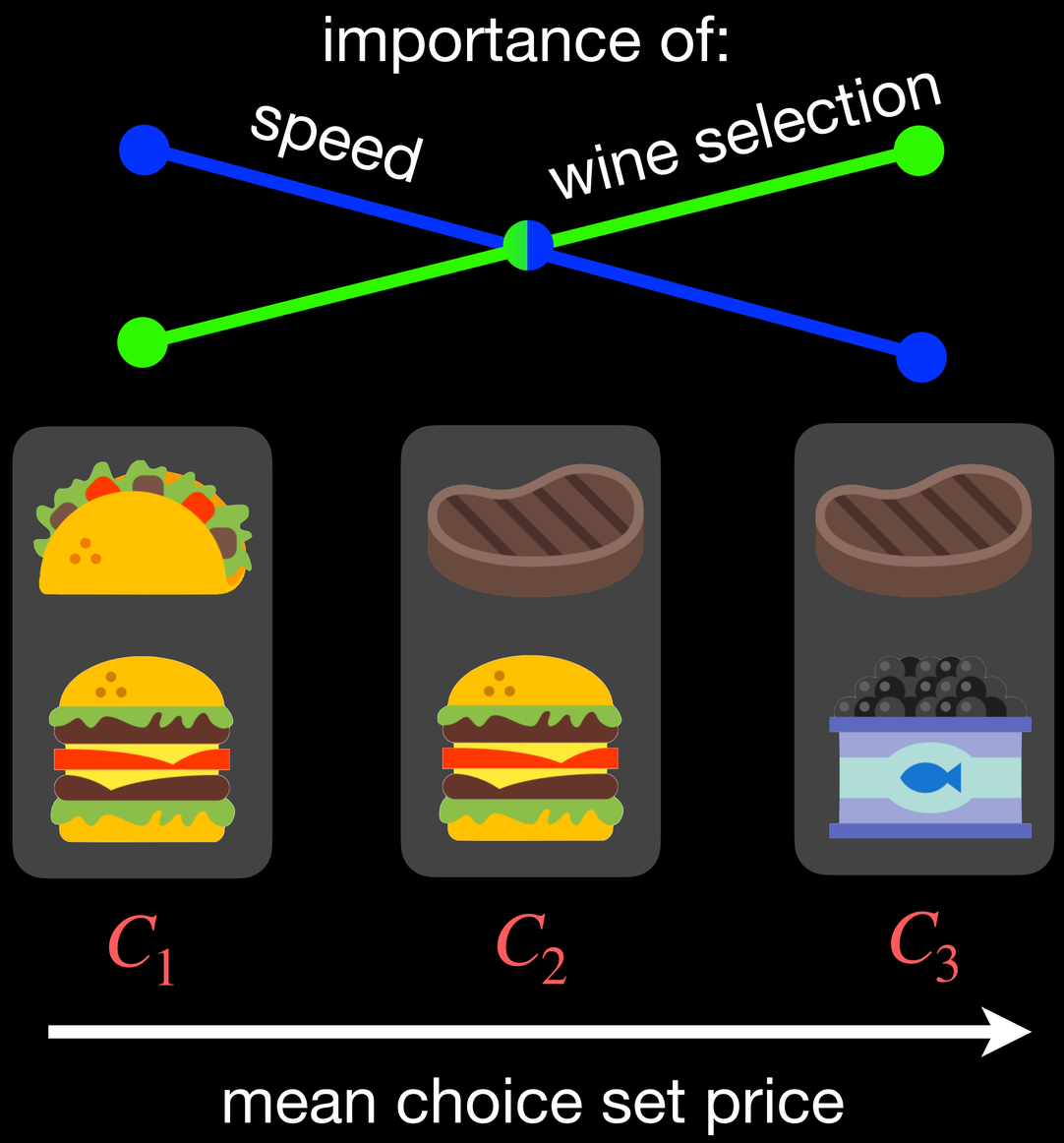
item features:
- price
- service speed
- wine selection



Icons by icons8.com

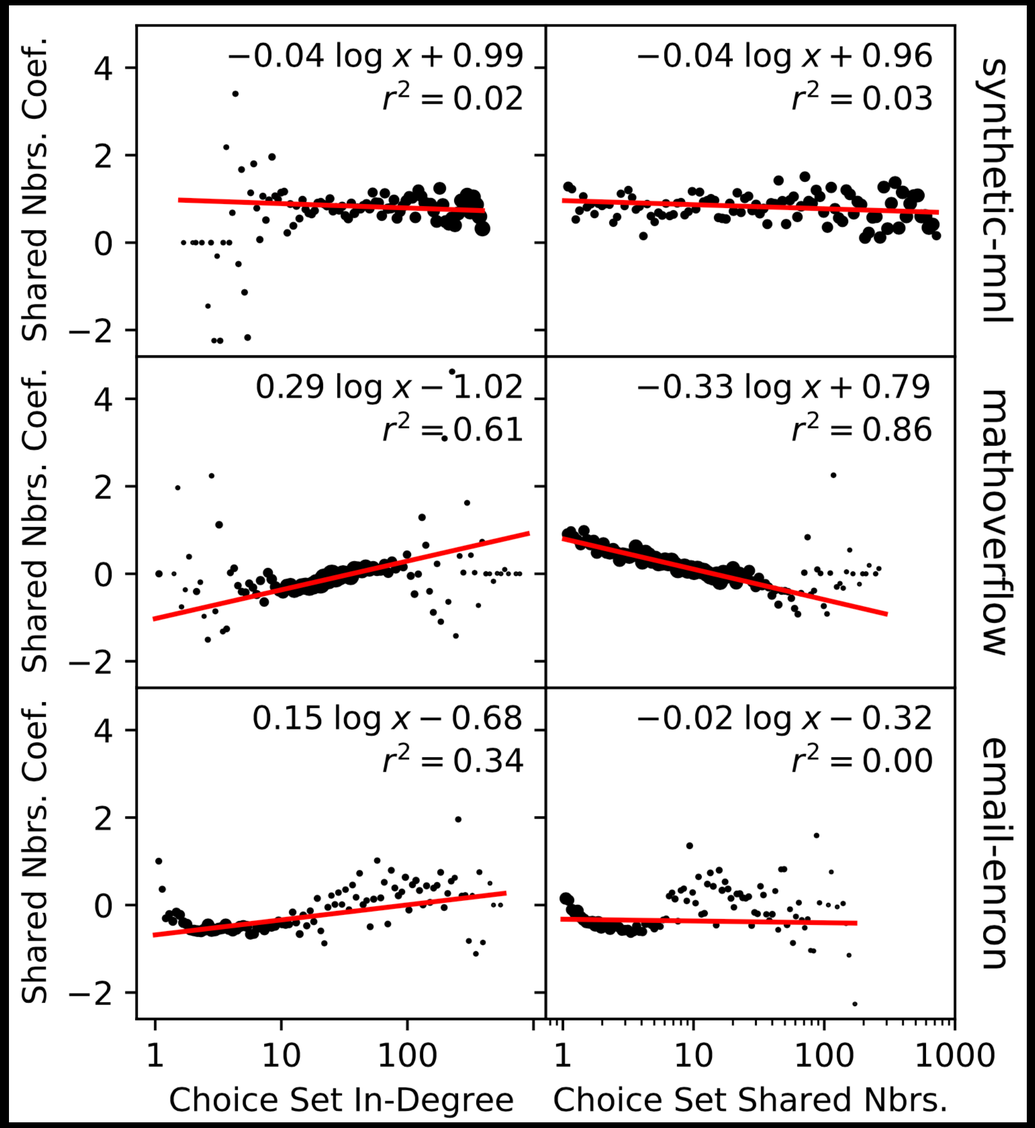
LCL example: restaurant selection

item features:
- price
- service speed
- wine selection



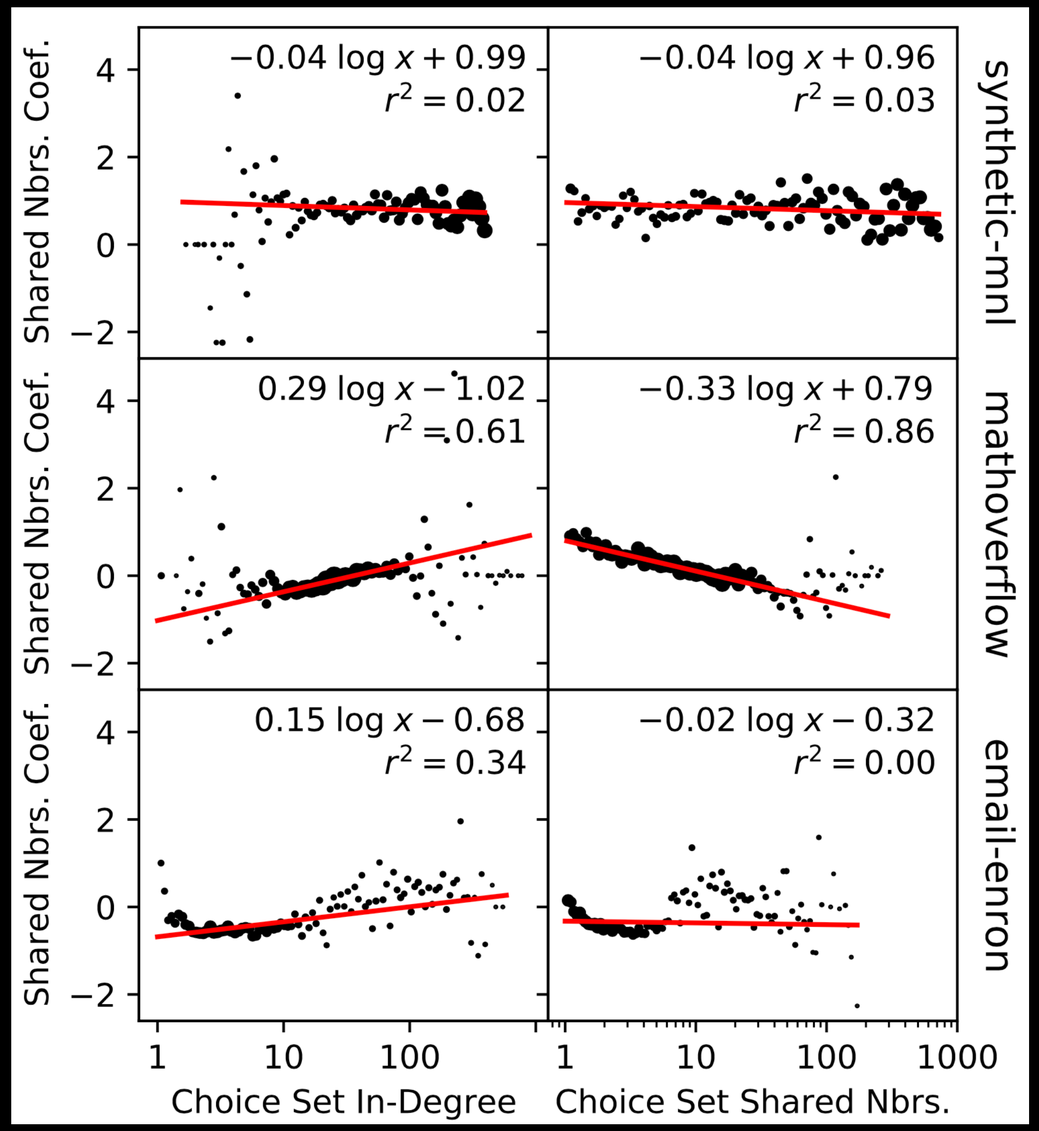
$$A = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Linear feature context effects appear in real data

