LinkedIn

A/B Testing for Recommender Systems in a Two-sided Marketplace

NeurIPS 2021

Preetam Nandy

Divya Venugopalan

Chun Lo

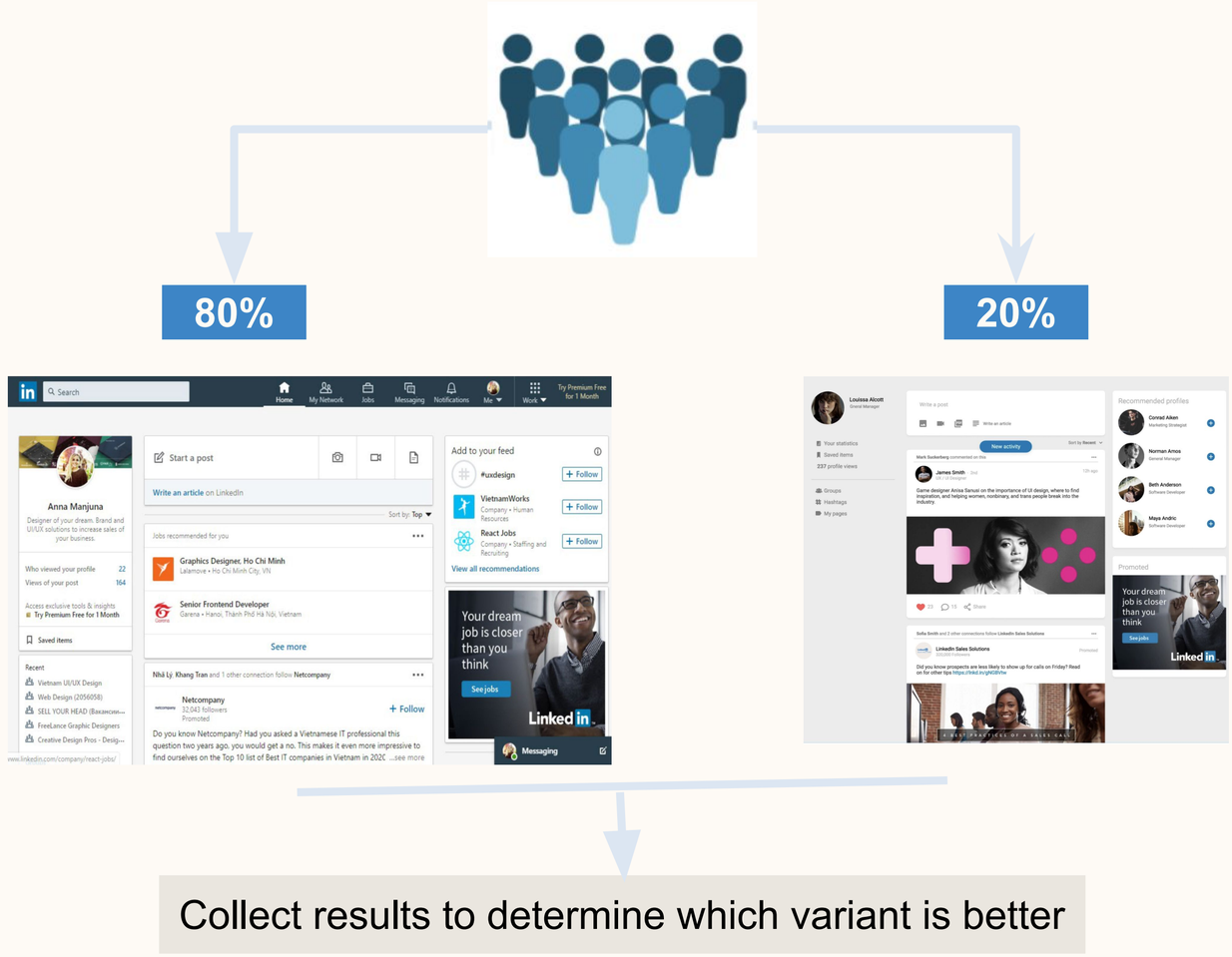
Shaunak Chatterjee

A/B Testing in Recommender system

- Recommender systems tries to serve users with an ordered list of items according to the underlying context and users' preference.

