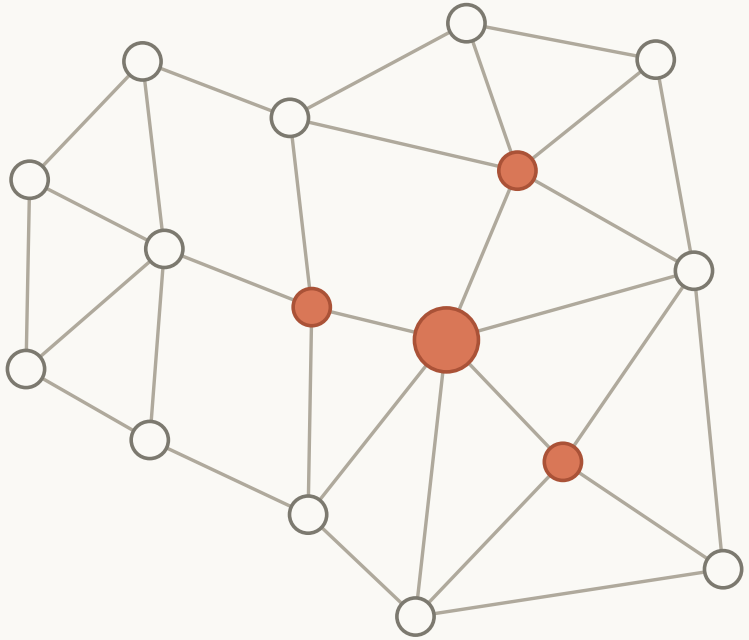
Ph.D. Mid-Term Defense

Causal Inference Under Network Interference



Presenter

Qianyi Chen

Supervisor

Bo Li

School of Economics and Management
Tsinghua University

How Platforms Decide What to Launch

At the first Practical Online Controlled Experiments Summit (Sunnyvale, 2018), 13 organizations reported testing over 100,000 treatments in one year.

Microsoft: the MSN mail button

The new button opened a mail app. Clicks rose 28%, but satisfaction showed no significant gain. The team stopped the test.



LinkedIn messaging

Control users can reply to treated senders. Mean lift in messages sent: 1.00% by standard A/B testing, 1.56% using sender–recipient pairs.



An illustrative launch question

If everyone could share with one tap, would more users return next week?

What Does a Randomized A/B Test Estimate?

Illustrative A/B test: 200 users assigned at random to two equal groups, with no interference

Control

Keep the existing sharing feature.

30 of 100 users return next week.

Treatment

Enable one-tap sharing.

35 of 100 users return next week.

Target: average treatment effect

$$\tau_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i(1) - Y_i(0)\right]$$

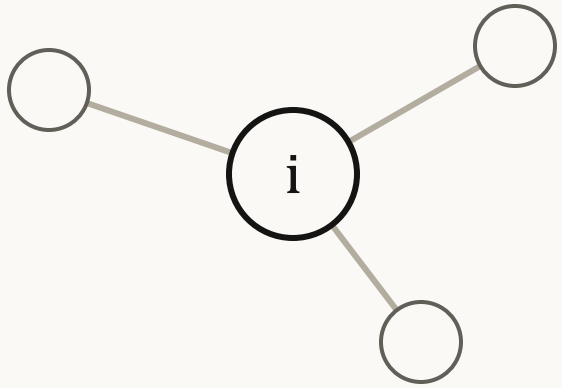
A return is coded as 1; no return as 0.

Estimate: +5 percentage points

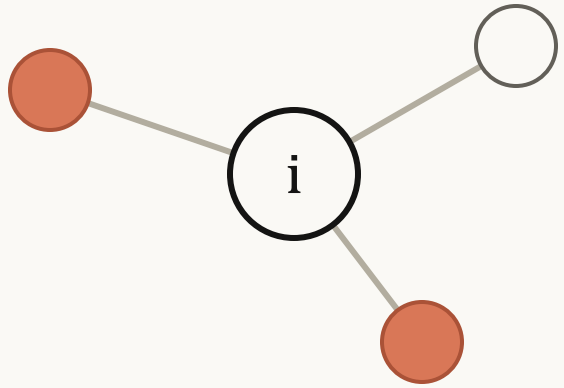
We see only one outcome per user. Under complete randomization, the gap in return rates is unbiased for the ATE.

When One User's Treatment Reaches Another

No contact has the feature



Two contacts have the feature



Treatment Control

The focal user stays in control. If treated friends send more messages, his own activity can still change.

Spillovers in Two Everyday Experiments

Social network

The focal user still uses the old interface.

- A friend gets one-tap sharing.
- He receives more messages and sends more replies.

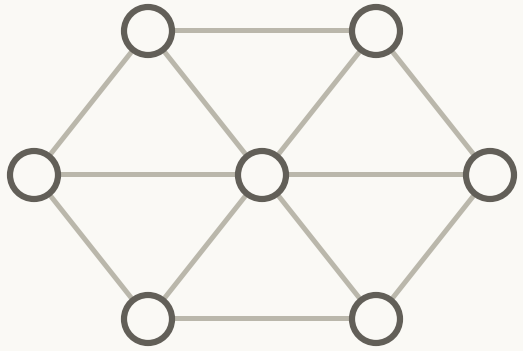
Online marketplace

The focal seller never receives the badge.

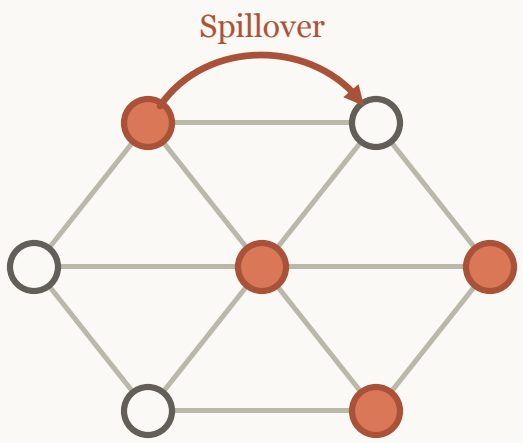
- Rival sellers get a promotion badge.
- Buyers switch to promoted listings, so the focal seller may lose sales.

The Rollout Decision Compares Two Global Regimes

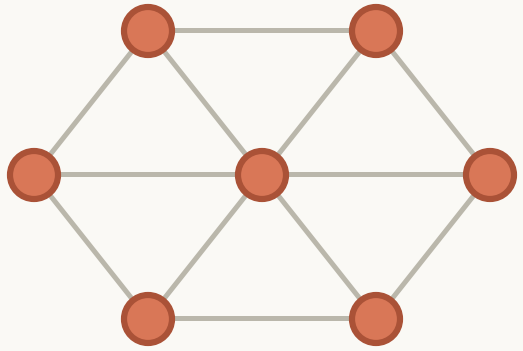
All control



Observed A/B test



All treated



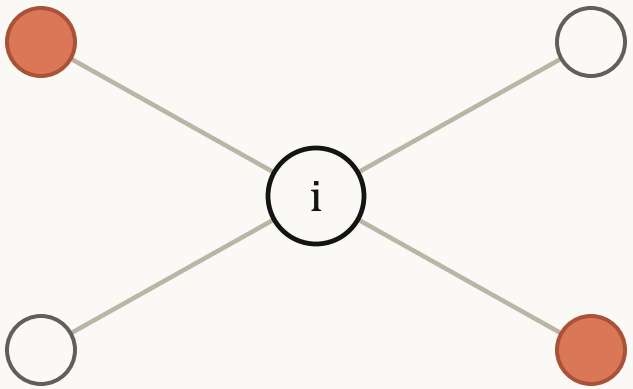
$$\tau_{\mathrm{GATE}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i(\mathbf{1}) - Y_i(\mathbf{0})\right]$$

Spillovers violate SUTVA's no-interference assumption: a unit's outcome may depend on all other units' treatments.