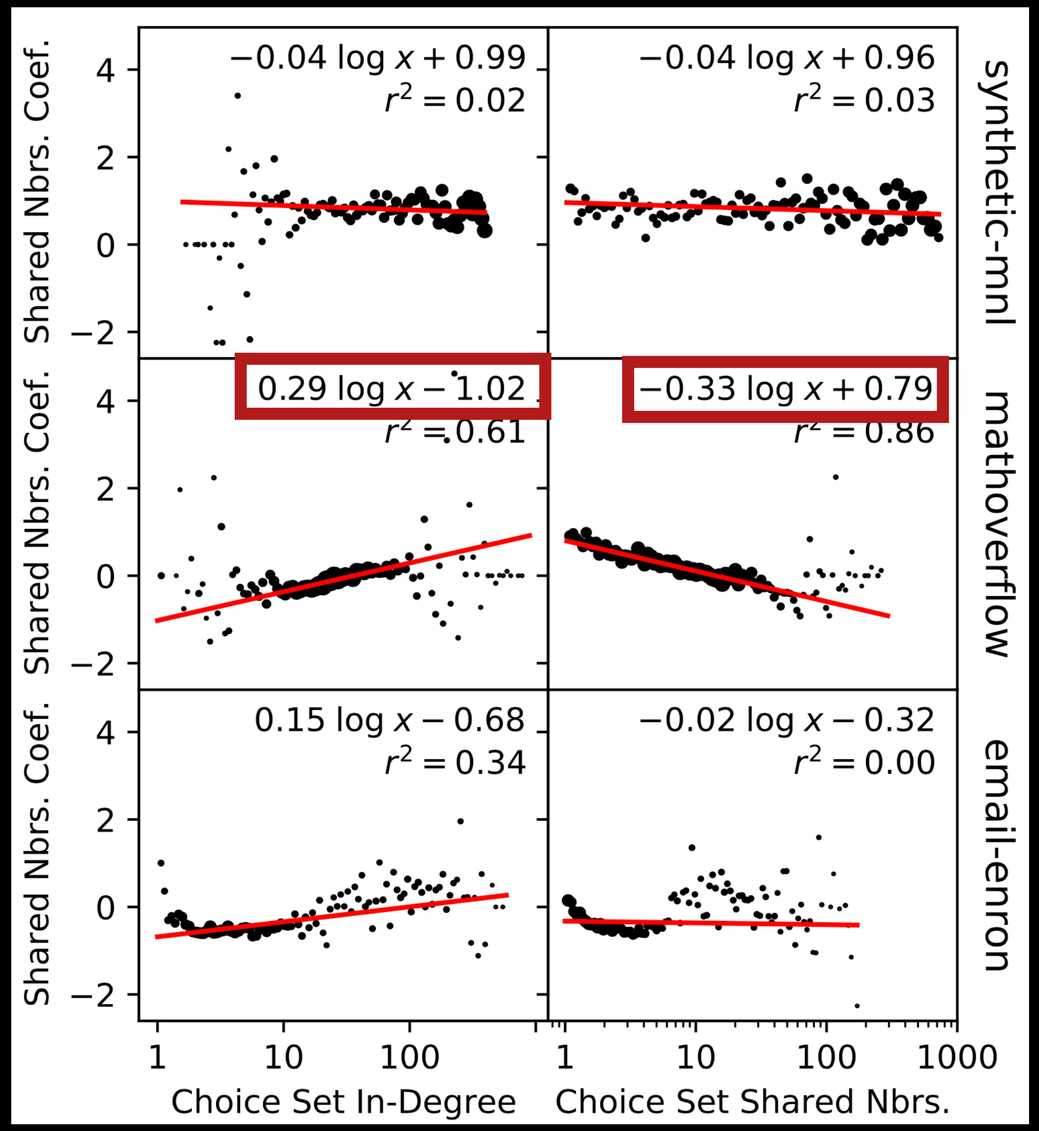


3. Decomposed LCL

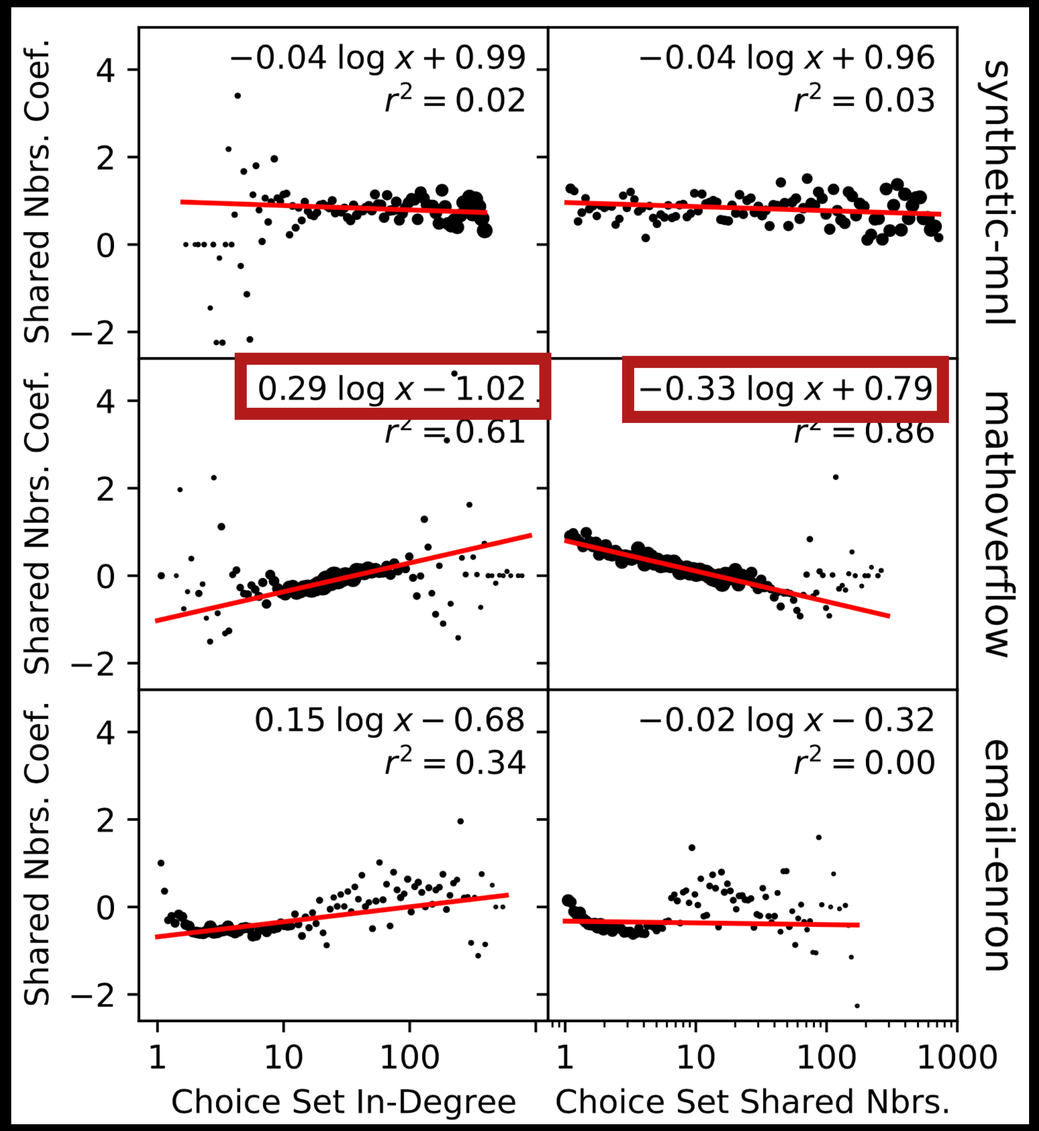
Motivation: the problem of intercepts



Motivation: the problem of intercepts

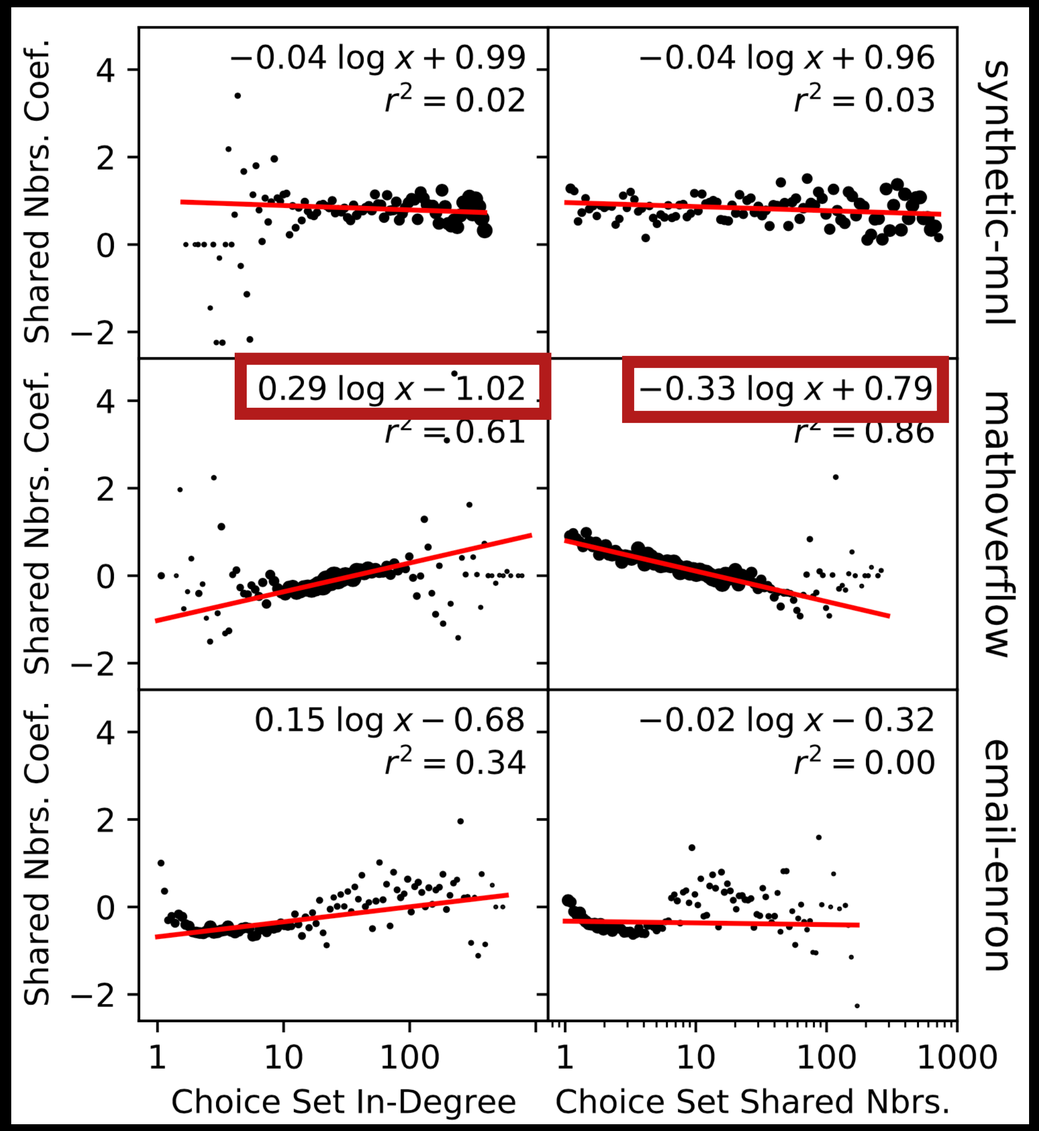


Motivation: the problem of intercepts



what happens with mean vector zero?

Motivation: the problem of intercepts



what happens with mean vector zero?

d intercepts d^2 slopes

$$\Pr(i \mid C) = \frac{\exp([\theta + Ax_C]^T x_i)}{\sum_{j \in C} \exp([\theta + Ax_C]^T x_j)}$$

One option: *Decomposed Linear Context Logit (DLCL)*

One option: *Decomposed Linear Context Logit (DLCL)*

sub-model for the context effect exerted by each feature

One option: *Decomposed Linear Context Logit (DLCL)*

sub-model for the context effect exerted by each feature

→ $u_{i,C,k}$: utility of i in C in the context of feature k

$$u_{i,C,k} = [B_k + F_k(C)]^T x_i$$

(matrix subscript = column)

One option: *Decomposed Linear Context Logit (DLCL)*

sub-model for the context effect exerted by each feature

→ $u_{i,C,k}$: utility of i in C in the context of feature k

$$u_{i,C,k} = [B_k + F_k(C)]^T x_i$$

(matrix subscript = column)

→ B: matrix of intercepts

One option: *Decomposed Linear Context Logit (DLCL)*

sub-model for the context effect exerted by each feature

→ $u_{i,C,k}$: utility of i in C in the context of feature k

$$u_{i,C,k} = [B_k + F_k(C)]^T x_i$$

(matrix subscript = column)

→ B: matrix of intercepts

→ LCL assumptions yields $F_k(C) = A_k(x_C)_k$

One option: *Decomposed Linear Context Logit (DLCL)*

sub-model for the context effect exerted by each feature

→ $u_{i,C,k}$: utility of i in C in the context of feature k

$$u_{i,C,k} = [B_k + F_k(C)]^T x_i$$

(matrix subscript = column)

→ B: matrix of intercepts

→ LCL assumptions yields $F_k(C) = A_k(x_C)_k$

→ combine sub-models in mixture (w/ proportions π)

One option: *Decomposed Linear Context Logit (DLCL)*

sub-model for the context effect exerted by each feature

→ $u_{i,C,k}$: utility of i in C in the context of feature k

$$u_{i,C,k} = [B_k + F_k(C)]^T x_i$$

(matrix subscript = column)

→ B: matrix of intercepts

→ LCL assumptions yields $F_k(C) = A_k(x_C)_k$

→ combine sub-models in mixture (w/ proportions π)

$$\Pr(i \mid C) = \sum_{k=1}^{d} \pi_k \frac{\exp([B_k + A_k(x_C)_k]^T x_i)}{\sum_{j \in C} \exp([B_k + A_k(x_C)_k]^T x_j)}$$

4. Identifiability and estimation

LCL identifiability, fully characterized

LCL identifiability, fully characterized

model is *identifiable* from dataset $\mathscr{D}$ if no two parameter values result in the same probability distribution

LCL identifiability, fully characterized

model is *identifiable* from dataset $\mathscr{D}$ if no two parameter values result in the same probability distribution

→ important for inference and interpretation

LCL identifiability, fully characterized

model is *identifiable* from dataset $\mathscr{D}$ if no two parameter values result in the same probability distribution

→ important for inference and interpretation

Theorem 1. A d-feature linear context logit is identifiable from a dataset $\mathcal{D}$ if and only if

$$\operatorname{span}\left\{\begin{bmatrix} x_C \\ 1 \end{bmatrix} \otimes (x_i - x_C) \mid C \in \mathcal{C}_{\mathcal{D}}, i \in C\right\} = \mathbb{R}^{d^2+d}. \tag{6}$$

($\mathscr{C}_{\mathscr{D}}$: unique choice sets in $\mathscr{D}$, $\otimes$: Kronecker product)

LCL identifiability, fully characterized

model is *identifiable* from dataset $\mathscr{D}$ if no two parameter values result in the same probability distribution

→ important for inference and interpretation

Theorem 1. A d-feature linear context logit is identifiable from a dataset $\mathcal{D}$ if and only if

$$\operatorname{span}\left\{\begin{bmatrix} x_C \\ 1 \end{bmatrix} \otimes (x_i - x_C) \mid C \in \mathcal{C}_{\mathcal{D}}, i \in C\right\} = \mathbb{R}^{d^2+d}. \qquad (6)$$

($\mathscr{C}_{\mathscr{D}}$: unique choice sets in $\mathscr{D}$, $\otimes$: Kronecker product)

intuition: need varied choice sets containing varied items

LCL identifiability, more intuition

LCL identifiability, more intuition

necessary:

PROPOSITION 2. *No d-feature linear context logit is identifiable from a dataset $\mathcal{D}$ if it does not include a set of $d + 1$ choice sets with affinely independent mean feature vectors.*

LCL identifiability, more intuition

necessary:

Proposition 2. *No d-feature linear context logit is identifiable from a dataset $\mathcal{D}$ if it does not include a set of $d+1$ choice sets with affinely independent mean feature vectors.*

sufficient:

Proposition 3. *If a dataset contains $d+1$ distinct choice sets $C_0, \ldots, C_d$ such that*

i. *the set of mean feature vectors $\{x_{C_0}, \ldots, x_{C_d}\}$ is affinely independent (the necessary condition from Proposition 2) and*
ii. *in each choice set C_i, there is some set of $d+1$ items with affinely independent features,*

then we can uniquely identify a d-feature LCL.

DLCL identifiability…. for another day

Mixture models are rough

see:

Chierichetti, F., Kumar, R., & Tomkins, A. (2018). Learning a mixture of two multinomial logits. In *International Conference on Machine Learning* (pp. 961-969).

Zhao, Z., Piech, P., & Xia, L. (2016). Learning mixtures of Plackett-Luce models. In *International Conference on Machine Learning* (pp. 2906-2914).

Grün, B., & Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, *25*(2), 225-247.