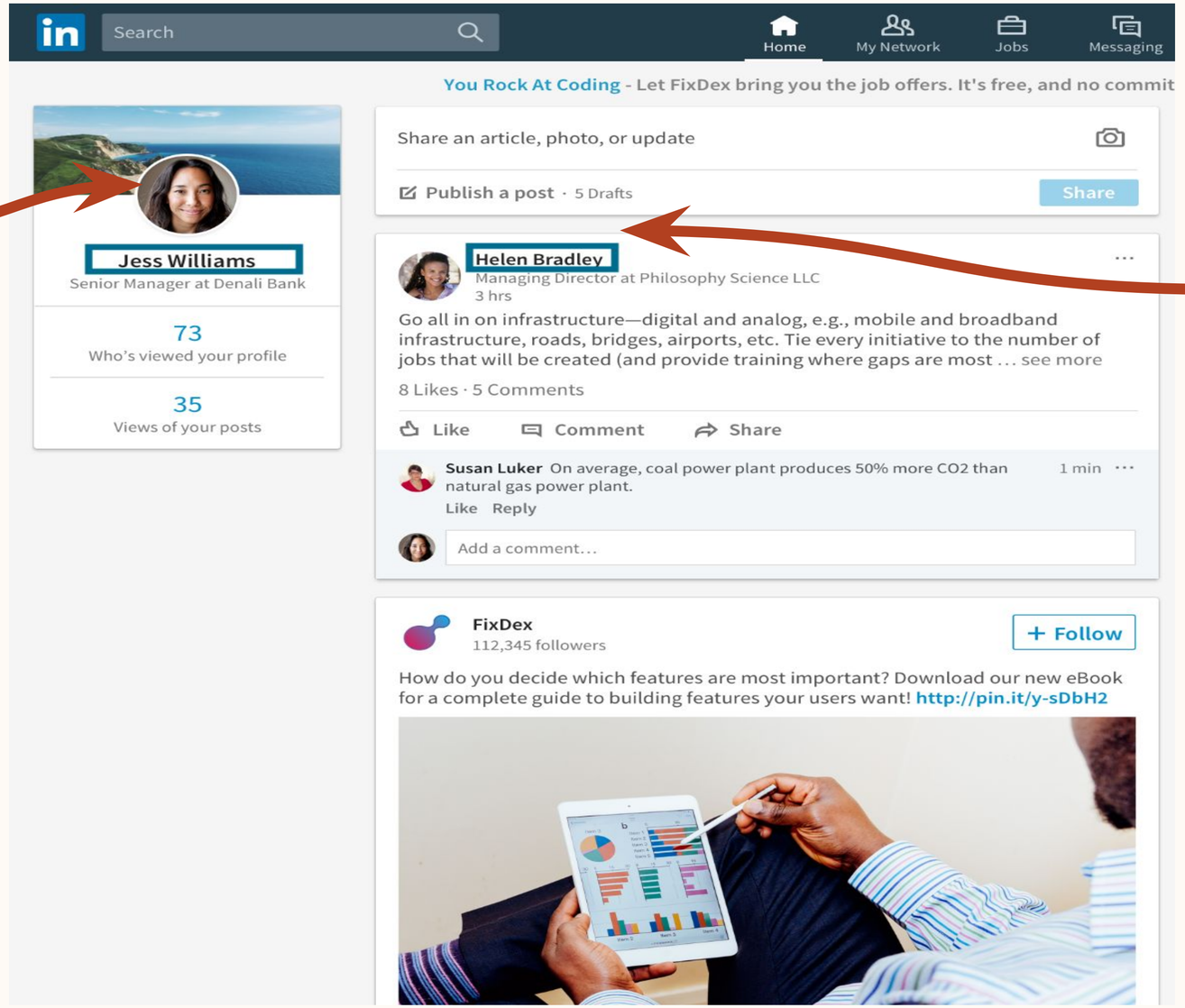
- **A/B Testing:** Randomized experiments with two variants A and B for evaluating model performance with respect to certain metrics measuring users' engagement.

- **Example: Feed ranking change**
  Suppose we believe users want to see more visual content and have made some changes to reflect the same. The best way to determine if this new ranking (relative to the old ranking) is driving more engagement is to run an A/B test.