The experiment observes one mixed network. The policy question concerns both endpoints.

Network Exposure and the Rollout Effect

Count the treated neighbors



$$e_i := \frac{1}{d_i} \sum_{j \in N_i} z_j$$

A nonlinear exposure response

$$Y_i(\mathbf{z}) = z_i + e_i + e_i^2$$

Here e is network exposure: the share of treated neighbors. The last term allows spillovers to grow faster at high exposure.

Within the test: 1

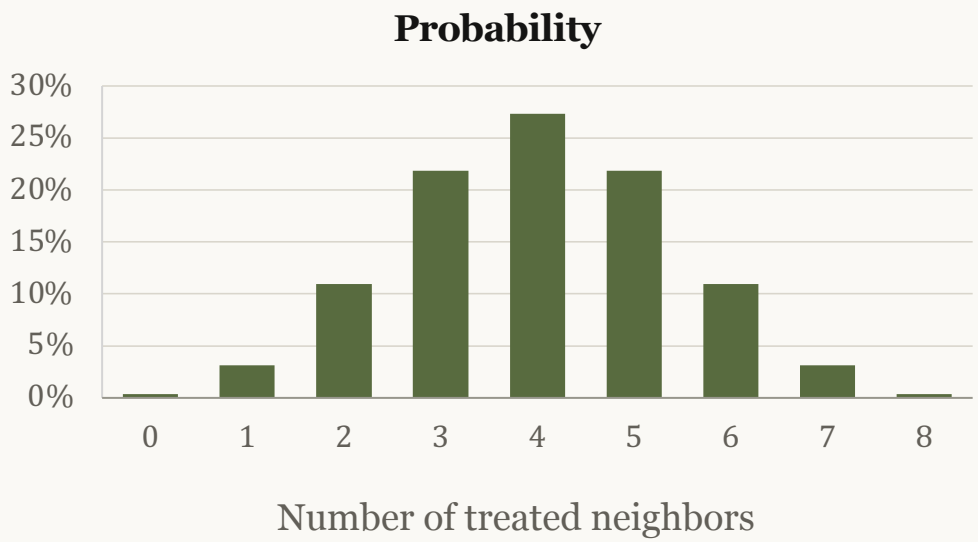
Both arms face the same exposure distribution.

Full rollout: 3 1 direct + 1 linear + 1 quadratic

How Rare Is Full Exposure?

Illustration: eight neighbors

Each user is independently treated with $p = 0.5$.



All eight treated: $1/256 \approx 0.39\%$.

Real topology: Facebook Stanford

11,586 users · 568,309 edges
Median user: 73 neighbors

$$\Pr(e_i = 1 \mid d_i) = 2^{-d_i}$$

3.22% of users are expected to have all their neighbors treated.

1.61% of users are expected to be treated along with all their neighbors.

Four Studies of the Rollout Effect

Experimental design

How should we assign treatment
to account for network interference?

OCD

NeurIPS 2023

Correlate treatment assignments
across cluster boundaries.

COSTA

arXiv:2609.02032

Design covariance across units and time
to account for carryover.

Estimation

What structure can we use
to estimate GATE with lower MSE?

AMII

ICLR 2026

Change the estimator: use interior outcomes
to correct predictions for the full network.

Ramp-Up

NeurIPS 2026

Change the training data: pool stages
with different treatment shares.

The Four Papers

OCD	**NeurIPS 2023**	Optimized Covariance Design for AB Test on Social Network Under Interference Qianyi Chen, Bo Li, Lu Deng, Yong Wang · with Tencent
COSTA	**arXiv:2609.02032**	COSTA: Covariance-Optimized Design and Causal Inference Under Network-Temporal Interference Qianyi Chen, Bo Li, Yongli Qin, Jinyong Ma · with ByteDance
AMII	**ICLR 2026**	Journey to the Centre of Cluster: Harnessing Interior Nodes for A/B Testing Under Network Interference Qianyi Chen, Anpeng Wu, Bo Li, Lu Deng, Yong Wang · with Tencent
Ramp-Up	**NeurIPS 2026**	Just Ramp-Up: Debiasing Regression-Based Estimator for A/B Tests Under Network Interference Qianyi Chen, Bo Li

Experimental Design

How should we assign treatment to account for network interference?

---

01 / OCD

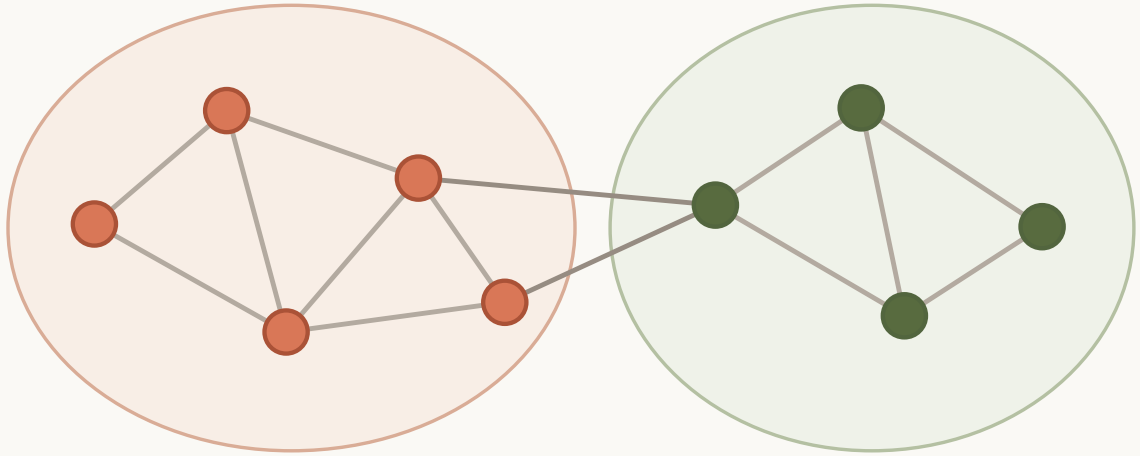
Correlate assignments across a network.

02 / COSTA

Coordinate assignments across units and time.

Randomizing Graph Clusters

One coin flip per cluster



Users in a cluster share the same assignment.
Spillovers can still cross cluster boundaries.

Two foundations

Hudgens & Halloran / JASA 2008

"no interference between individuals in different groups"

Ugander et al. / KDD 2013

"using graph clustering to analyze average treatment effects under social interference"

Why Cluster Randomization Helps

01	**Bias–variance tradeoff**	Within-cluster assignment correlation is one. This can reduce spillover bias at the cost of fewer independent assignments.
02	**Optimization**	A design over K clusters replaces a design over n individuals. The smaller covariance matrix makes optimization tractable.
03	**Endpoint extrapolation**	Many local neighborhoods now match all treatment or all control. Their outcomes anchor the global rollout contrast.
04	**Learning a response curve**	A wider spread of network exposure gives regression models, including GNNs, more information about spillovers.

Bias as a Trace of the Assignment Covariance

The standard Horvitz–Thompson estimator targets the all-treated versus all-control contrast.

A one-hop working model

$$Y_i(\mathbf{Z}) = \alpha_i + \beta_i Z_i + \gamma \sum_{j \in N_i} Z_j$$

Z_i is the treatment of unit i, inherited from its cluster. Each treated neighbor adds γ.