LCL and DLCL Estimation

LCL and DLCL Estimation

LCL

negative log-likelihood convex, so use gradient descent

LCL and DLCL Estimation

LCL

negative log-likelihood convex, so use gradient descent

$$-\ell(\theta, A; \mathscr{D}) = \sum_{(i,C)\in\mathscr{D}} -\left(\theta + Ax_C\right)^T x_i + \log \sum_{j\in C} \exp\left(\left[\theta + Ax_C\right]^T x_j\right)$$

LCL and DLCL Estimation

LCL

negative log-likelihood convex, so use gradient descent

$$-\ell(\theta, A; \mathcal{D}) = \sum_{(i,C)\in\mathcal{D}} -\left(\theta + Ax_C\right)^T x_i + \log \sum_{j\in C} \exp\left(\left[\theta + Ax_C\right]^T x_j\right)$$

DLCL

log-likelihood not convex…
but, *expectation-maximation (EM)* algorithm performs well

Table 9: Adam vs. EM algorithm for DLCL

Dataset	Adam NLL	EM NLL
DISTRICT	3206	**3041**
DISTRICT-SMART	3303	**3144**
EXPEDIA	837569	**805055**
SUSHI	9764	**9709**
CAR-A	1692	**1684**
CAR-B	1284	**1246**
CAR-ALT	7011	**6369**

5. Results on choice data

Choice datasets

Dataset	Choices	Features	Largest Choice Set
DISTRICT	5376	27	2
DISTRICT-SMART	5376	6	2
SUSHI	5000	6	10
EXPEDIA	276593	5	38
CAR-A	2675	4	2
CAR-B	2206	5	2
CAR-ALT	4654	21	6

Choice datasets

favorite sushi types
(random choice sets in survey)

Dataset	Choices	Features	Largest Choice Set
DISTRICT	5376	27	2
DISTRICT-SMART	5376	6	2
SUSHI	5000	6	10
EXPEDIA	276593	5	38
CAR-A	2675	4	2
CAR-B	2206	5	2
CAR-ALT	4654	21	6

Choice datasets

favorite sushi types
(random choice sets in survey)

hotel bookings
(choice sets = search results)

Dataset	Choices	Features	Largest Choice Set
DISTRICT	5376	27	2
DISTRICT-SMART	5376	6	2
SUSHI	5000	6	10
EXPEDIA	276593	5	38
CAR-A	2675	4	2
CAR-B	2206	5	2
CAR-ALT	4654	21	6

Choice datasets

favorite sushi types
(random choice sets in survey)

hotel bookings
(choice sets = search results)

car purchasing survey
(handcrafted choice sets)

Dataset	Choices	Features	Largest Choice Set
DISTRICT	5376	27	2
DISTRICT-SMART	5376	6	2
SUSHI	5000	6	10
EXPEDIA	276593	5	38
CAR-A	2675	4	2
CAR-B	2206	5	2
CAR-ALT	4654	21	6

LCL improves model fit

whole-dataset negative log-likelihood (lower = better)

	MNL	LCL	Mixed logit	DLCL
DISTRICT	3313	**3130**	3258	3206
DISTRICT-SMART	3426	**3278***	3351	3303†
EXPEDIA	839505	837649*	839055	**837569**†
SUSHI	9821	9773*	9793	**9764**
CAR-A	1702	1694	1696	**1692**
CAR-B	1305	1295	1297	**1284**
CAR-ALT	7393	**6733***	7301	7011†

*significant likelihood ratio test vs MNL ($p < 0.001$)

†significant likelihood ratio test vs mixed logit ($p < 0.001$)

LCL can improve out-of-sample prediction performance

mean (std. dev) relative rank (0 = perfect prediction, 1 = always wrong)

	MNL	LCL	Mixed logit	DLCL
DISTRICT	.3680 (.4823)	.3253 (.4685)	**.3188** (.4660)	.3225 (.4674)
DISTRICT-SMART	.4006 (.4900)	.3764 (.4845)	**.3271** (.4692)	.3448 (.4753)
EXPEDIA	.3859 (.2954)	.3800* (.2945)	.3201 (.2825)	**.3195**† (.2823)
SUSHI	.2727 (.2751)	.2737 (.2781)	**.2724** (.2765)	.2732 (.2765)
CAR-A	**.3570** (.4791)	**.3570** (.4791)	**.3570** (.4791)	**.3570** (.4791)
CAR-B	.3326 (.4711)	**.3213** (.4670)	.3303 (.4703)	.3235 (.4678)
CAR-ALT	.2944 (.2875)	**.2661*** (.2819)	.2931 (.2966)	.2798 (.2837)

*significant Wilcoxon test vs MNL ($p < 0.001$)

LCL can test individual effects for significance

LCL can test individual effects for significance

constrain all but one entry of A to 0

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

$\overline{A}_{pq}$: learned param in constrained model

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

$\overline{A}_{pq}$: learned param in constrained model

Table 5: Five largest context effects in SUSHI.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	**LRT**	p-value
popularity on *popularity*	−0.28	−0.11	3.0	0.081
availability on *is maki*	0.24	0.04	0.39	0.53
oiliness on *oiliness*	−0.20	−0.26	23	1.5×10^{-6}
popularity on *availability*	0.19	0.09	2.3	0.13
availability on *oiliness*	−0.18	−0.04	0.74	0.39

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

$\overline{A}_{pq}$: learned param in constrained model

Table 5: Five largest context effects in SUSHI.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
popularity on *popularity*	−0.28	−0.11	3.0	0.081
availability on *is maki*	0.24	0.04	0.39	0.53
oiliness on *oiliness*	−0.20	−0.26	23	1.5×10^{-6}
popularity on *availability*	0.19	0.09	2.3	0.13
availability on *oiliness*	−0.18	−0.04	0.74	0.39

Table 6: Five largest context effects in EXPEDIA.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
location score on *price*	−0.47	−0.13	10	0.002
on promotion on *price*	0.27	0.13	17	3.5×10^{-5}
review score on *price*	−0.19	−0.13	29	8.6×10^{-8}
star rating on *price*	0.15	0.20	65	6.0×10^{-16}
price on *star rating*	0.10	0.03	4.0	0.046

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

$\overline{A}_{pq}$: learned param in constrained model

Table 5: Five largest context effects in SUSHI.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
popularity on *popularity*	−0.28	−0.11	3.0	0.081
availability on *is maki*	0.24	0.04	0.39	0.53
oiliness on *oiliness*	−0.20	−0.26	23	1.5×10^{-6}
popularity on *availability*	0.19	0.09	2.3	0.13
availability on *oiliness*	−0.18	−0.04	0.74	0.39

Table 6: Five largest context effects in EXPEDIA.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
location score on *price*	−0.47	−0.13	10	0.002
on promotion on *price*	0.27	0.13	17	3.5×10^{-5}
review score on *price*	−0.19	−0.13	29	8.6×10^{-8}
star rating on *price*	0.15	0.20	65	6.0×10^{-16}
price on *star rating*	0.10	0.03	4.0	0.046

Table 7: Five largest context effects in CAR-ALT.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
truck on *truck*	1.06	0.83	239	$< 10^{-16}$
van on *van*	0.94	0.97	309	$< 10^{-16}$
suv on *station wagon*	0.89	0.98	−0.21	1.0
station wagon on *station wagon*	0.88	0.93	153	$< 10^{-16}$
sports car on *station wagon*	0.86	0.96	−0.21	1.0

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

$\overline{A}_{pq}$: learned param in constrained model

are these effects *causal*?

Table 5: Five largest context effects in SUSHI.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
popularity on *popularity*	−0.28	−0.11	3.0	0.081
availability on *is maki*	0.24	0.04	0.39	0.53
oiliness on *oiliness*	−0.20	−0.26	23	1.5×10^{-6}
popularity on *availability*	0.19	0.09	2.3	0.13
availability on *oiliness*	−0.18	−0.04	0.74	0.39

Table 6: Five largest context effects in EXPEDIA.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
location score on *price*	−0.47	−0.13	10	0.002
on promotion on *price*	0.27	0.13	17	3.5×10^{-5}
review score on *price*	−0.19	−0.13	29	8.6×10^{-8}
star rating on *price*	0.15	0.20	65	6.0×10^{-16}
price on *star rating*	0.10	0.03	4.0	0.046

Table 7: Five largest context effects in CAR-ALT.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
truck on *truck*	1.06	0.83	239	$< 10^{-16}$
van on *van*	0.94	0.97	309	$< 10^{-16}$
suv on *station wagon*	0.89	0.98	−0.21	1.0
station wagon on *station wagon*	0.88	0.93	153	$< 10^{-16}$
sports car on *station wagon*	0.86	0.96	−0.21	1.0

LCL can test individual effects for significance

constrain all but one entry of A to 0

→ loss still convex

→ test *strength* of single effect

→ constrained model nests cond. logit

$\overline{A}_{pq}$: learned param in constrained model

are these effects *causal*?

→ *choice set assignment* (stay tuned)

Table 5: Five largest context effects in SUSHI.

Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
popularity on *popularity*	−0.28	−0.11	3.0	0.081
availability on *is maki*	0.24	0.04	0.39	0.53
oiliness on *oiliness*	−0.20	−0.26	23	1.5×10^{-6}
popularity on *availability*	0.19	0.09	2.3	0.13
availability on *oiliness*	−0.18	−0.04	0.74	0.39

Table 6: Five largest context effects in EXPEDIA.

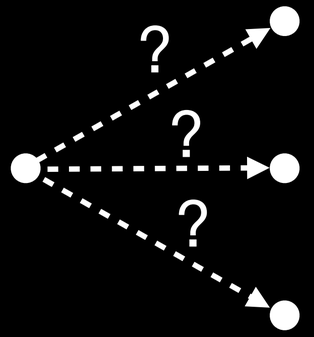
Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
location score on *price*	−0.47	−0.13	10	0.002
on promotion on *price*	0.27	0.13	17	3.5×10^{-5}
review score on *price*	−0.19	−0.13	29	8.6×10^{-8}
star rating on *price*	0.15	0.20	65	6.0×10^{-16}
price on *star rating*	0.10	0.03	4.0	0.046

Table 7: Five largest context effects in CAR-ALT.

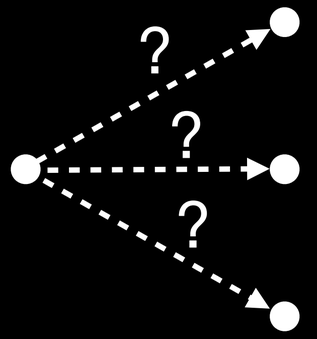
Effect (q on p)	A_{pq}	$\overline{A}_{pq}$	LRT	p-value
truck on *truck*	1.06	0.83	239	$< 10^{-16}$
van on *van*	0.94	0.97	309	$< 10^{-16}$
suv on *station wagon*	0.89	0.98	−0.21	1.0
station wagon on *station wagon*	0.88	0.93	153	$< 10^{-16}$
sports car on *station wagon*	0.86	0.96	−0.21	1.0

6. Social network application

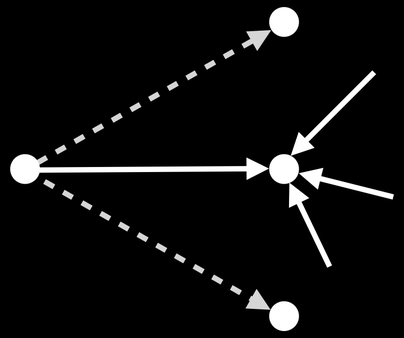
What factors drive edge formation?



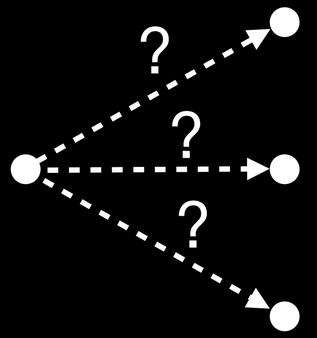
What factors drive edge formation?



Preferential attachment
(Barabási & Albert, *Science* 1999)

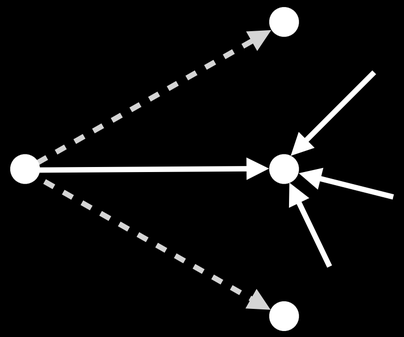


What factors drive edge formation?



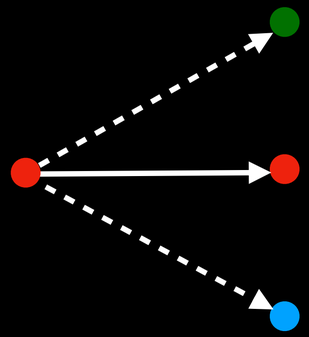
Preferential attachment

(Barabási & Albert, *Science* 1999)

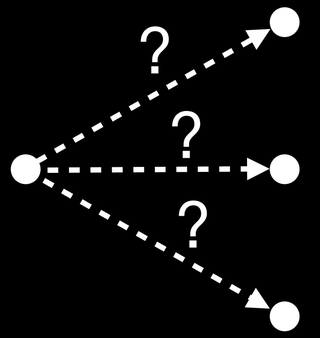


Homophily

(McPherson et al., *Annual Review of Sociology* 2001)
(Papadopoulos et al., *Nature* 2012)

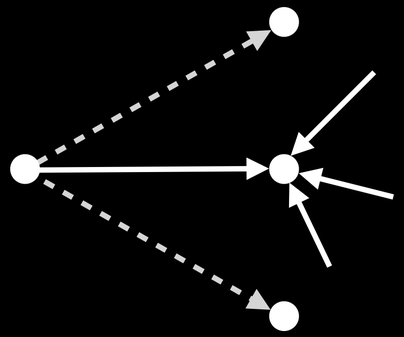


What factors drive edge formation?



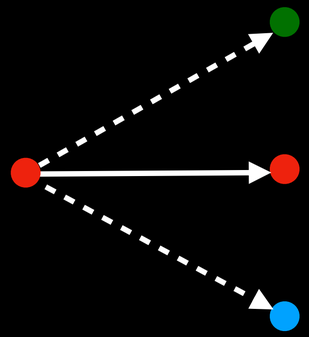
Preferential attachment

(Barabási & Albert, *Science* 1999)



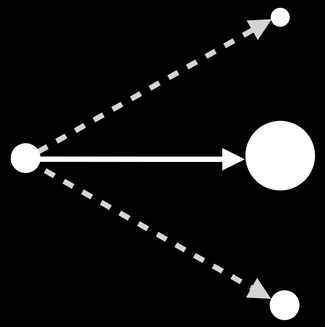
Homophily

(McPherson et al., *Annual Review of Sociology* 2001)
(Papadopoulos et al., *Nature* 2012)

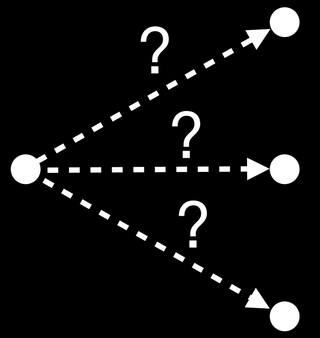


Fitness

(Bianconi & Barabási, *Europhysics Letters* 2001)
(Caldarelli et al., *Physical Review Letters* 2002)

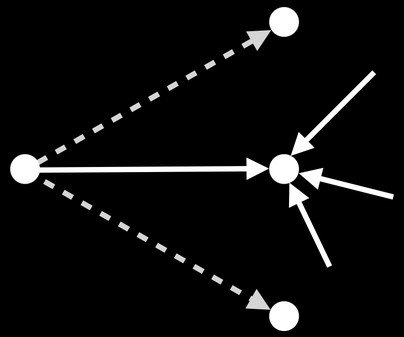


What factors drive edge formation?