Two-sided Marketplace (Newsfeed)

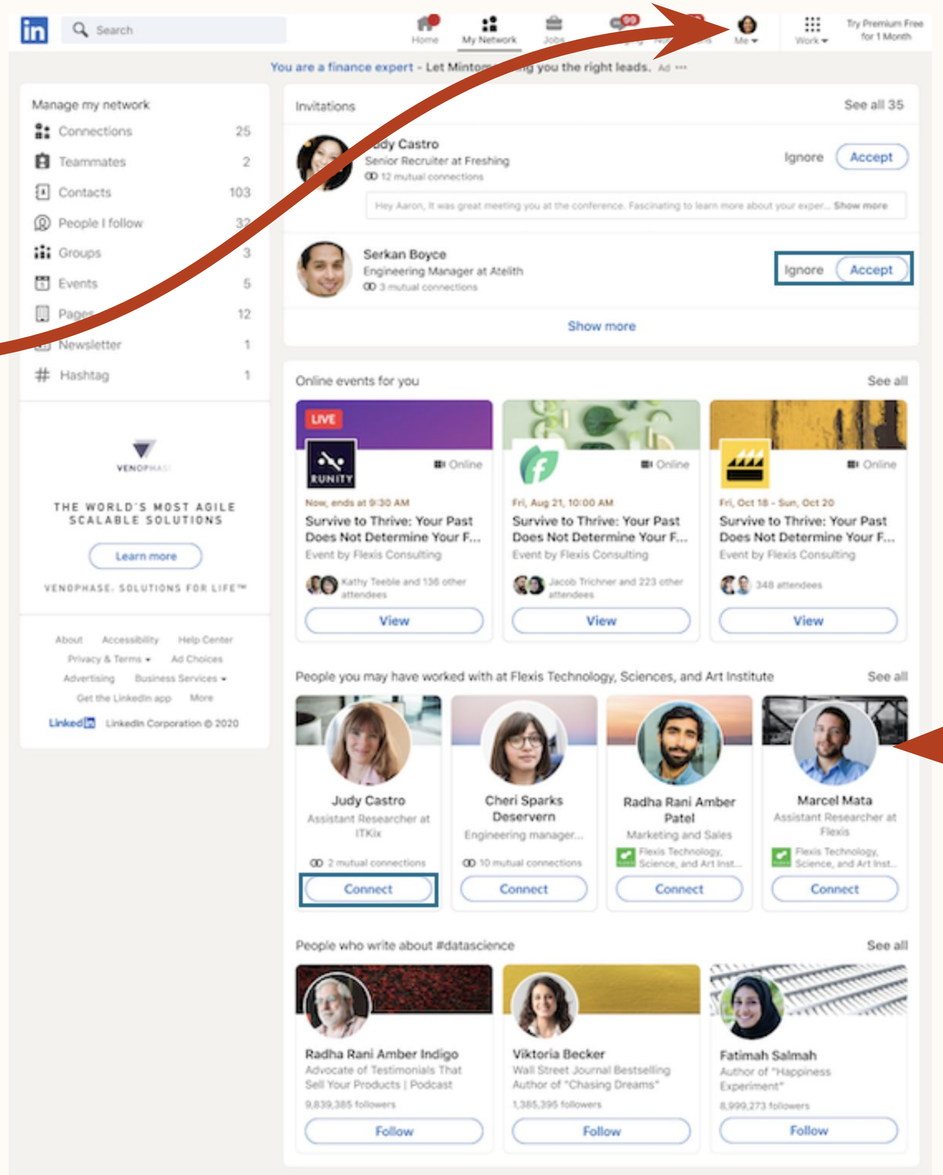
Content Viewer



Content Creator

Two-sided Marketplace (Connection Recommendation)