Exact bias under balanced assignment

$$B(V) = \frac{\gamma}{n}\left[4\,\mathrm{tr}(CV) - \mathbf{1}^\top C \mathbf{1}\right]$$

C counts links between clusters. The cluster-level treatment vector is $\mathbf{z}$.

$$V_{k\ell} = \mathrm{Cov}(z_k, z_\ell), \qquad V = \mathrm{Cov}(\mathbf{z})$$

Within-cluster edges cancel. For a linked pair of clusters, stronger assignment correlation reduces the magnitude of its bias contribution.

Third and Fourth Moments in the Exact Variance

After baseline adjustment, the HT estimator has a linear and a quadratic term in the cluster assignments.

$$L = h^\top \mathbf{z}, \qquad Q = \mathbf{z}^\top C \mathbf{z}, \qquad \widehat{\tau}_{-\alpha} = \frac{2}{n}(L + 2\gamma Q)$$

$$\frac{n^2}{4}\operatorname{Var}(\widehat{\tau}_{-\alpha}) = \operatorname{tr}(hh^\top V) + 4\gamma \operatorname{Cov}(L, Q) + 4\gamma^2 \operatorname{Var}(Q)$$

Second order

The trace depends on the assignment covariance V.

Third order

The cross term combines one assignment with a pair.

Fourth order

The variance of Q involves products of four assignments.

OCD, Assumption 2 and Proposition 2, equation (12), p. 5; proof in Appendix A.2

A Variance Bound With Only Second Moments

Cluster-level comparability controls effect sizes. Binary assignments bound higher-order terms.

Direct and spillover effects by cluster

$$h_k = \sum_{i \in S_k} (\beta_i - \gamma d_i), \qquad |h_k| \le \omega \gamma d_k$$

Here d_k is the sum of node degrees in cluster k. The comparison allows individual effects to vary.

Bounding the cross and quadratic terms

$$(L + 2\gamma Q)^2 \le 2L^2 + 8\gamma^2 Q^2, \qquad 0 \le Q \le d^\top \mathbf{z}$$

C and the joint probabilities are entrywise nonnegative. Each assignment is zero or one.

$$\mathrm{Var}(\widehat{\tau}_{-\alpha}) \le \frac{8\gamma^2(\omega^2 + 4)}{n^2}\, d^\top \left(V + \tfrac{1}{4}\mathbf{1}\mathbf{1}^\top\right) d$$

Squared bias and the bound share γ^2. Given ω, the design needs no fitted effect sizes.

OCD, Assumption 3, Proposition 3 and Theorem 1, pp. 6–7; proof in Appendix A.3

A Feasible Family of Bernoulli Covariances

The Gaussian vector $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \Omega)$ contains one latent score per cluster. Treat cluster k when $G_k > 0$.

Grothendieck's identity for two Gaussian scores

$$\operatorname{Cov}(\operatorname{sgn} G_k, \operatorname{sgn} G_\ell) = \frac{2}{\pi} \arcsin(\Omega_{k\ell})$$

Scores have unit variance and correlation $\Omega_{k\ell}$. Coding signs as $0/1$ divides their covariance by four.

$$V = \frac{1}{2\pi} \arcsin(RR^\top)$$

OCD fits a full matrix R with unit-length rows. Its Gram matrix $RR^\intercal$ is positive semidefinite with unit diagonal. A triangular Cholesky factor is optional.

Projected Optimization and Gaussian Sampling

Optimize the Gaussian factor once, then reuse it to draw new cluster assignments.

Projected updates of the covariance factor

$$\min_{\|R_{k\cdot}\|_2=1} \; B(V)^2 + \overline{V}_\omega(V)$$

Constraints are imposed on the diagonal terms: variance of Bernoulli random variable is $p(1-p)$.

We prove that a projection operator to the feasible domain is row normalization on R.

Draw the assignments

$$\mathbf{z} = \mathbf{1}\{R\boldsymbol{\xi} > 0\}, \quad \boldsymbol{\xi} \sim \mathcal{N}(0, I_K)$$

Apply the fitted factor to fresh Gaussian noise. Treat clusters whose scores are positive.

Each cluster has marginal treatment probability ½.

$$\Pr(z_k = 1) = \Pr(G_k > 0) = \tfrac{1}{2}$$

OCD, equations (19)–(20) and Algorithm 1, pp. 7–8

Results Under Multiplicative Interference

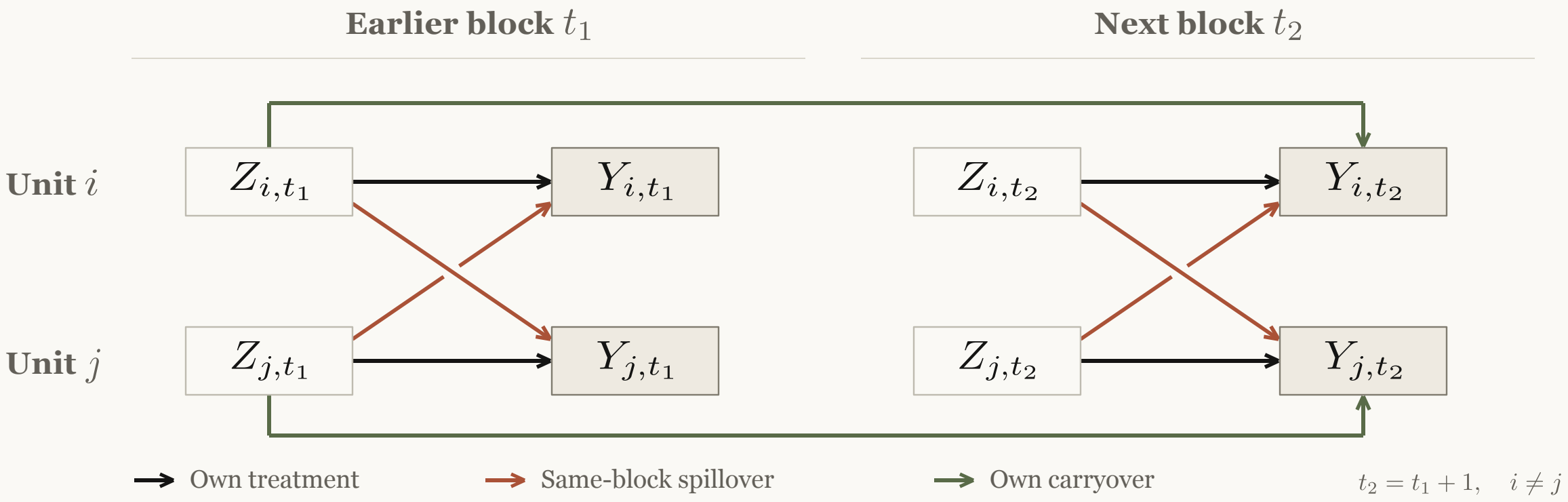
OCD has the lowest mean squared error at each of the three spillover strengths in Table 2.