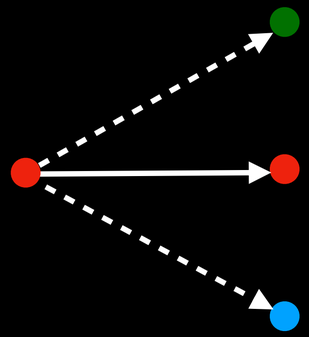
Preferential attachment

(Barabási & Albert, *Science* 1999)



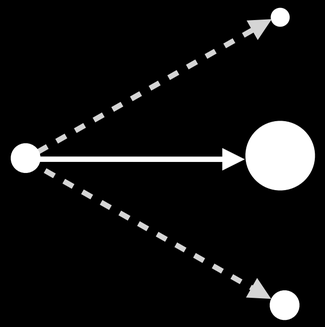
Homophily

(McPherson et al., *Annual Review of Sociology* 2001)
(Papadopoulos et al., *Nature* 2012)



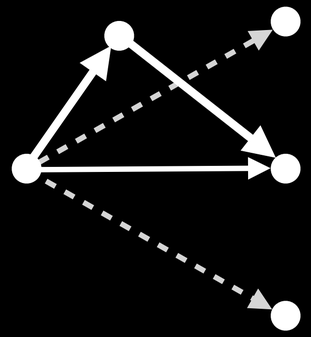
Fitness

(Bianconi & Barabási, *Europhysics Letters* 2001)
(Caldarelli et al., *Physical Review Letters* 2002)



Triadic closure

(Rapoport, *Bulletin of Mathematical Biophysics* 1953)
(Jin et al., *Physical Review E* 2001)



“Choosing to grow a graph”

(Overgoor et al., *SINM* ’19 & *WWW* ’19)

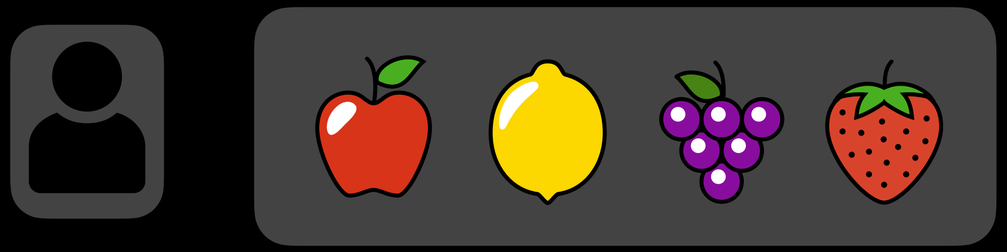
(Gupta & Porter, *arXiv* 2020)

“Choosing to grow a graph”

(Overgoor et al., *SINM* ’19 & *WWW* ’19)

(Gupta & Porter, *arXiv* 2020)

so far:



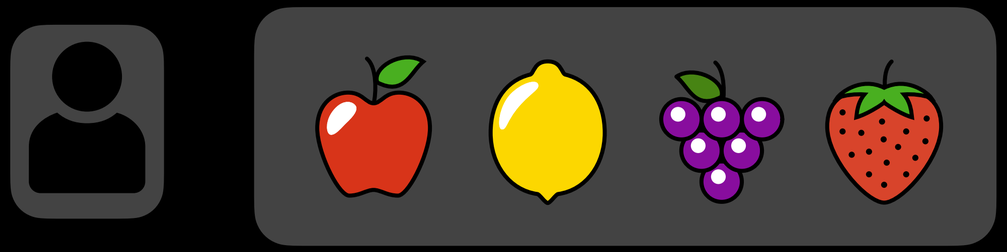
chooser *choice set*

“Choosing to grow a graph”

(Overgoor et al., *SINM* ’19 & *WWW* ’19)

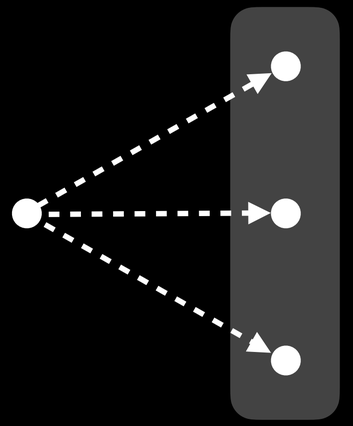
(Gupta & Porter, *arXiv* 2020)

so far:



chooser *choice set*

in network growth:

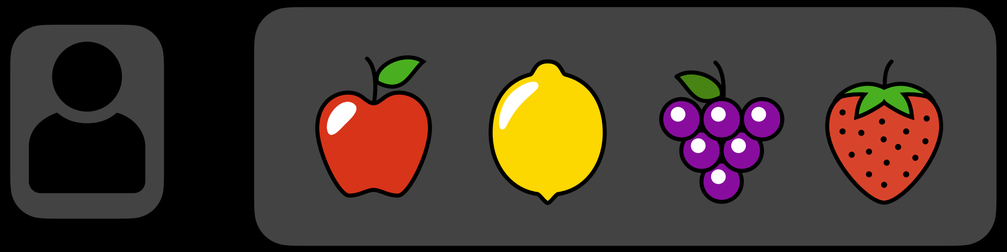


choice set

“Choosing to grow a graph”

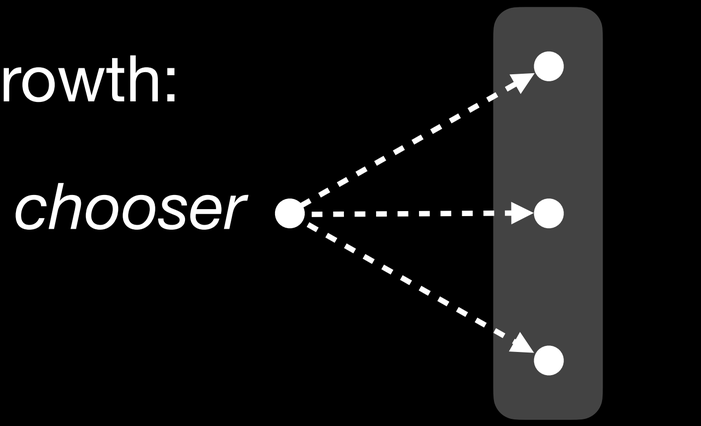
(Overgoor et al., *SINM* ’19 & *WWW* ’19)
(Gupta & Porter, *arXiv* 2020)

so far:



chooser *choice set*

in network growth:



choice set

Key usage

Timestamped edges
→ meaningful choice sets

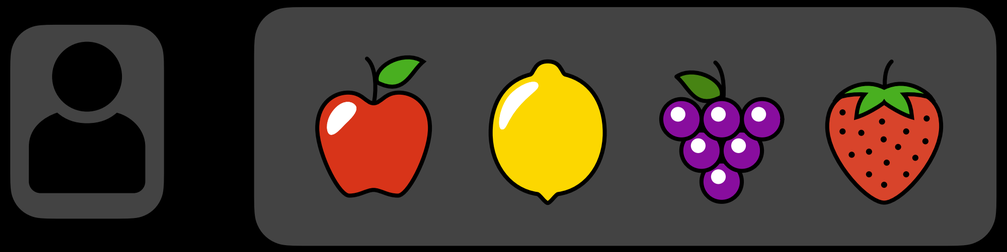
Infer relative importance of edge formation mechanisms from data

“Choosing to grow a graph”

(Overgoor et al., *SINM* ’19 & *WWW* ’19)

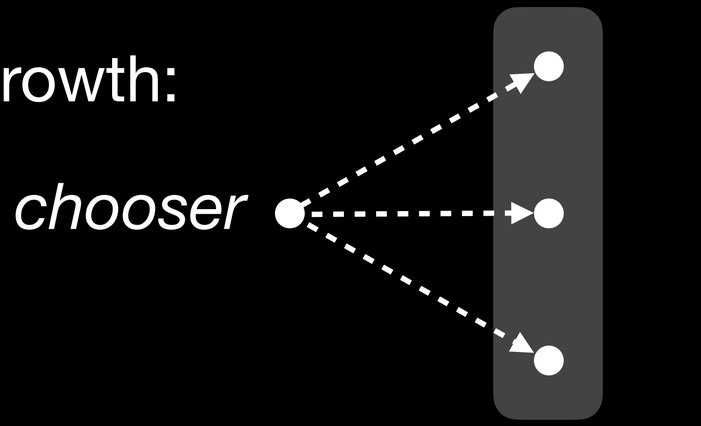
(Gupta & Porter, *arXiv* 2020)

so far:



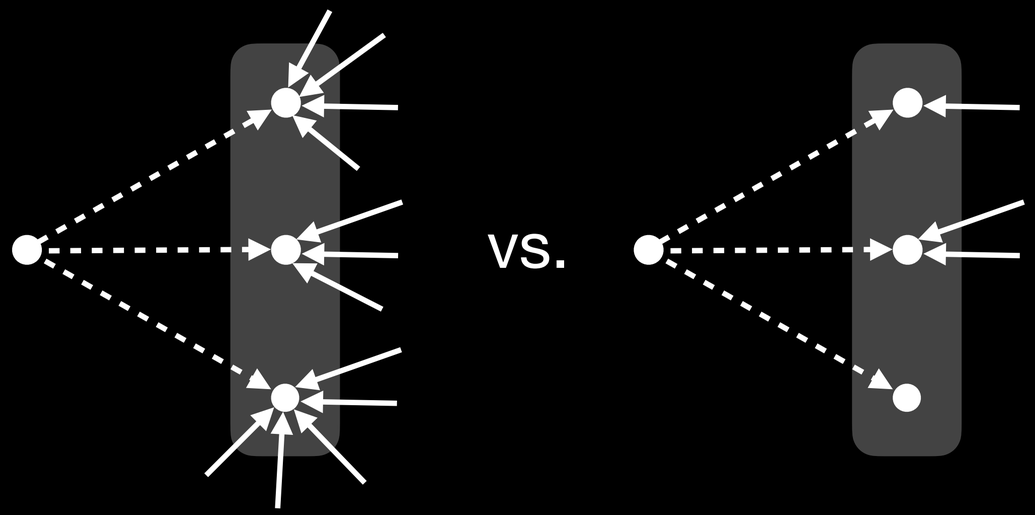
chooser *choice set*

in network growth:



choice set

feature context effects:



Key usage

Timestamped edges
→ meaningful choice sets

Infer relative importance of edge formation mechanisms from data

Choosing to close triangles

Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets

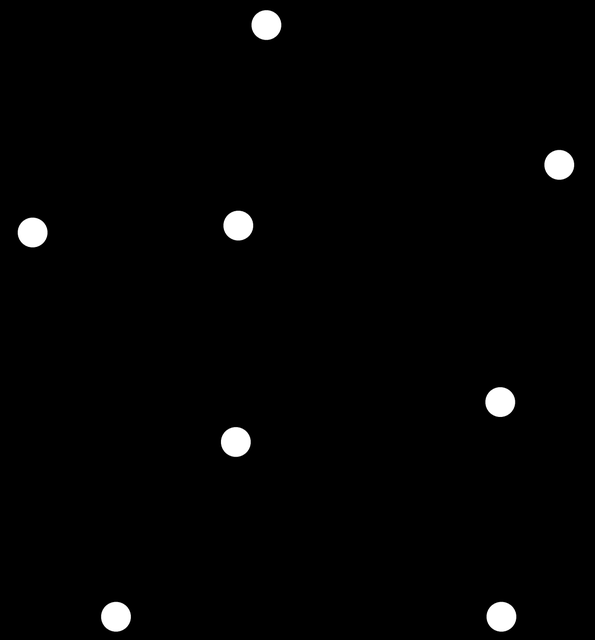
Choosing to close triangles

Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets

Our data
Timestamped edges
(including repeats)

Choosing to close triangles

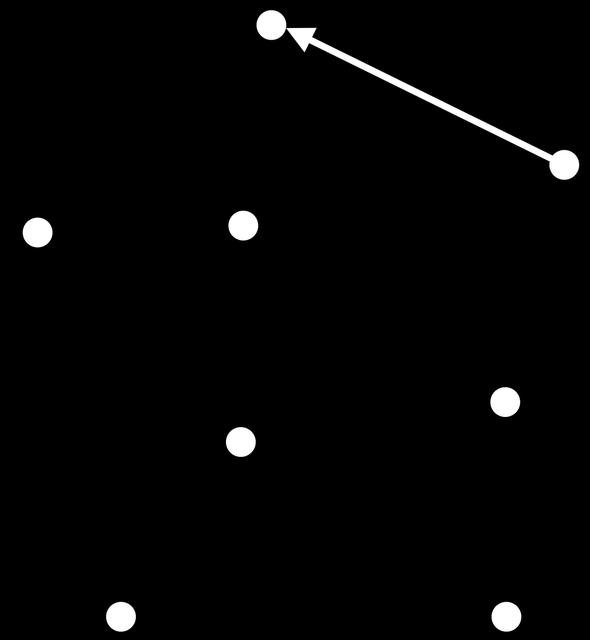
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

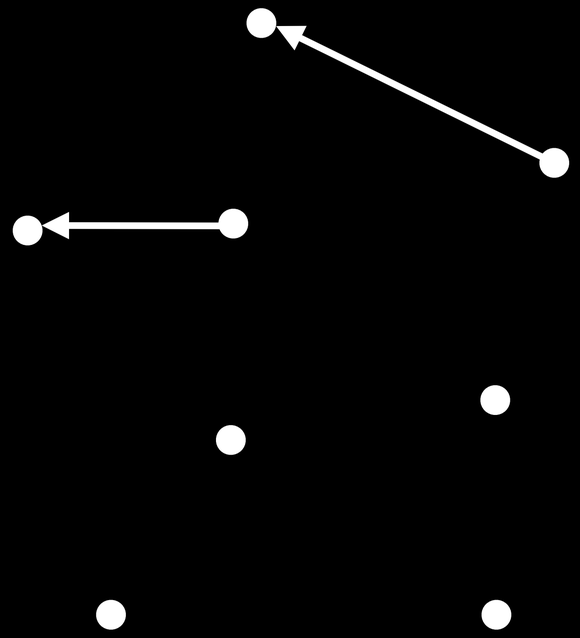
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

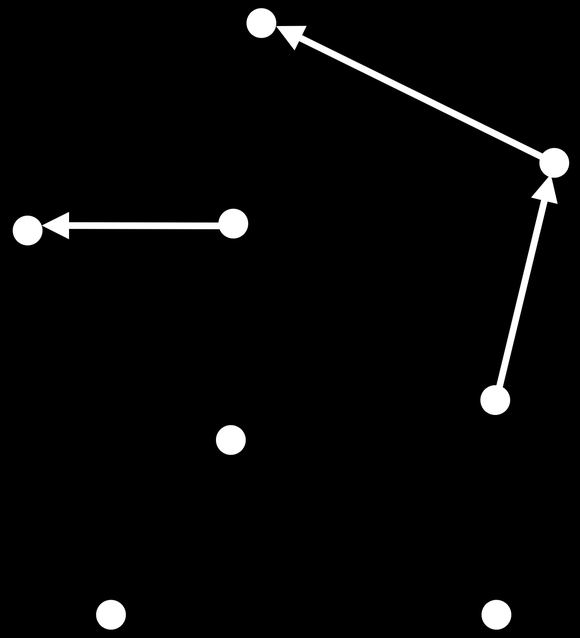
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