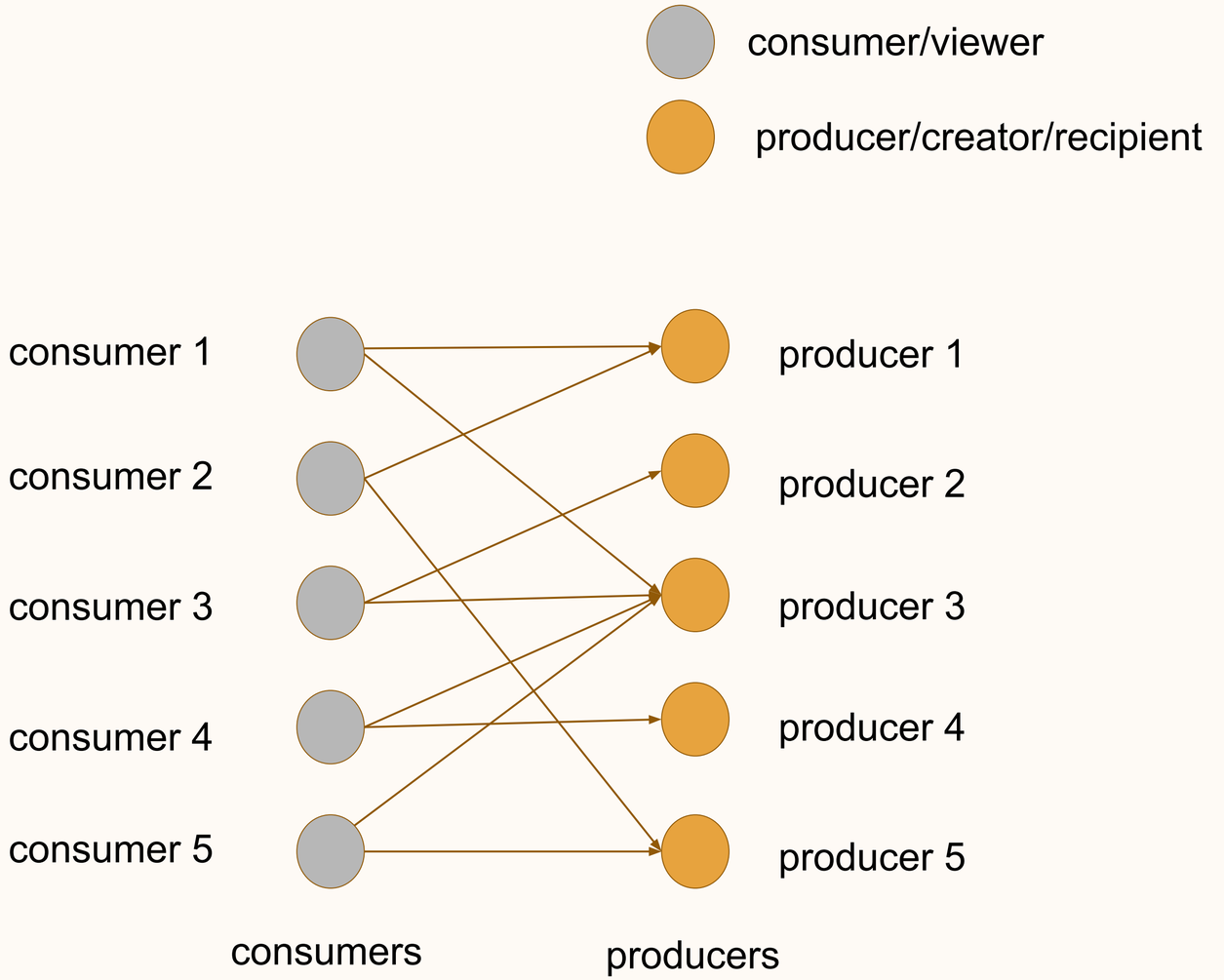
Connection Recommendation Viewer