	$\gamma = 0.5$			$\gamma = 1.0$			$\gamma = 2.0$		
Design	**Bias**	**SD**	**MSE**	**Bias**	**SD**	**MSE**	**Bias**	**SD**	**MSE**
Ber	−0.365	0.348	0.255	−0.736	0.394	0.698	−1.475	0.493	2.421
CR	−0.368	0.235	0.191	−0.744	0.274	0.629	−1.477	0.336	2.297
ReAR	−0.402	0.178	0.194	−0.809	0.174	0.685	−1.548	0.226	2.450
PSR	−0.366	0.134	0.152	−0.738	0.153	0.569	−1.479	0.192	2.227
IBR	−0.369	0.155	0.161	−0.737	0.178	0.576	−1.484	0.221	2.252
IBR-p	−0.368	0.163	0.163	−0.739	0.185	0.581	−1.482	0.232	2.252
OCD	**−0.258**	**0.040**	**0.069**	**−0.517**	**0.050**	**0.271**	**−1.034**	**0.054**	**1.073**

At $\gamma = 2$, OCD reduces MSE from 2.227 with pairwise-sequential randomization to 1.073.

OCD, Table 2, FB-Stanford3 network, 192 Louvain clusters, 10,000 randomizations per design

Network Spillovers and Temporal Carryover



One-lag working model

$$Y_{it} = \mu_{it} + \beta_{it} Z_{it} + \gamma E_{it} + \eta Z_{i,t-1}$$

Z is treatment, Y is outcome and E is same-block neighbor exposure. μ is the baseline; β, γ and η are direct, network and carryover coefficients.

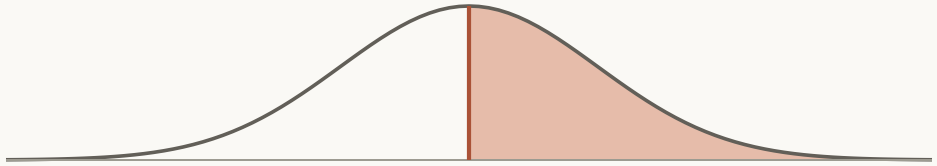
COSTA, §§2–3, eqs. (1)–(10)

A Gaussian Copula for Any Treatment Probability

Both methods draw correlated Gaussian scores. COSTA can target any common treatment probability.

OCD uses a zero threshold

$$z_k = \mathbf{1}\{G_k > 0\}, \qquad \Pr(z_k = 1) = \tfrac{1}{2}$$

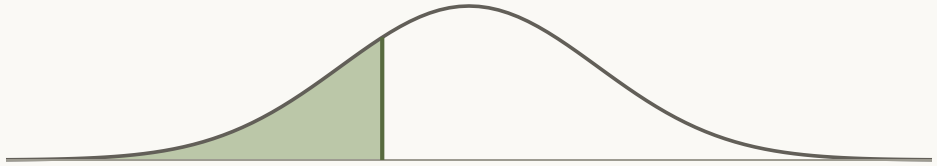


Upper half receives treatment

OCD uses Grothendieck's identity for Gaussian signs. Cluster k is equally likely to fall on either side of zero.

COSTA chooses a quantile

$$Z_a = \mathbf{1}\{\Phi(G_a) \le p\} = \mathbf{1}\{G_a \le \Phi^{-1}(p)\}$$



Lowest quarter treated in this example

For cell a, the normal CDF Φ gives its uniform percentile. The cutoff at p makes its treatment probability exactly p.

At fixed p, Ω controls dependence across cells. The realized treatment share can vary across draws.

From Gaussian Scores to Treatment Correlation

Two cells receive treatment together when both Gaussian scores fall below the same cutoff.

Joint probability

$$\Pr(Z_a = 1, Z_c = 1) = \Phi_2(t_p, t_p; \Omega_{ac}), \qquad t_p = \Phi^{-1}(p)$$

Binary correlation

$$R_{ac} = \frac{\operatorname{Cov}(Z_a, Z_c)}{p(1-p)} = \frac{\Phi_2(t_p, t_p; \Omega_{ac}) - p^2}{p(1-p)}$$

Φ_2 is the bivariate standard normal CDF. Its correlation parameter Ω_{ac} belongs to the latent scores.

Balanced treatment

$$p = \tfrac{1}{2}: \qquad R_{ac} = \frac{2}{\pi}\arcsin(\Omega_{ac})$$

At one half, the binary variance is one quarter. This arcsine map gives four times the raw covariance in OCD.

Ten percent treated

$$p = 0.1: \qquad -\tfrac{1}{9} \le R_{ac} \le 1$$

The lower bound occurs when the cells are never treated together; the upper bound when they always agree.

Pairwise limits alone do not ensure a joint design. COSTA uses one positive semidefinite latent correlation matrix.

A Cut Edge Has a Bias Cost

The HT score links assignments to bias

$$\psi_a(Z_a) = \frac{Z_a - p}{p(1-p)}, \qquad \mathbb{E}_D[\psi_a(Z_a)Z_c] = R_{ac}$$

Exact bias under linear exposure

$$\mathbb{E}_D[\widehat{\tau}] - \tau = -\sum_{a,c} h_{ac} \frac{\Pr_D(Z_a \neq Z_c)}{2p(1-p)}$$

The cross-moment equals the treatment correlation under any assignment law with common marginals. With positive exposure weights, a higher cut probability lowers the expected HT contrast.

Covariance and the cut probability

$$\Pr_D(Z_a \neq Z_c) = 2p(1-p)(1-R_{ac})$$

R_{ac} is the treatment correlation on this edge. With nonnegative weights, zero bias requires cells linked by positive edges to move together.

A second-moment design criterion

$$\mathrm{MSE}_D(\widehat{\tau}) \leq B(R)^2 + \overline{V}(R)$$

$B(R)$ is the exact bias. A bound on effect sizes gives the variance envelope. COSTA also discourages imbalanced blocks and too few switches.

COSTA, Theorems 3.2 and 3.5, Corollary 3.6 and Proposition 3.8

Kronecker Factors Make Joint Design Highly Scalable

Kronecker factors parameterize the latent Gaussian correlation, before the nonlinear Bernoulli map.

Latent correlation

$$\Omega_{(i,t),(j,r)} = (f_i^\top f_j)(g_t^\top g_r)$$

Item factor f_i is shared across blocks; time factor g_t is shared across items. Both have unit length. Ω is the latent correlation matrix.

A valid joint Bernoulli draw

$$Z_{it} = \mathbf{1}\{f_i^\top W g_t \le \Phi^{-1}(p)\}$$

Draw one matrix W of independent standard normal scores. The normal quantile sets each cell's treatment probability to p.

6.4 billion entries

A full covariance matrix for 10,000 items observed over 8 blocks.

160,056 coefficients

The default design uses 16 factor coefficients per item and 7 per block.

COSTA, §§4–5, eqs. (44)–(51) and Algorithm 1

Six Designs on Two Item Networks

Reported RMSE for three simulated outcome models

	RetailRocket			MovieLens		
Design	**Linear**	**Nonlinear**	**Demand**	**Linear**	**Nonlinear**	**Demand**
Independent	0.185	0.239	0.125	0.185	0.235	0.123
Balanced switchback	0.196	0.255	0.166	0.198	0.256	0.164
Graph clusters	0.123	0.166	0.194	0.141	0.185	0.217
Static OCD	0.131	0.173	0.197	0.159	0.200	0.233
COSTA	**0.066**	**0.071**	**0.056**	**0.102**	**0.111**	**0.108**
COSTA-LP	**0.073**	**0.091**	**0.062**	**0.100**	**0.117**	**0.098**

In all six settings, COSTA and its locality-penalized variant have lower RMSE than the four baselines.

COSTA, Tables 1–2, §7.3

Which Randomness Does the Interval Describe?

IC-0: under repeated assignments, is the HT estimate approximately normal around its design mean?

Conditional central limit theorem

$$\frac{\widehat{\tau} - \mu_D}{s_D} \Rightarrow \mathcal{N}(0,1)$$

μ_D and s_D are the mean and standard deviation under this design. The theorem requires overlap, bounded moments of order greater than four, nondegenerate variance and dependence that decays fast enough relative to graph growth.