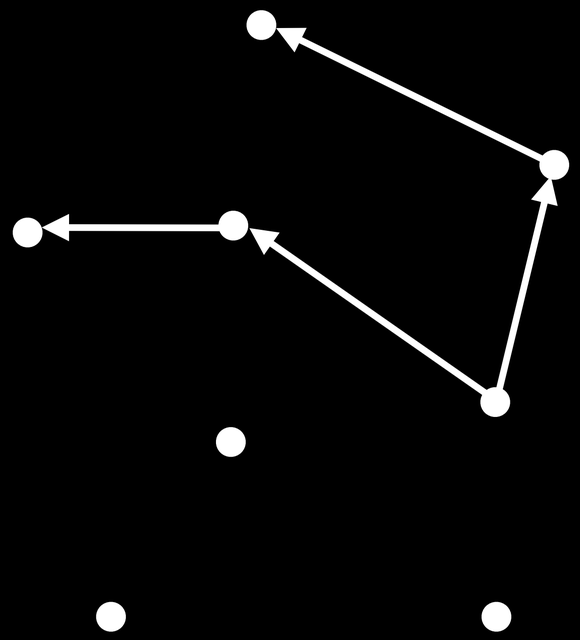
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

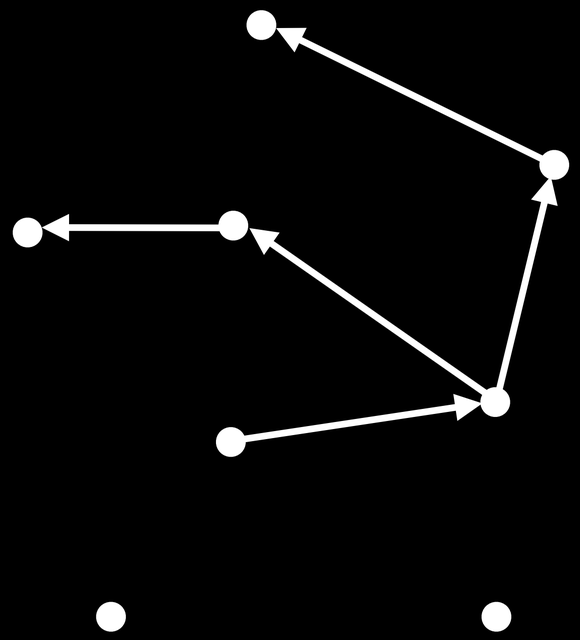
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

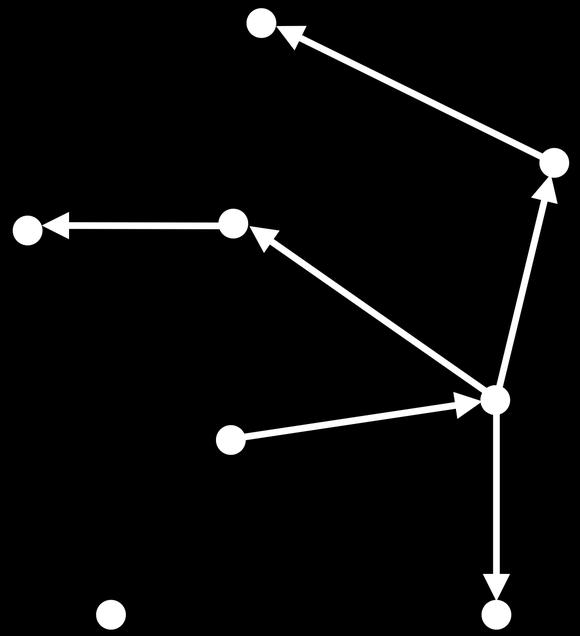
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

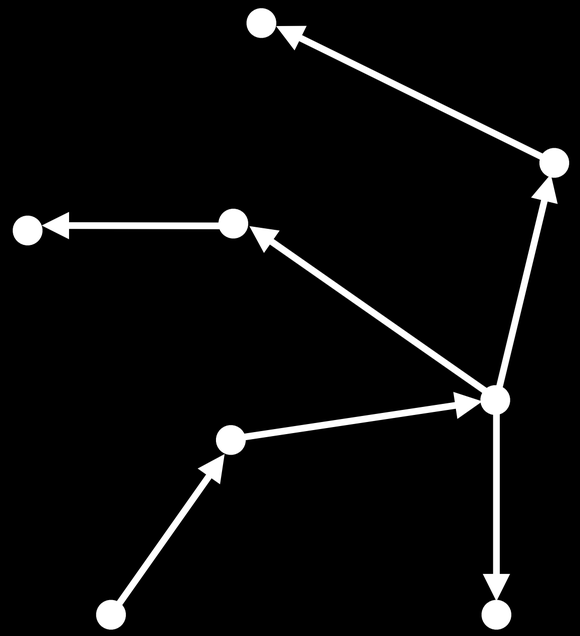
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

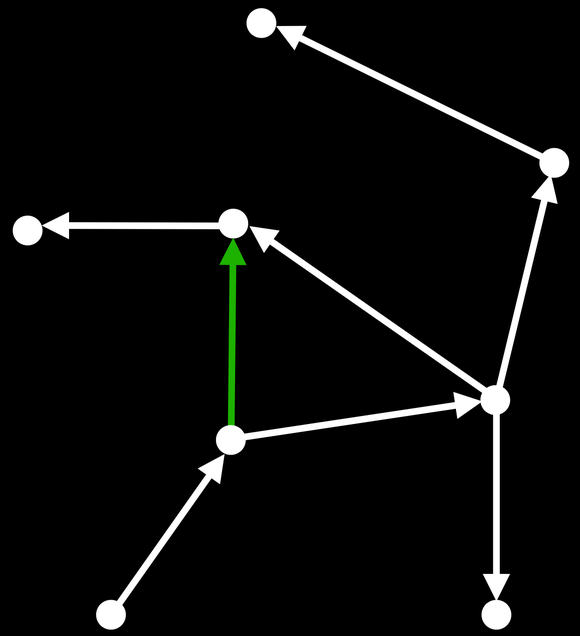
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

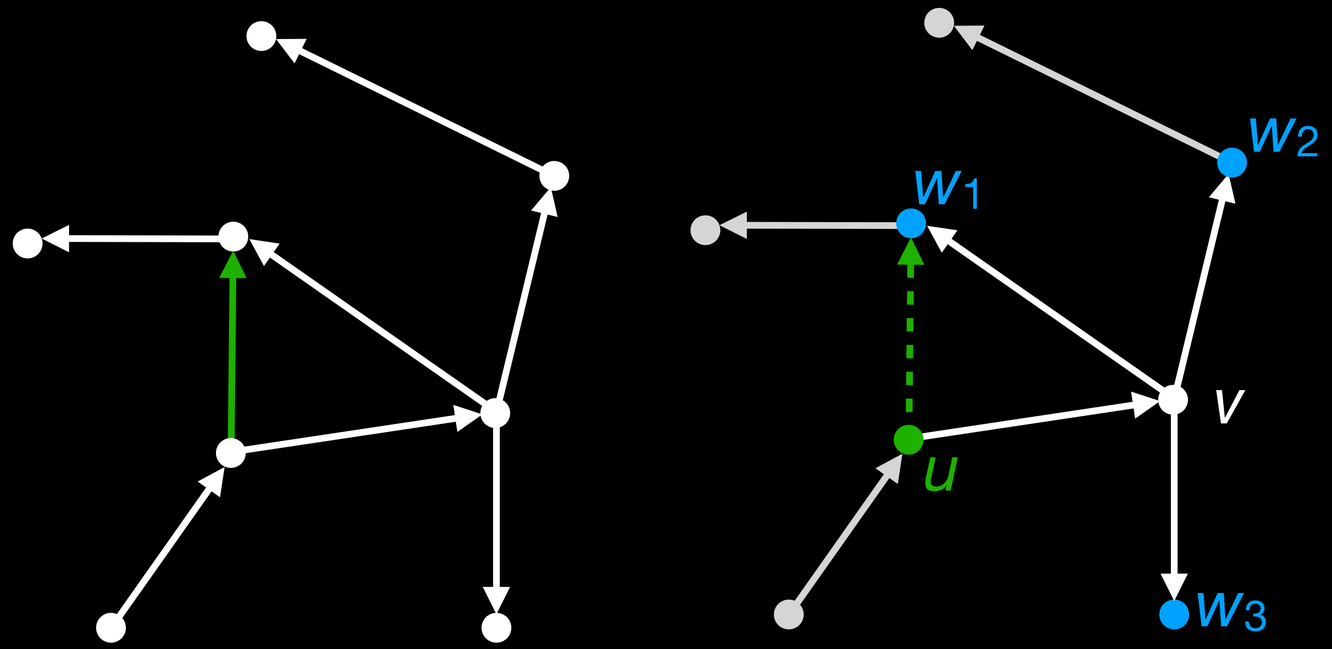
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

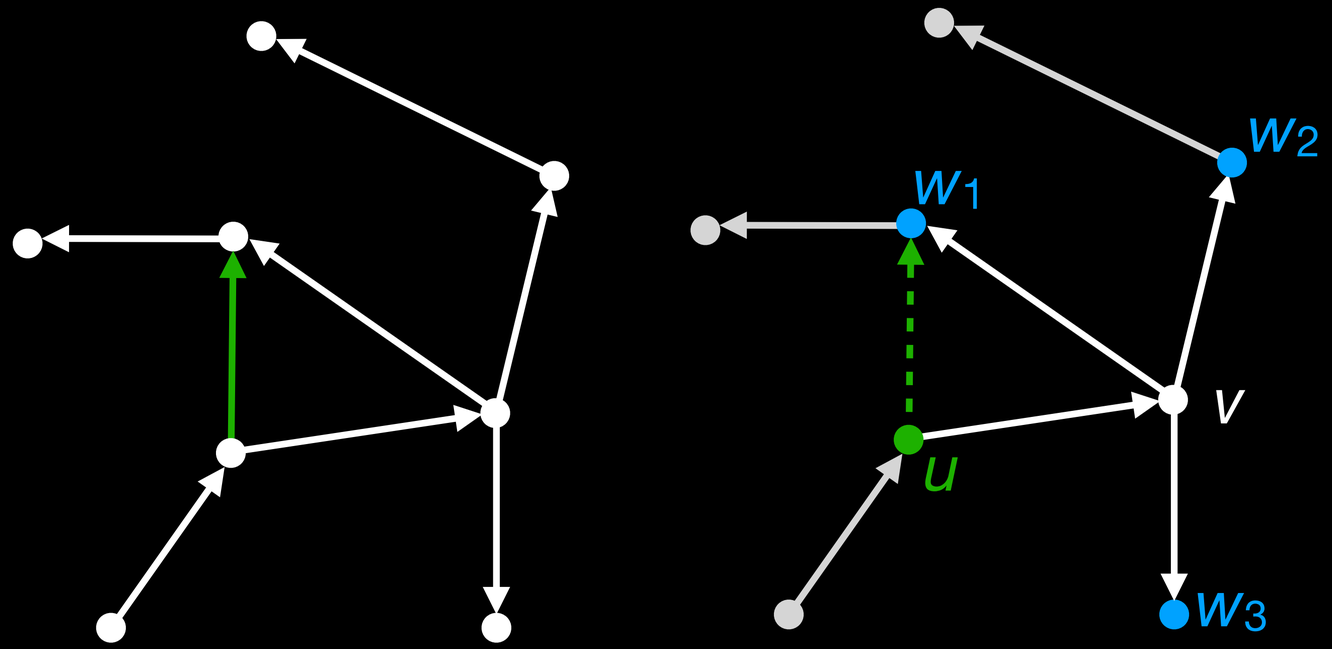
Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



Our data
Timestamped edges
(including repeats)

Choosing to close triangles

Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets

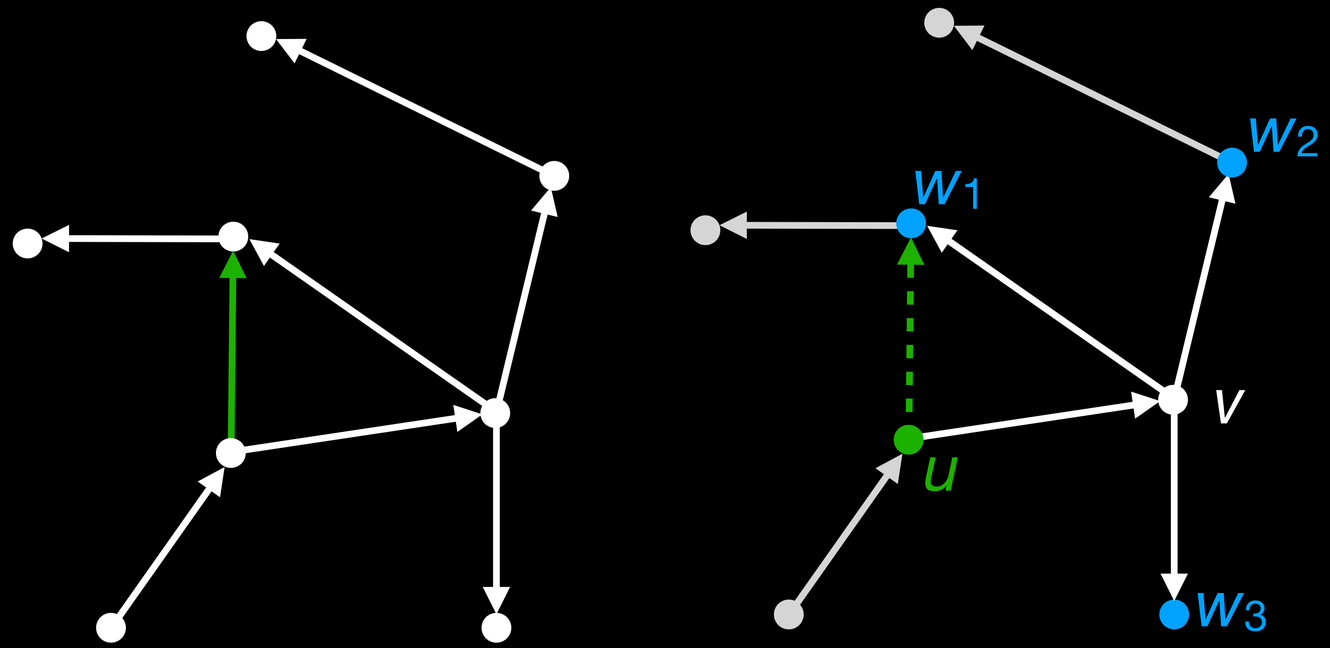


chooser
u

Our data
Timestamped edges
(including repeats)