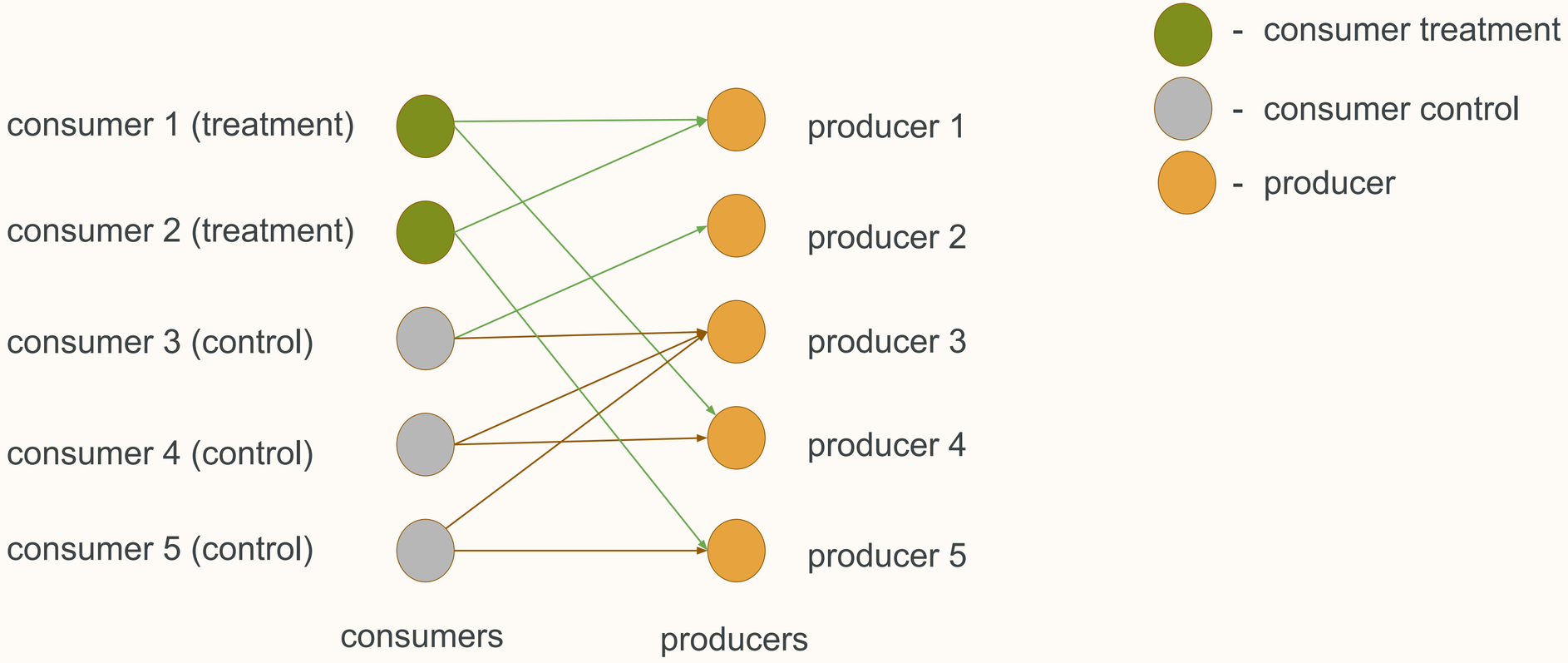
Potential Recipients of Invitations Sent by the Viewer