Dependence under the assignment law

$$\rho(\ell) = \sup_{d(A,B)\geq\ell} \mathrm{CanCorr}(G_{A^+}, G_{B^+})$$

A and B are outcome sets at least ℓ edges apart.
The + adds their assignment neighbors.
G contains the latent scores.
Canonical correlation is the strongest correlation between projections of the two score vectors.

In the shared-factor example, pairwise correlation is 0.01, yet the two 1,000-score means correlate at 0.91.

When the Design Center Misses the GATE

IC-1 requires the design bias to vanish in standard-error units

$$\frac{\widehat{\tau} - \tau}{s_D} = \frac{\widehat{\tau} - \mu_D}{s_D} + \frac{\mu_D - \tau}{s_D}$$

τ is the sustained GATE. The first term fluctuates around zero.
For a raw HT interval to attain nominal GATE coverage, the bias in standard-error units must vanish.

Why the path matters



Different endpoint assignments force a cut.
Positive exposure weights turn each cut into bias.

When centering fails for the GATE

$$\frac{|\mu_D - \tau|}{s_D} \longrightarrow \infty$$

Assume persistent positive spillovers and decorrelation at long distances. The standard error has order $1/\sqrt{M}$ for M outcome cells.

Estimating Variance Under Designed Dependence

IC-2 adds a consistent estimate of the design variance

$$\frac{\widehat{V}_D}{V_D} \xrightarrow{p} 1 \quad \Longrightarrow \quad \frac{\widehat{\tau} - \mu_D}{\sqrt{\widehat{V}_D}} \Rightarrow \mathcal{N}(0,1) \quad \text{(IC-0)}$$

V_D is the variance across repeated assignments. With IC-0, a consistent standard error gives an interval for the design mean. Coverage of the GATE also requires IC-1 or a valid bias correction.

Estimate local covariance

Subtract each cell's own expected HT contribution. Graph HAC sums centered products of nearby cells.

Add the omitted covariance tail

$$\widehat{s}_D^2 = \left[\widehat{V}_{\text{local}} + \max(0, \Delta)\right]_+$$

Δ is full minus local variance under fresh simulated assignments and outcomes. Adding its positive part is conservative. Consistency requires the negative tail to be negligible.

Coverage Depends on the Interval Target

Positive-tail intervals, nominal coverage 90%

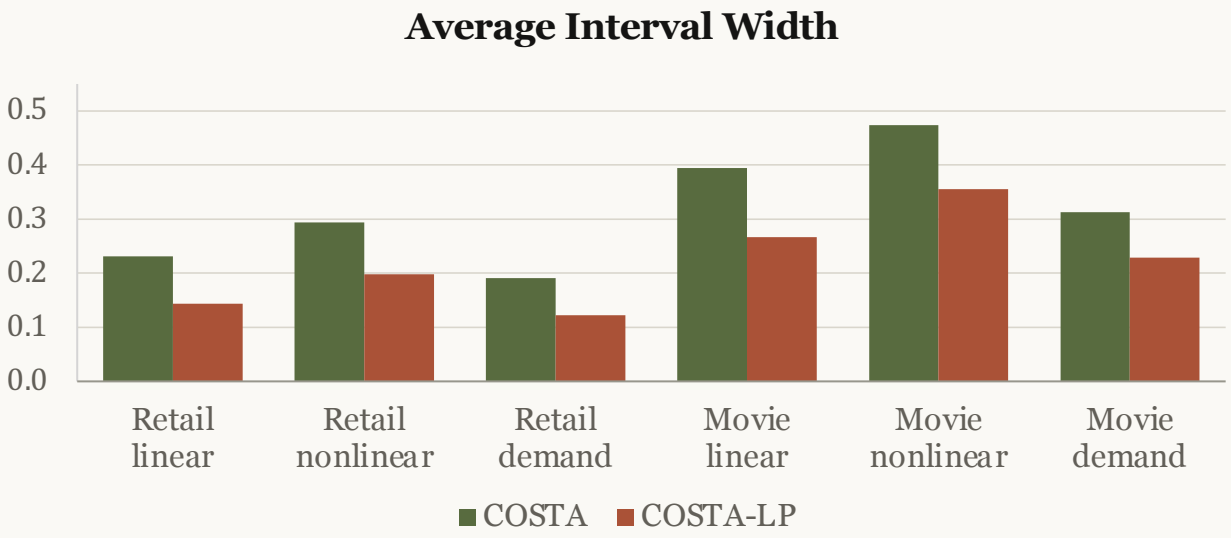
		COSTA		COSTA-LP	
Dataset	**Outcome**	**Design / RC**	**Raw GATE**	**Design / RC**	**Raw GATE**
RetailRocket	Linear	92.0%	84.3%	90.4%	71.3%
	Nonlinear	90.3%	85.1%	92.0%	69.6%
	Demand	89.9%	88.2%	89.1%	78.8%
MovieLens	Linear	92.2%	90.9%	89.7%	84.1%
	Nonlinear	91.3%	89.7%	93.2%	84.2%
	Demand	90.0%	89.9%	88.8%	86.4%

Recentered GATE coverage (RC) matches design-centered coverage after the same calibrated bias shift.

COSTA, Table 5, positive-tail model-assisted intervals

Interval Width and Standard-Error Calibration

Mean width of the nominal 90% interval



Retail = RetailRocket; Movie = MovieLens.

Estimated SE / Monte Carlo SD

Dataset / model	COSTA	LP
RetailRocket / linear	1.015	1.001
Nonlinear	1.002	1.063
Demand	0.994	0.986
MovieLens / linear	1.030	0.981
Nonlinear	0.999	1.078
Demand	0.968	0.991

A ratio of 1 means the estimated SE matches the standard deviation across repeated draws.

COSTA-LP narrows intervals by 25–38%, with estimated standard errors close to the Monte Carlo SDs.

Estimation

Which features of the observed data can we use to estimate GATE with lower mean squared error?

---

03 / AMII

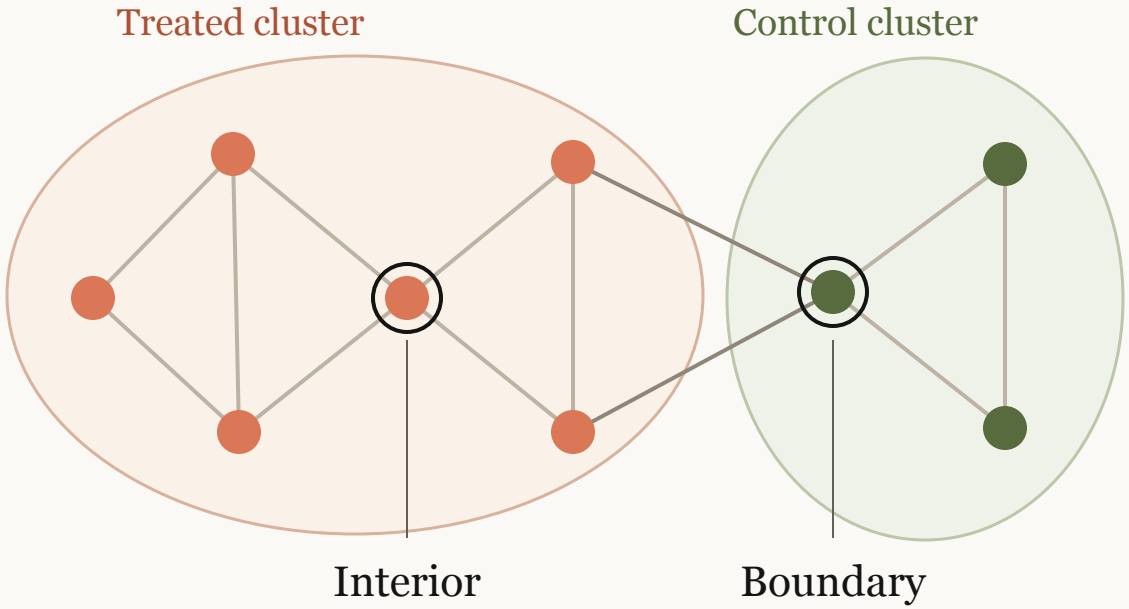
Correct predictions with reliable interior labels.