Choosing to close triangles

Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



chooser u

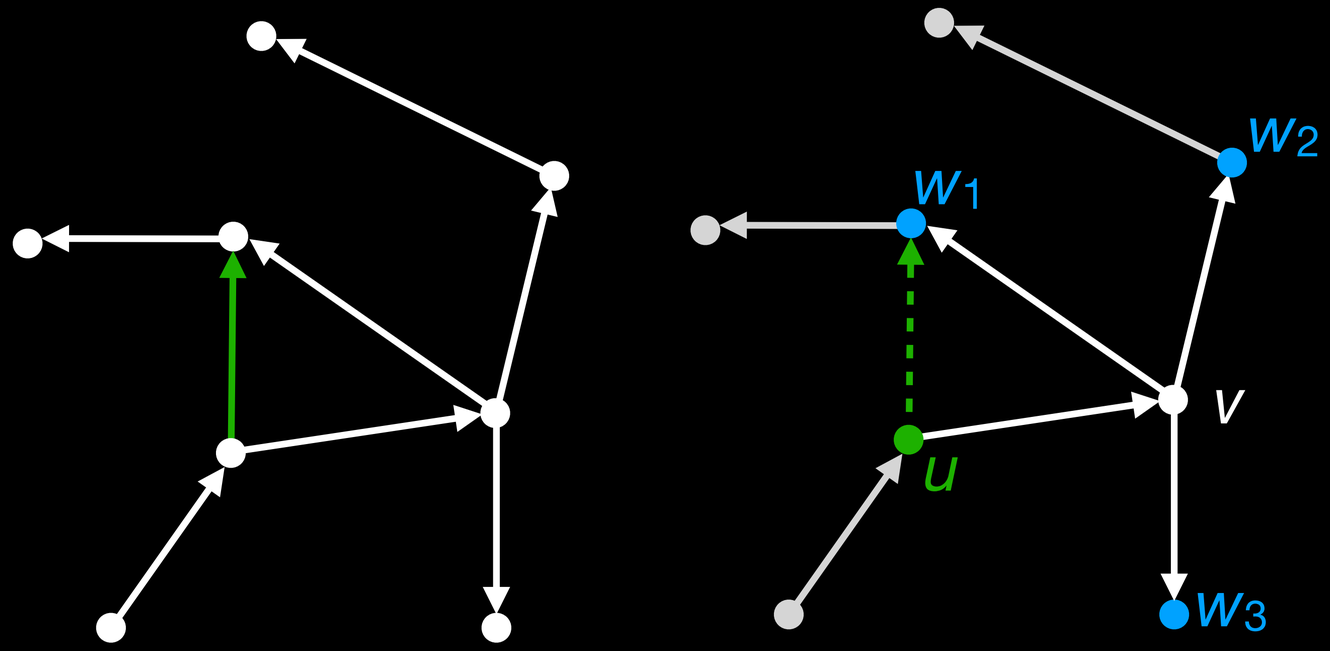
choice set $\{w_1, w_2, w_3\}$

Our data
Timestamped edges
(including repeats)

Choosing to close triangles

Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets

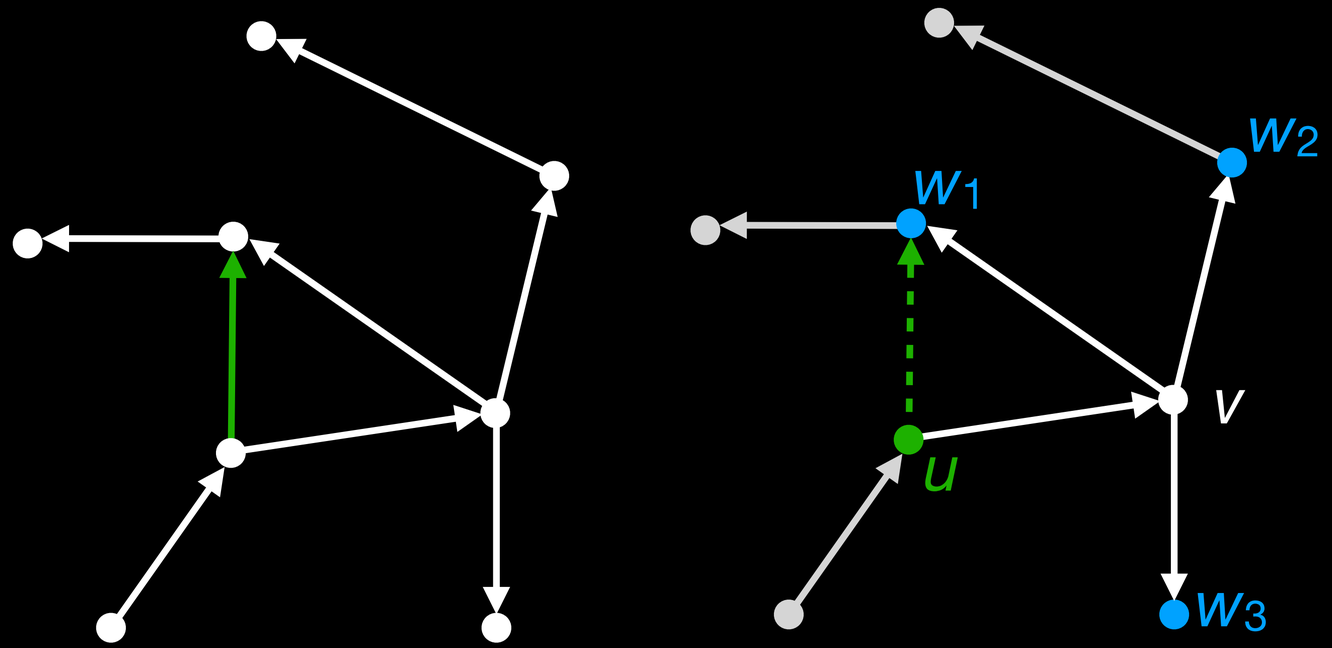
Our data
Timestamped edges
(including repeats)



chooser	*choice set*	*choice*
u	$\{w_1, w_2, w_3\}$	w_1

Choosing to close triangles

Triadic closure offers small choice sets
→ tractable inference
→ varied choice sets



chooser	*choice set*	*choice*
u	$\{w_1, w_2, w_3\}$	w_1

Our data
Timestamped edges
(including repeats)

Node features

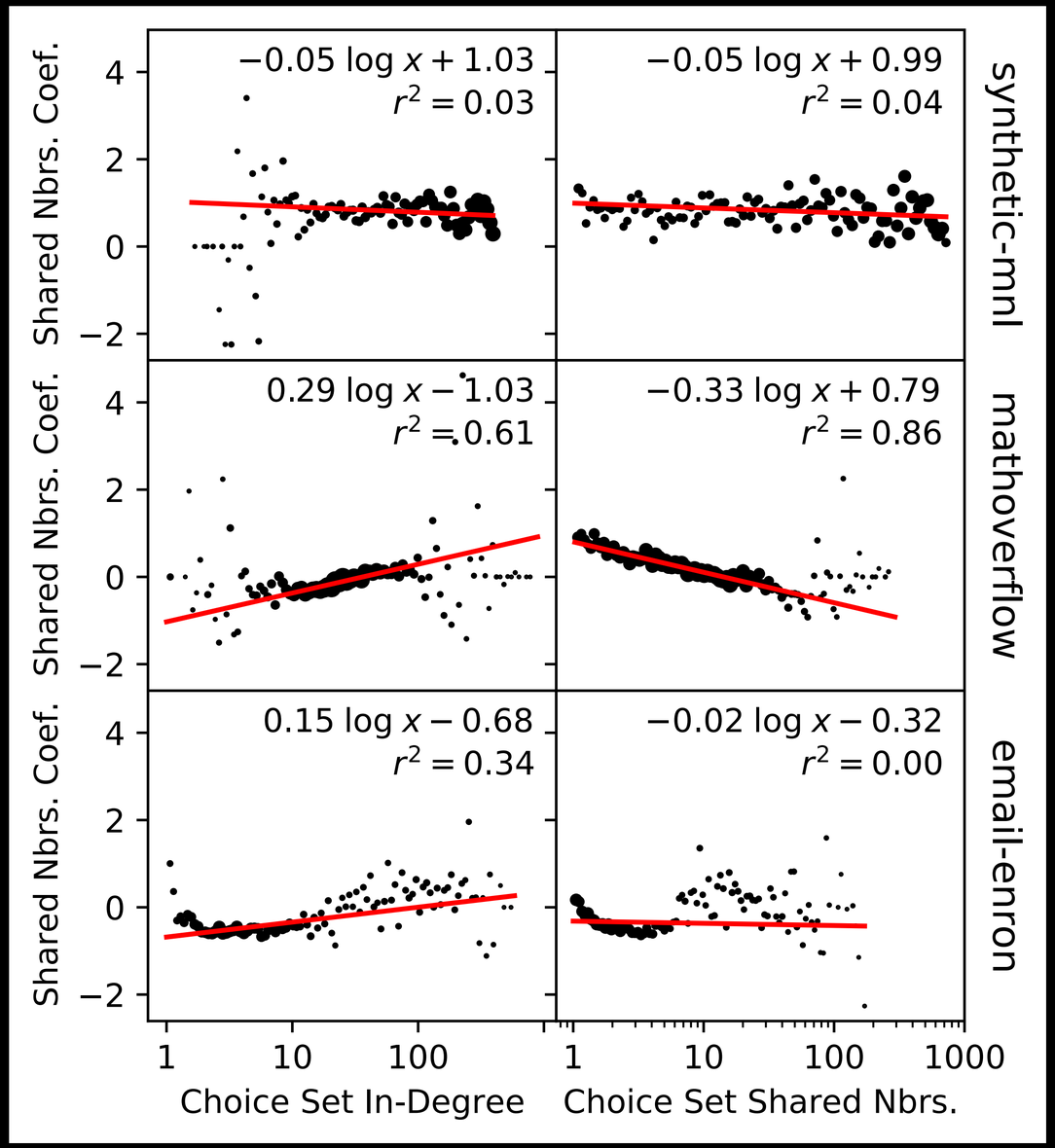
1. in-degree of w
2. # shared neighbors of u, w
3. weight of edge $w \rightarrow u$
4. time since last edge into w
5. time since last edge out of w
6. time since last $w \rightarrow u$ edge

Context matters in triadic closure

Context matters in triadic closure

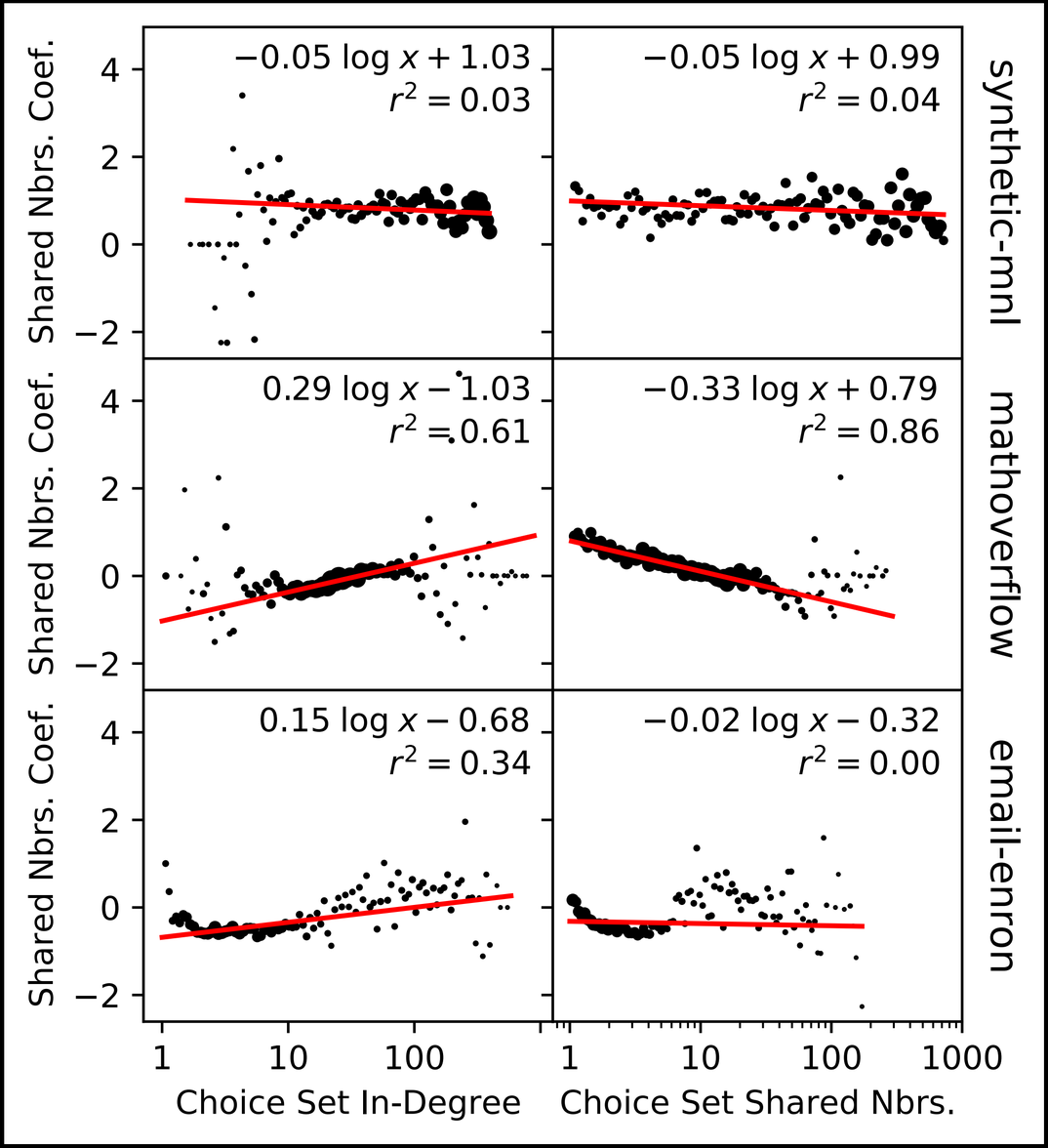
Datasets
email-enron
email-eu
email-w3c
wiki-talk
reddit-hyperlink
bitcoin-alpha
bitcoin-otc
mathoverflow
college-msg
facebook-wall
sms-a
sms-b
sms-c

Context matters in triadic closure



Datasets
email-enron
email-eu
email-w3c
wiki-talk
reddit-hyperlink
bitcoin-alpha
bitcoin-otc
mathoverflow
college-msg
facebook-wall
sms-a
sms-b
sms-c