Graphical Representation

- Consumer 1’s recommendation consists of items from Producers 1 and 3.
- Consumer 2’s recommendation consists of items from Producers 1 and 5
- ...

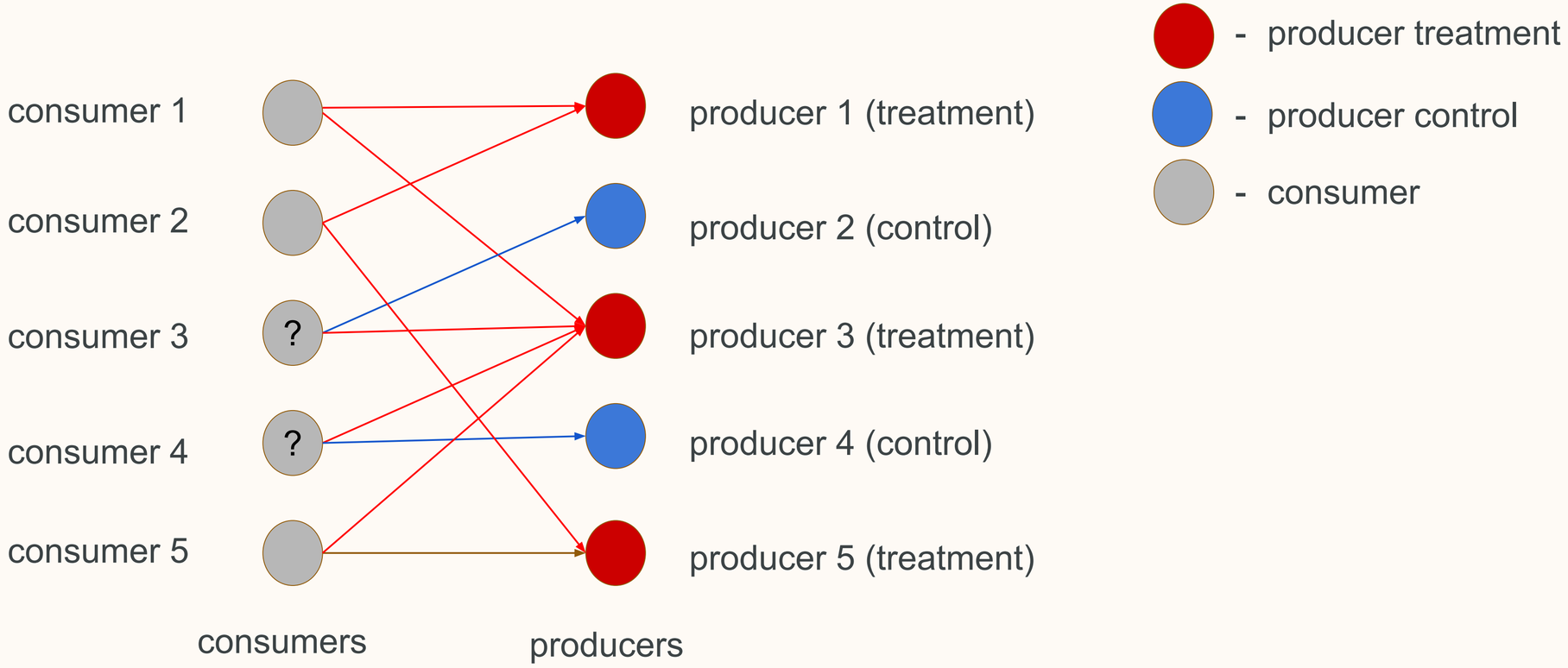


Consumer-side A/B testing in a Two-sided Marketplace



- Generate Consumers 3, 4 and 5’s recommendations using Control Model A
- Generate Consumers 1 and 2’s recommendations using Treatment Model B

Producer-side A/B testing in a Two-sided Marketplace



- Consumers 1, 2 and 5's recommendation can be based on Treatment Model B to make sure that producers 1, 3 and 5 receive treatment experience.
- How to define Consumer 3 and 4's recommendation? There is a conflict because they have producers in control as well as producers in treatment.

Counterfactual Rankings

Producers in the control group: P1, P3, P5, P7
Producers in the treatment group: P2, P4, P6, P8

Ranking based on Control Model A

P1	P2	P3	P4	P5	P6	P7	P8

Ranking based on Treatment Model B

P1	P3	P4	P2	P8	P7	P5	P6

(1) P1, P3, P5, P7 should be placed in positions 1, 3, 5 and 7 (according to the control model ranking)
(2) P2, P4, P6, P8 should be placed in positions 4, 3, 8 and 5 (according to the treatment model ranking)

(1) and (2) cannot be achieved simultaneously because
both P3 and P4 are demanding position 3 and
both P5 and P8 are demanding position 5.

Unifying Counterfactual Rankings (UniCoRn)

- P1, P3, P5, P7 should be placed in position 1, 3, 5 and 7 (according to the control model ranking)
- P2, P4, P6, P8 should be placed in position 4, 3, 8 and 5 (according to the treatment model ranking)

“Ideal but unrealizable” ranking

P1		P3, P4	P2	P5, P8		P7	P6

UniCoRn (breaking ties randomly)

P1	P4	P3	P2	P8	P5	P7	P6

- Theoretically optimal with respect to the mean squared error (when compared with the ideal ranking)
- Provided bias and variance bounds
- Can handle multiple treatments
- Provided cost-efficient versions (next slide)

Cost and Accuracy Trade-off (UniCoRn(α))

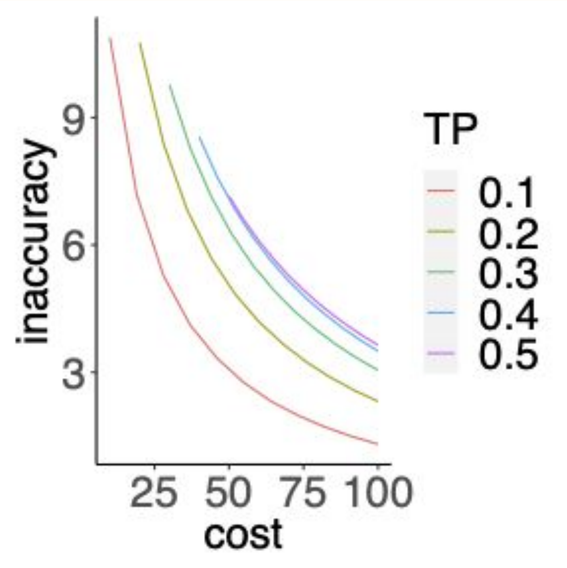
Cost: computational load of generating the counter factual rankings involving scoring the items based on the multiple treatment and control models.
Alpha: tuning parameter which lets you control the tradeoff between cost and accuracy based on your application

Cost-efficient variants of UniCoRn:

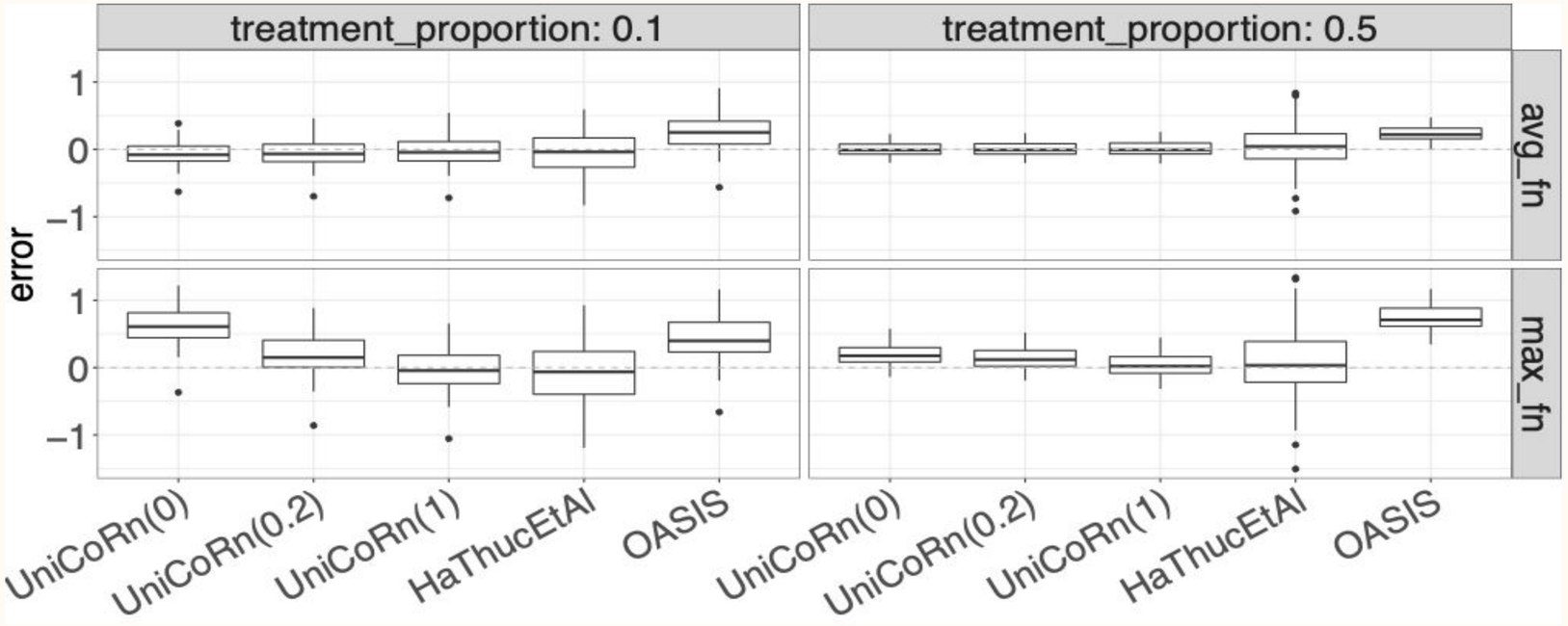
- Step one: We rank, all items across the treatments and control groups according to the control model.
- Step two: We randomly select a subset of items from the control group and fix their ranks/positions.
- Step three: We run UniCoRn exactly as described earlier with one key difference. All items from step 2 are excluded.

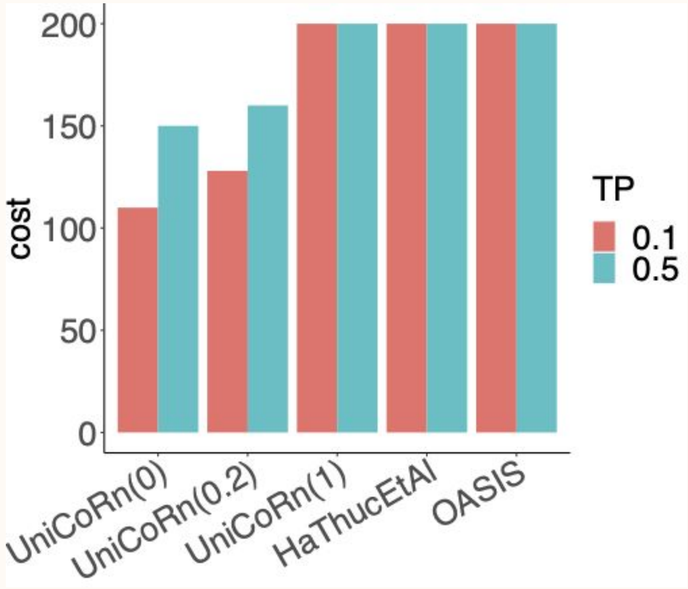
- α denotes the proportion of items scored/ranked in step three using UniCoRn(α)

- UniCoRn(1) is the same as the original UniCoRn (most expensive) and UniCoRn(0) is the least expensive version.