04 / Ramp-Up

Pool stages with distinct traffic shares.

Cluster Interiors Supply Reliable Outcome Labels

Two clusters, one treatment each



Interior All one-hop neighbors lie in the same cluster.

Boundary All other nodes have a cross-cluster neighbor.

The observed label

Every interior node shares its assignment with all of its neighbors.

Under one-hop interference, the node observes its all-treated or all-control outcome, according to the assignment of its own cluster.

FB-Stanford3: 7.9% interior across 95 clusters.

What the Baseline Estimators Average

Standard difference in means

Average observed outcomes by own treatment; mixed neighborhoods remain in both arms.

Full-exposure Horvitz–Thompson

$$\widehat{\mu}_a^{\mathrm{HT}} = \frac{1}{n} \sum_i \frac{\delta_i(a) Y_i}{\Pr\{\delta_i(a) = 1\}}$$

At 10% traffic, four independent clusters in a closed neighborhood give a treated label weight of 10,000.

Cluster-adaptive estimator (CAE)

$$\bar{Y}_{k,a}^{\mathrm{clean}} = \frac{\sum_{i \in C_k} \delta_i(a) Y_i}{\sum_{i \in C_k} \delta_i(a)}$$

$$\widehat{\mu}_a^{\mathrm{CAE}} = \frac{1}{K_a} \sum_{k : t_k = a} \bar{Y}_{k,a}^{\mathrm{clean}}$$

First average clean outcomes within each cluster; then average the clusters assigned to that arm.

δ marks a clean outcome; K_a counts clusters in arm a.

AMII, §2.2, Eqs. (3–6) · clean = own and all neighbors' assignments match the arm

Interior Labels Can Be Selectively Observed

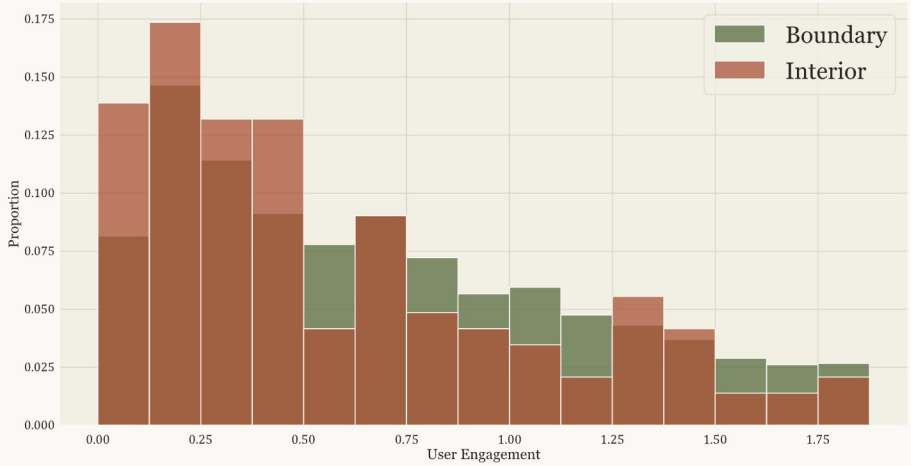
Mean-in-interior (MII)

$$\widehat{\tau}_{\mathrm{MII}} = \frac{1}{|I_1|}\sum_{i\in I_1} Y_i - \frac{1}{|I_0|}\sum_{i\in I_0} Y_i$$

$$I_1 = \{i \in I : z_i = 1\}, \qquad I_0 = \{i \in I : z_i = 0\}$$

A clean label need not be representative

Interior and boundary users differ in engagement.
If engagement modifies the treatment effect,
MII can miss the population effect.



WeChat social network subgraph · anonymized data

AMII Adjusts for Selection With Predictions

Form 1 / Correct the interior mean

$$\widehat{\mu}_a^{\mathrm{AMII}} = \underbrace{\bar{Y}_{I_a}}_{\text{interior labels}} + \underbrace{\bar{f}_{\mathrm{all}}(a) - \bar{f}_{I_a}(a)}_{\text{predicted population gap}}$$

1. Fit the outcome model

Fit the model to the observed graph, covariates and assignments. A neural network can learn how degree modifies the treatment response.

2. Predict one global arm

Give every unit assignment a.
Compare the mean prediction for all users with the mean for interior users in arm a.

AMII as Semi-Supervised Estimation

Form 2 / Correct the population predictions

$$\widehat{\mu}_z = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_i(z\mathbf{1}) + \frac{1}{|I_z|}\sum_{i\in I_z}\left[Y_i - \widehat{f}_i(z\mathbf{1})\right]$$

Predict outcomes for all users. Use interior labels to correct each arm's mean prediction.

Why the adjustment can reduce bias

$$\mathrm{Bias}(\widehat{\tau}_{\mathrm{MII}}) = \alpha\,\Delta$$

$$\mathrm{Bias}(\widehat{\tau}_{\mathrm{AMII}}) = \left(\alpha - \mathbb{E}[\widehat{\alpha}]\right)\Delta$$

Δ is the interior–population covariate gap; α is its treatment interaction. The result assumes the paper's interaction model.

AMII Across Traffic Levels and Spillover Ranges

Mean squared error / lower is better

Estimator	1-hop · 10%	1-hop · 30%	1-hop · 50%	2-hop · 10%	2-hop · 30%	2-hop · 50%
Hájek	1.243	1.001	0.901	2.520	1.814	1.416
CAE	1.014	0.937	0.917	2.538	2.076	1.692
MII	0.971	0.940	0.933	2.239	1.822	1.544
GNN	3.235	0.452	0.198	4.494	1.336	0.641
AMII	**0.526**	**0.088**	**0.037**	**1.311**	**0.563**	**0.241**

At 30% traffic, AMII reduces one-hop MSE from 0.940 to 0.088 relative to MII.
Under two-hop spillover, AMII still has absolute bias 0.731.

When Does Prediction Adjustment Pay Off?