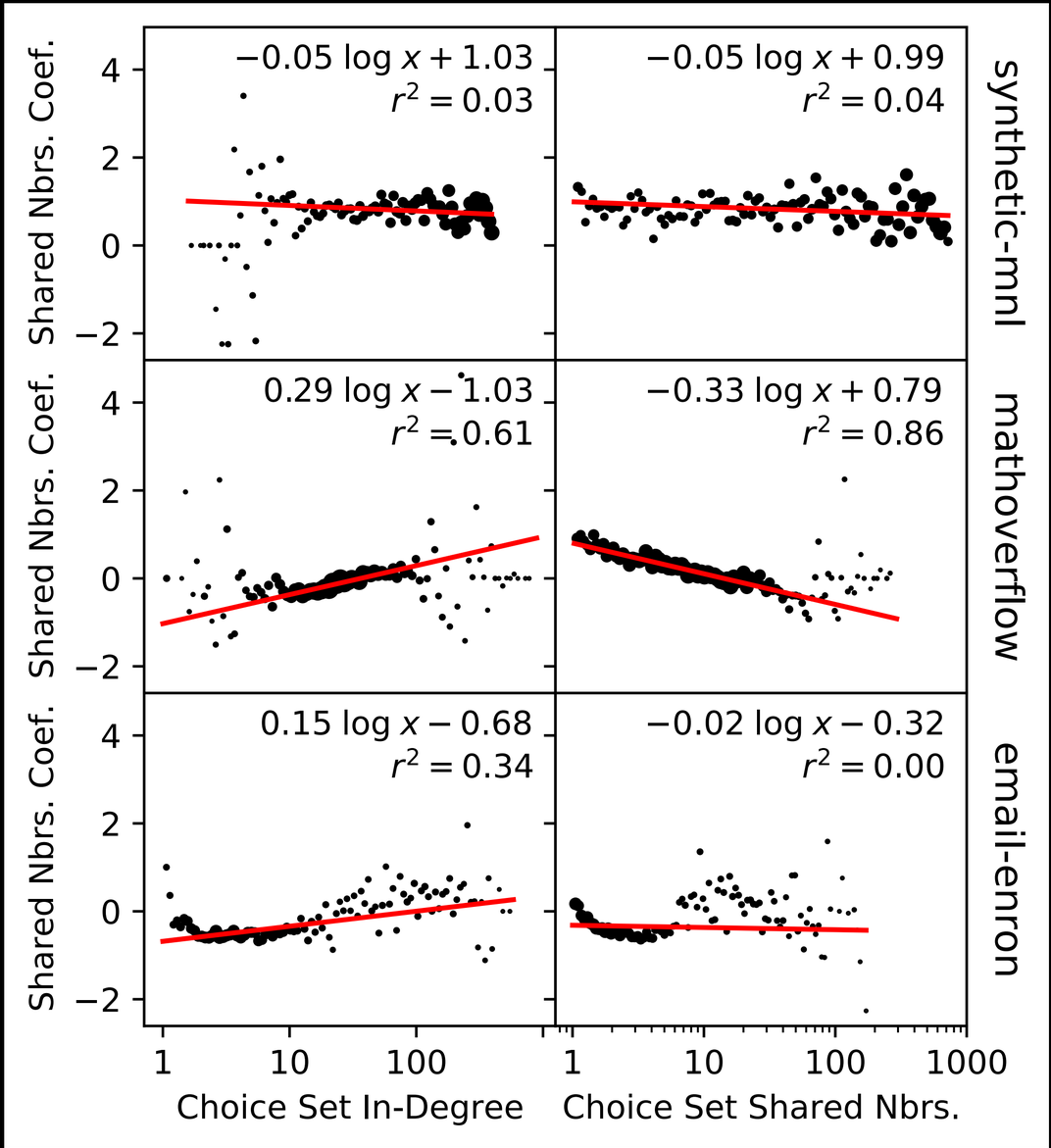
Context matters in triadic closure



Synthetic data,
no context effects

Datasets
email-enron
email-eu
email-w3c
wiki-talk
reddit-hyperlink
bitcoin-alpha
bitcoin-otc
mathoverflow
college-msg
facebook-wall
sms-a
sms-b
sms-c

Context matters in triadic closure

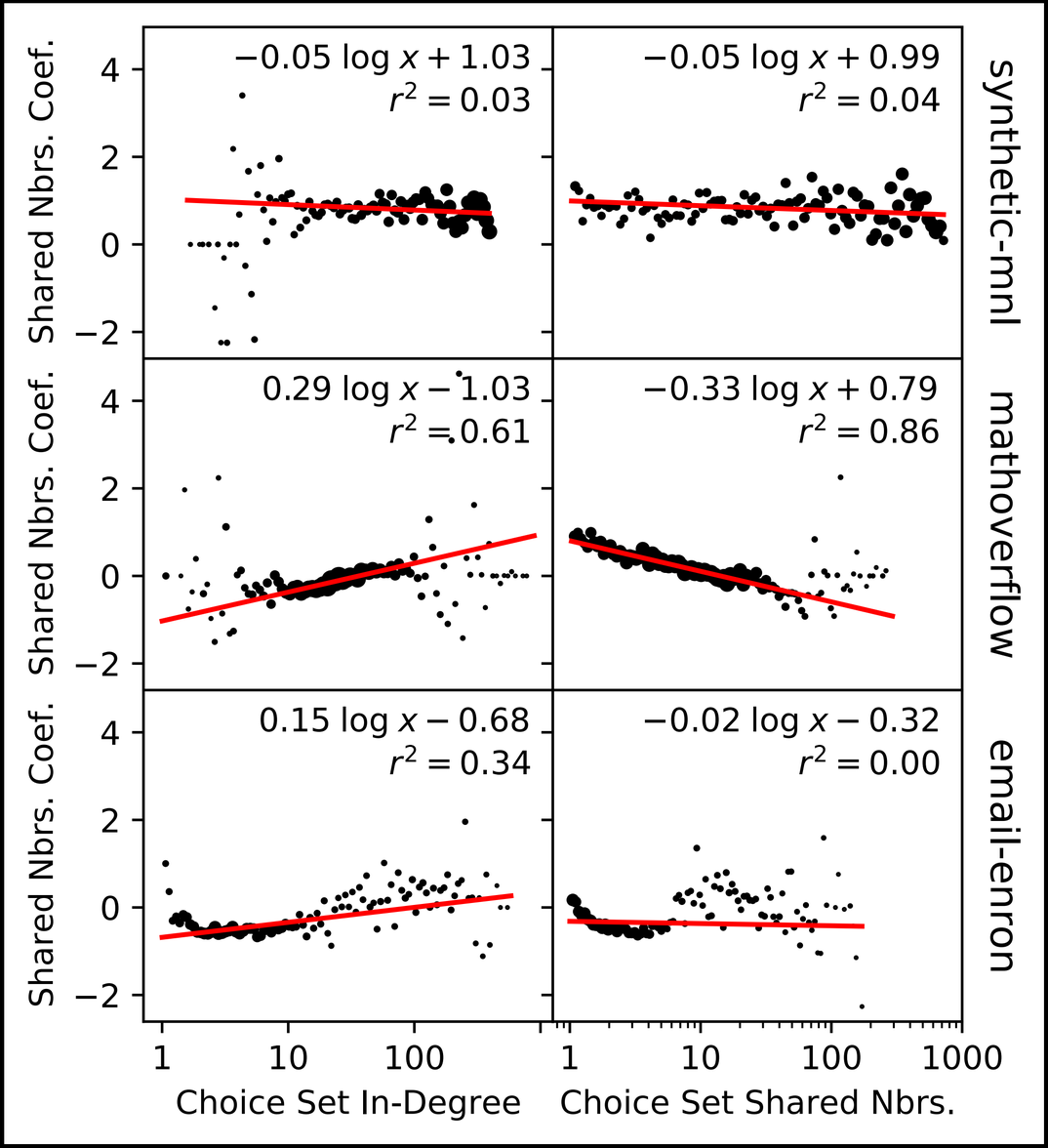


Synthetic data,
no context effects

Commenting network,
linear context effects

Datasets
email-enron
email-eu
email-w3c
wiki-talk
reddit-hyperlink
bitcoin-alpha
bitcoin-otc
mathoverflow
college-msg
facebook-wall
sms-a
sms-b
sms-c

Context matters in triadic closure



Synthetic data,
no context effects

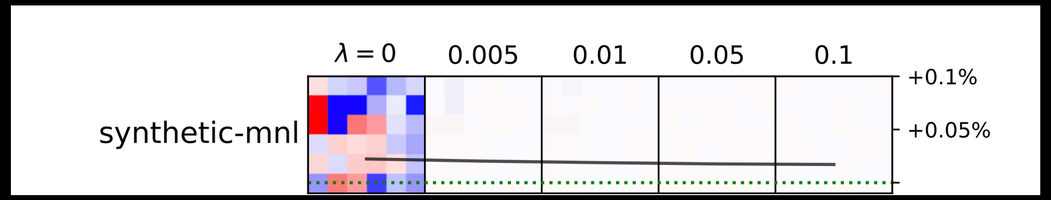
Commenting network,
linear context effects

Email network,
nonlinear context effects?

Datasets
email-enron
email-eu
email-w3c
wiki-talk
reddit-hyperlink
bitcoin-alpha
bitcoin-otc
mathoverflow
college-msg
facebook-wall
sms-a
sms-b
sms-c

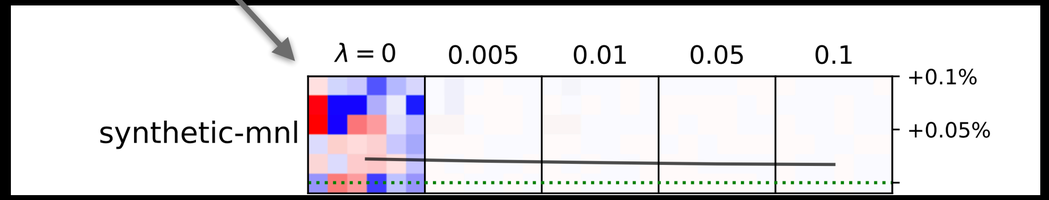
LCL reveals interpretable feature context effects

LCL reveals interpretable feature context effects



LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)



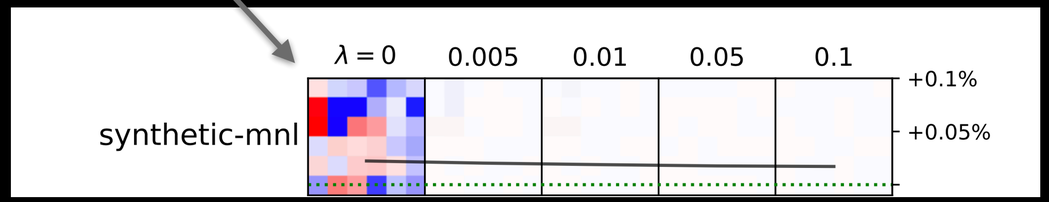
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



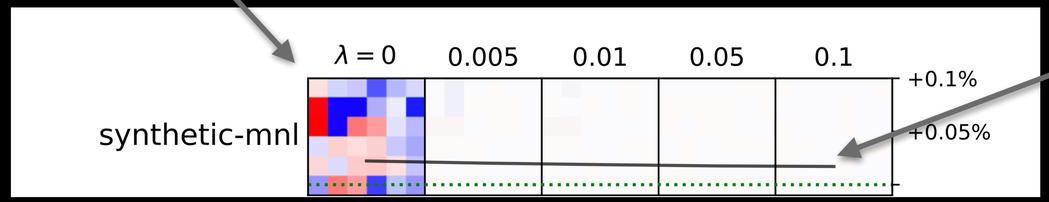
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



NLL (lower = better fit)

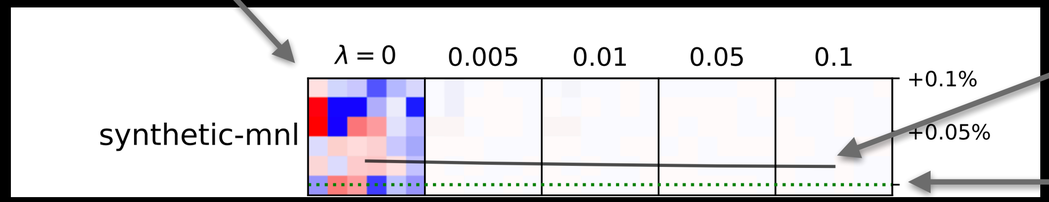
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



NLL (lower = better fit)

$p = 0.001$ LRT threshold

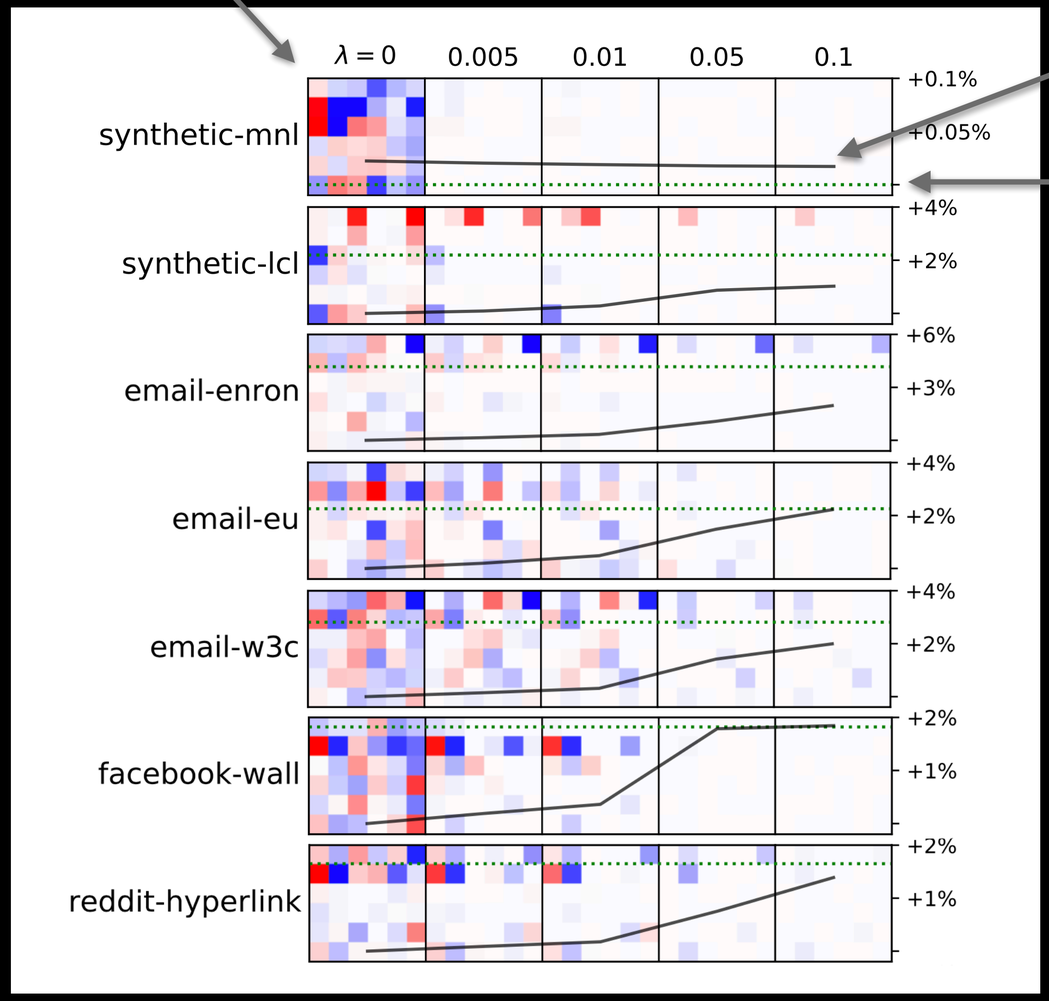
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



NLL (lower = better fit)