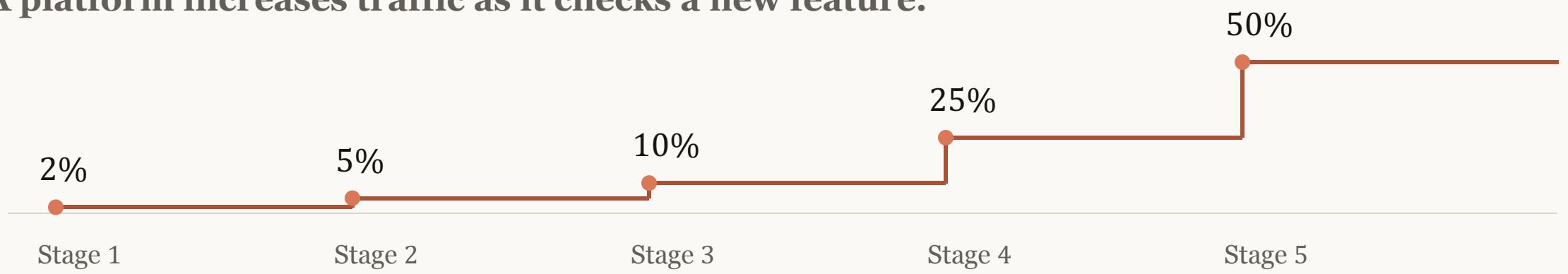
Ablation / matched traffic and graph partition

Covariates modify effects?	Spillover	MII MSE	AMII MSE
Yes	One hop	0.940	0.088
Yes	Two hops	1.822	0.563
No	One hop	0.017	0.035
No	Two hops	0.165	0.247

When the interior already represents the population, adjustment can cost precision.
In the one-hop ablation, SD rises from 0.130 to 0.177.

Most Recent Stage or the Whole Rollout?

A platform increases traffic as it checks a new feature.



Latest stage only

Fit the predictor to the 50% stage.
Earlier observations do not enter the fit.

Merge stages

Stack observations from several stages and fit one predictor. Each stage contributes another exposure regime.

A Regression Target Under General Linear Interference

The response model permits multihop spillover.

$$Y_t = \beta_0 \mathbf{1} + \beta_1 z_t + B z_t + \varepsilon_t, \qquad B_{ii} = 0$$

In a network of n users, each entry of B gives the effect of one user's assignment on another user's outcome. Errors are independent across users and stages, with common variance.

$$\tau = \beta_1 + \bar{b}, \qquad \bar{b} = \frac{1}{n} \sum_{i,j} B_{ij}$$

For theory, fit OLS to an intercept and own treatment. Its slope estimates GATE.
In simulations, a GNN also uses the network.

One Stage Leaves a Finite-Sample Spillover Bias

Complete randomization assigns exactly a fraction c to treatment.

$$D = nc(1-c), \qquad H = (z - c\mathbf{1})^\top B z$$

Exact bias

$$\mathrm{Bias}(\widehat{\tau}_{(1)}) = -\frac{n\bar{b}}{n-1}$$

Exact variance decomposition

$$\mathrm{Var}(\widehat{\tau}_{(1)}) = \frac{\mathrm{Var}_z(H)}{D^2} + \frac{\sigma_e^2}{D}$$

Under regular interference, bias persists as the network grows while variance shrinks in proportion to $1/n$. Assignment variability and outcome noise both contribute to variance.

Just Ramp-Up, Thm. 3.1, Cor. 3.3, Eqs. (9–13), Appendix C.1

Two Distinct Shares Reduce the Leading Bias

Stack the two stages and fit one slope.

$$\bar{c} = (c_1 + c_2)/2, \quad D_2 = 2n\bar{c}(1-\bar{c}), \quad H_t = (z_t - \bar{c}\mathbf{1})^\top B z_t$$

Exact finite-sample bias

$$\mathrm{Bias}(\widehat{\tau}_{(2)}) = -\frac{n\bar{b}}{n-1}\left[1 - \frac{(c_2-c_1)^2}{(c_1+c_2)(2-c_1-c_2)}\right]$$

Exact variance decomposition

$$\mathrm{Var}(\widehat{\tau}_{(2)}) = \frac{\mathrm{Var}_z(H_1) + \mathrm{Var}_z(H_2) + 2\,\mathrm{Cov}_z(H_1, H_2)}{D_2^2} + \frac{\sigma_e^2}{D_2}$$

Independent stage assignments remove the covariance term. Staggered rollout generally retains it.

Bias Depends on Dispersion Across Stages

Hold the number of stages fixed as the network grows.

$$\bar{c} = T^{-1}\sum_{t=1}^{T} c_t, \qquad v_c = T^{-1}\sum_{t=1}^{T}(c_t - \bar{c})^2$$

$$\mathrm{Bias}(\widehat{\tau}_{(T)}) = -\bar{b}\left[1 - \frac{v_c}{\bar{c}(1-\bar{c})}\right] + O(n^{-1}), \qquad \mathrm{Var}(\widehat{\tau}_{(T)}) = \Theta(n^{-1})$$

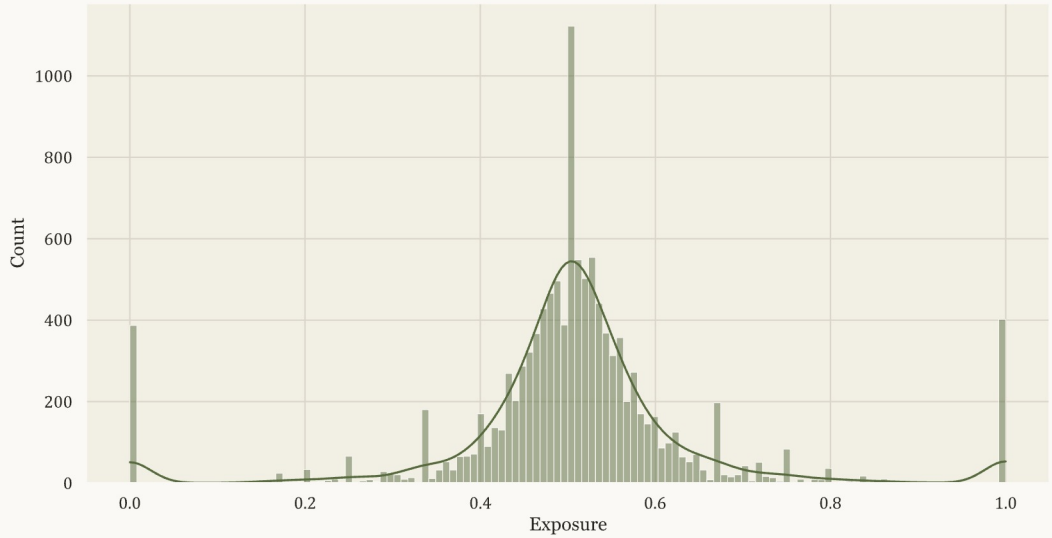
Traffic schedule	Neighborhood exposure variance	Leading absolute bias fraction
10%, 50%	0.0587	81.0%
10%, 25%, 50%	0.0459	86.6%

Adding the middle stage narrows neighborhood exposure here.
The last column expresses the leading OLS bias as a share of average spillover magnitude.

How Assignment Changes Neighborhood Exposure

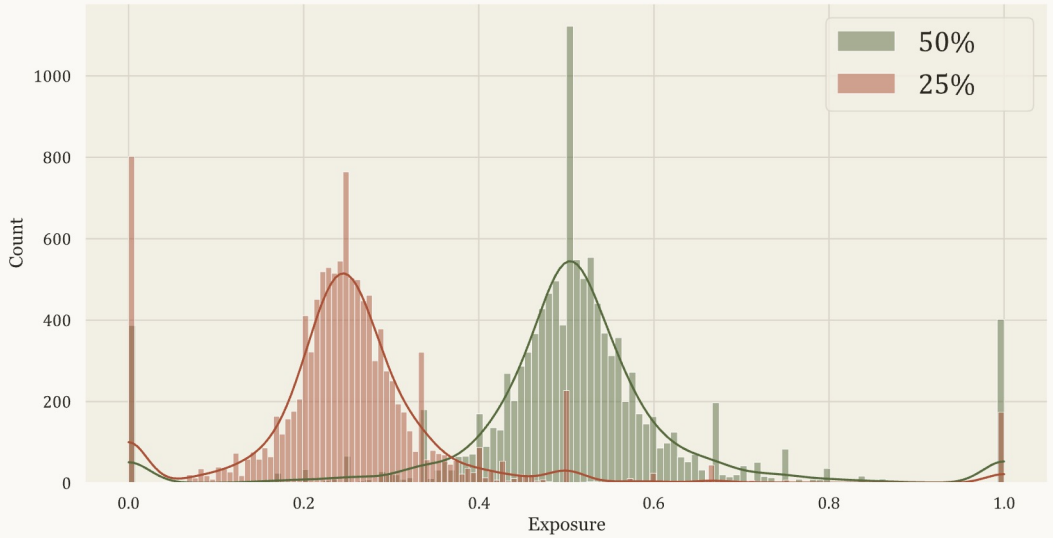
Exposure is the fraction of a user's neighbors receiving treatment.

Unit assignment at 50% traffic



Exposure variance 0.0243

Unit assignment at 25% and 50% traffic



Exposure variance 0.0377

At the same 50% traffic, assigning clusters instead of users yields exposure variance 0.0712.

Does More Data at Fixed Traffic Reduce Bias?

Two-hop simulation on FB-Stanford3 / unit randomization, 100 repetitions

Pooled stages	1	2	3	4	5
MSE, distinct shares	**1.963**	**0.374**	**0.100**	**0.025**	**0.026**
MSE, repeat 50%	1.968	2.132	2.189	2.208	2.214
Bias, repeat 50%	−1.402	−1.460	−1.480	−1.486	−1.488
SD, repeat 50%	0.034	0.015	0.012	0.012	0.012

Five rounds contribute the same number of training rows.
At fixed 50% traffic the estimate becomes less variable, but its absolute bias grows.

Just Ramp-Up, Tables 1(a) and 2, pp. 9–10

When Do Intermediate Stages Help?

GNN mean squared error under individual randomization, 100 repetitions

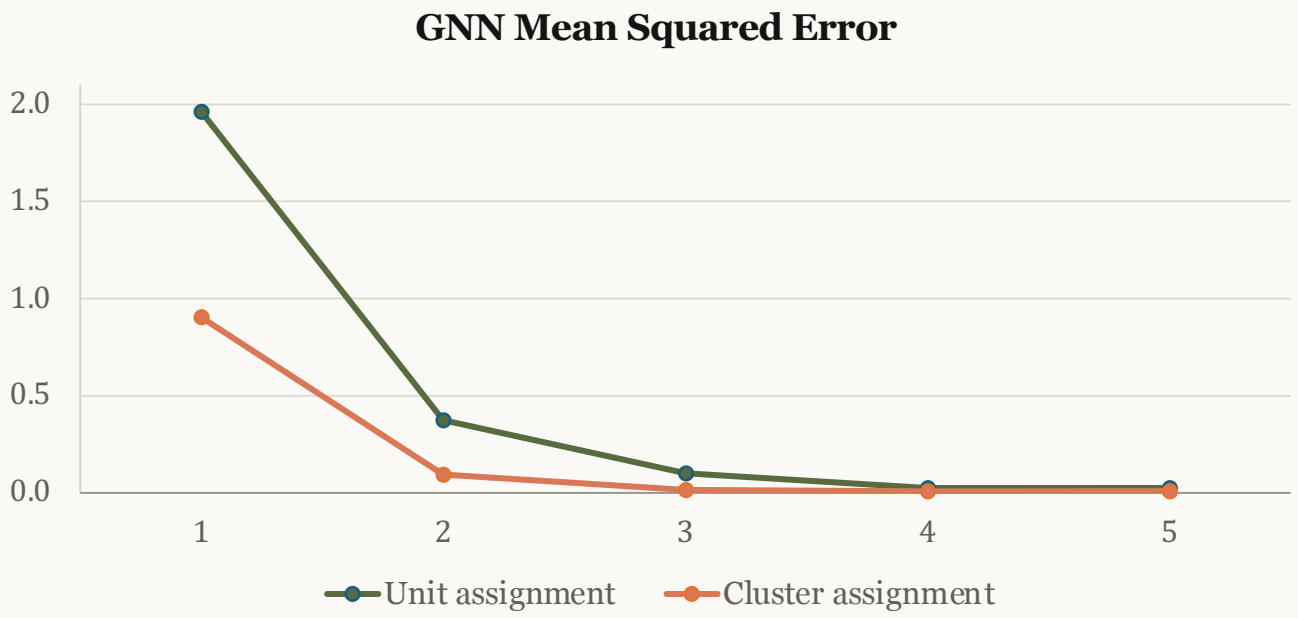
Outcome model	Latest: 50%	Near 0% + 50%	All five stages
Linear-in-means	0.909	0.009	0.013
Multiplicative	1.163	0.030	0.025
Quadratic	0.955	0.262	0.196
Square root	0.821	0.048	0.000*
Two-hop	1.963	0.033	0.026
Dynamic graph	0.453	0.011	0.024

The endpoint pair treats one user in its first stage. The five-stage schedule starts at 2%. Intermediate stages can help with curvature. The two-stage strategy wins in other settings.

* Source value rounded to three decimals.

Cluster Assignment and Pooling Work Together

Two-hop interference / 11,586 users, 568,309 observed edges, 100 simulated rollouts



Number of pooled stages ending at 50% traffic

One stage vs two stages

Assignment	50%	25% + 50%
Unit	1.963	0.374
Cluster	0.904	0.096

With both changes, MSE falls to 0.096. The fifth stage yields little additional information in this simulation.

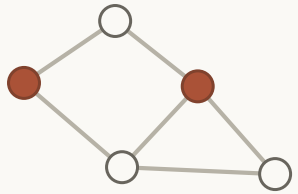
What Each Study Needs From the Experiment

Study	Required information	Use in the experiment
OCD	A known graph and a cluster partition	Before randomization, choose how to correlate assignments across clusters.
COSTA	A network–time exposure model	Coordinate assignment across units and blocks. Check which target the interval covers.
AMII	Cluster interiors and predictive covariates	Correct full-population predictions with observed interior residuals.
Ramp-Up	Comparable stages with distinct traffic shares	Pool stages to learn exposure responses closer to the global endpoints.

The Decision Behind a Full Rollout

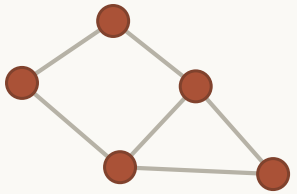
Platforms test features on a subset of users.
At launch, the change reaches those users' contacts as well.

What the pilot observes



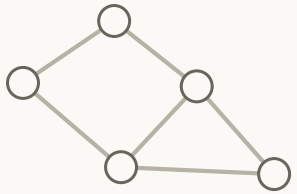
Mixed assignments

What the launch decision compares



New feature for all

versus



Current version for all

A rollout decision needs a network-wide effect and a measure of its uncertainty.
The four studies use the network to choose assignments and learn from the exposures a pilot produces.

From a Partial Test to a Full Rollout

Design

Which units should receive treatment together?

OCD correlates assignments across connected clusters to reduce spillover bias. COSTA adds time and carryover, distinguishing interval coverage for the design mean from coverage for the sustained rollout effect.

Estimation

Which observed exposures tell us about a full release?

Under one-hop spillover, interior units reveal global-regime outcomes. AMII uses those labels to correct population predictions. Ramp-Up pools stages at distinct traffic shares to learn the exposure response.

A sharing pilot observes a mixed network. Its design or observed exposures must support comparing next-week activity when everyone has the feature and when no one does.