$p = 0.001$ LRT threshold

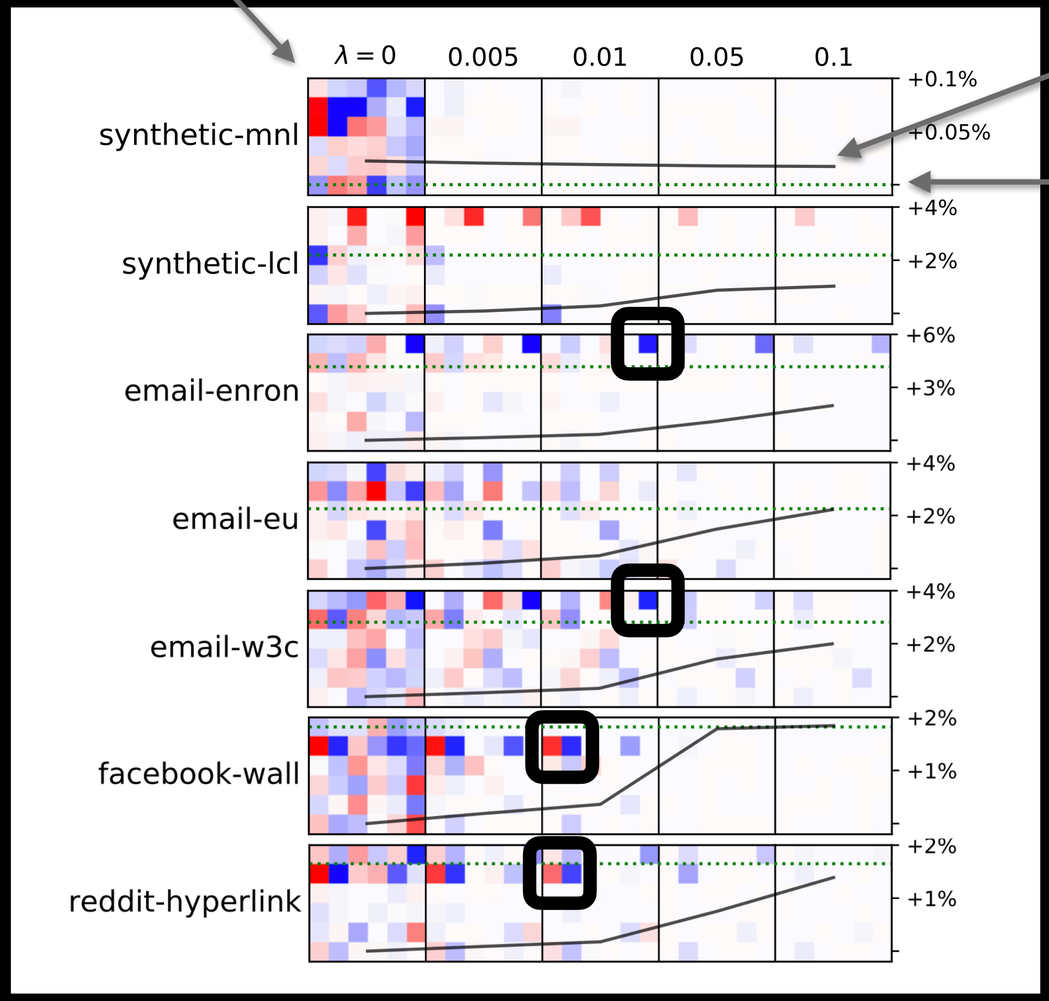
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



NLL (lower = better fit)

$p = 0.001$ LRT threshold

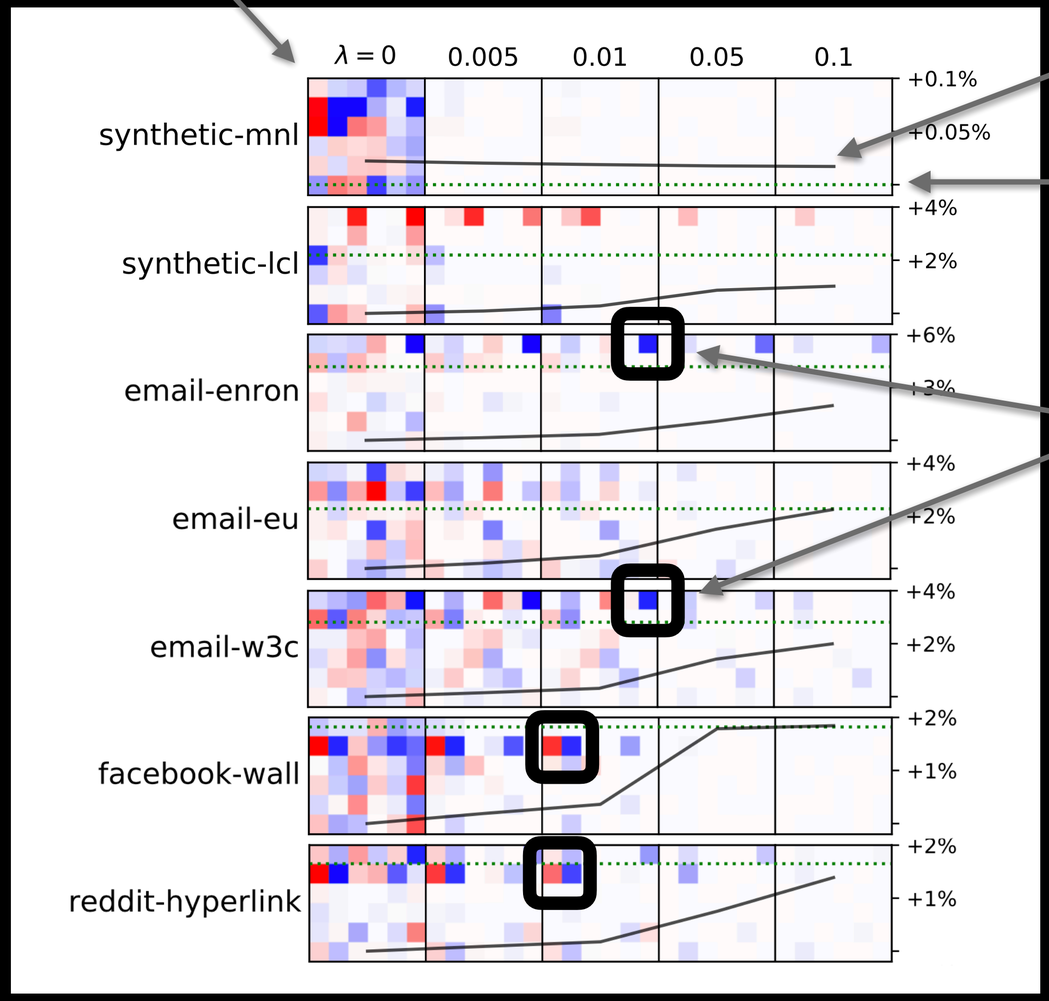
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



NLL (lower = better fit)

$p = 0.001$ LRT threshold

"cluttered inbox"
high choice set reciprocal recency
→ in-degree less important

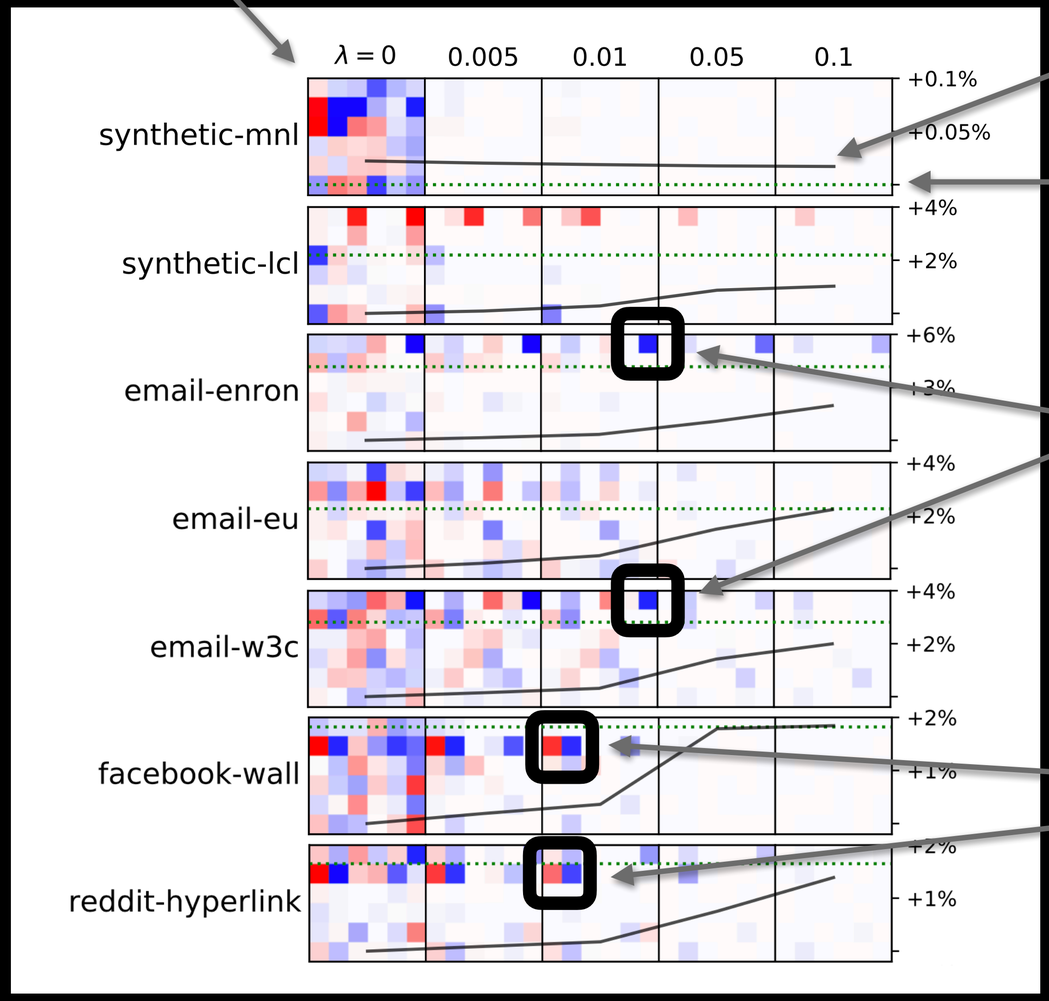
Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

LCL reveals interpretable feature context effects

context effect matrix A
red: +, blue: -, white: 0
(column acts on row)

increasing L_1 regularization on A



NLL (lower = better fit)

$p = 0.001$ LRT threshold

Node features
(left-right, top-bottom)

1. in-degree
2. shared neighbors
3. reciprocal weight
4. send recency
5. receive recency
6. reciprocal recency

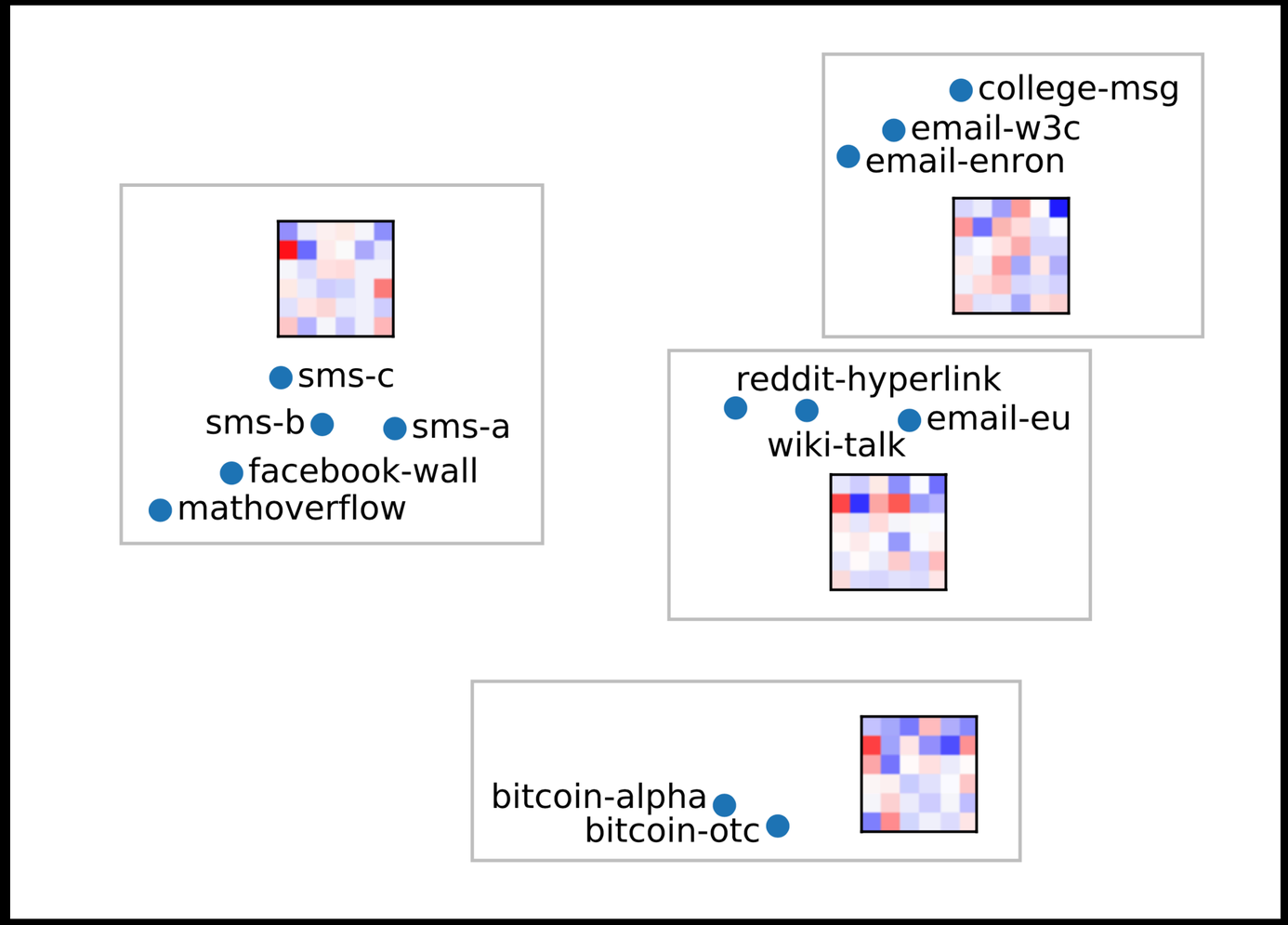
"cluttered inbox"
high choice set reciprocal recency
→ in-degree less important

red: *"cocktail party introduction"*
high choice set in-degree
→ shared neighbors more important

blue: *"familiarity saturation"*
high choice set shared neighbors
→ shared neighbors less important

Similar datasets have similar feature context effects

T-SNE embedding of learned A matrices



Concluding thoughts

Concluding thoughts

Key takeaways
Feature context effects extend item-level effects
LCL offers an interpretable and tractable way to reveal them

Concluding thoughts

Key takeaways
Feature context effects extend item-level effects
LCL offers an interpretable and tractable way to reveal them

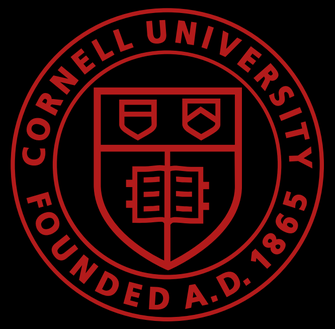
Challenges
Features correlate
Causal context effects?
Handling nonlinearity?

Concluding thoughts

Slides: bit.ly/lcl-slides-2
Preprint: bit.ly/lcl-paper
Code: bit.ly/lcl-code
Data: bit.ly/lcl-data

Key takeaways
Feature context effects extend item-level effects
LCL offers an interpretable and tractable way to reveal them

Challenges
Features correlate
Causal context effects?
Handling nonlinearity?



Thank you!

More questions or ideas?
Email me: kt@cs.cornell.edu
@kiran_tomlinson

Acknowledgments
Funding from NSF, ARO

Thanks to Johan Ugander and Jan Overgoor