Empirical Evaluation





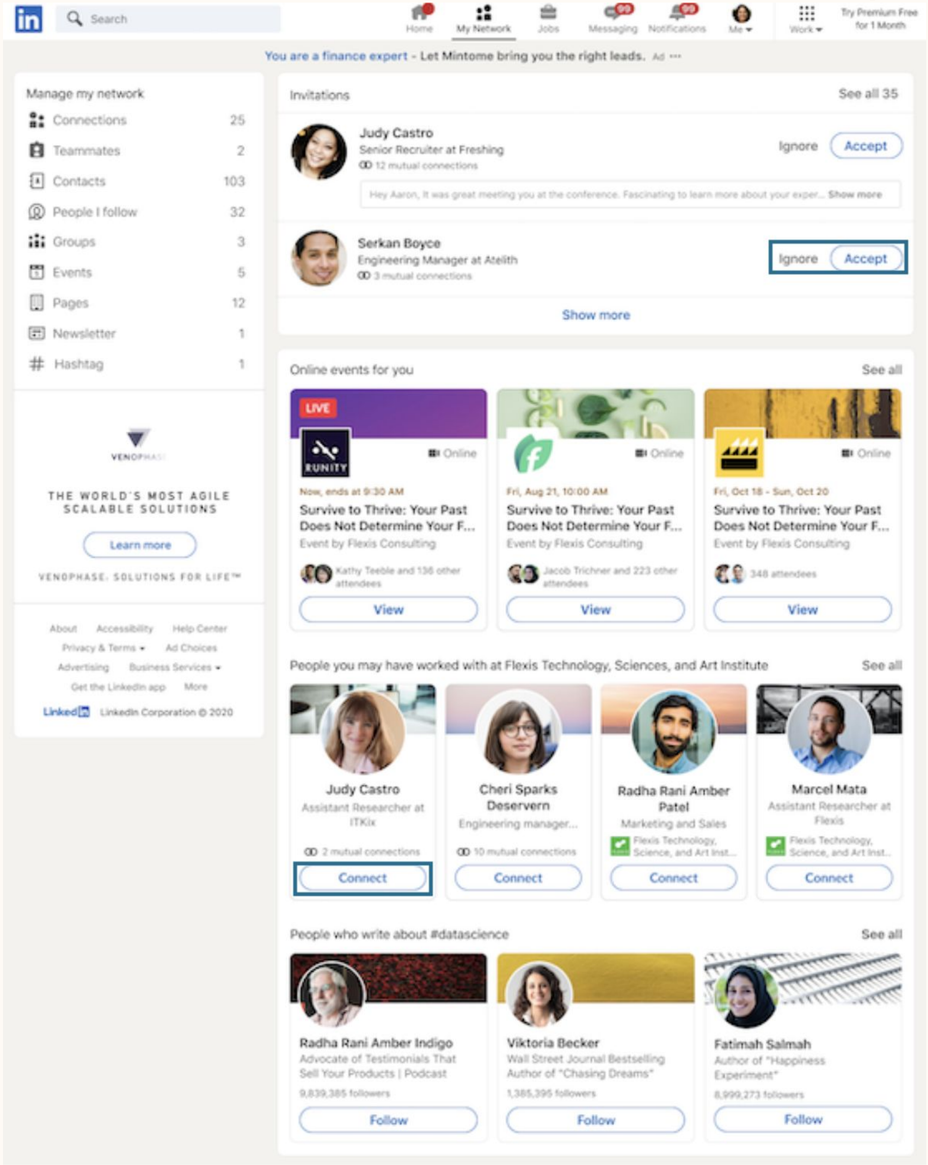
TP: Treatment proportion
avg_fn and max_fn: Mapping of rankings to the response
Error in estimating the treatment effect

Observations:

1. UniCoRn(1) compares favourably to state of the art methods
2. UniCoRn(α) is more sensitive to the choice of α at TP = 0.1
3. Rank to response max_fn is more sensitivity to the choice of α
   - We should choose an alpha based on the treatment proportion and the kind of function between ranking and response in our application
4. Unicorn (0.2 or 0) provide significant cost reduction if we are willing to sacrifice some accuracy

Large-scale Application (LinkedIn Connection Recommendation)

- We implemented UniCoRn in an online edge recommender system that serves tens of millions of members, and billions of edge recommendations daily.
- Viewers as consumers and recommendations/viewees as producers
- In the experiment we have 50-50 treatment and control proportion on recipient/producer side
- We chose α = 0 (i.e., UniCoRn(0)) to minimize the online scoring latency increase



Large-scale Application (LinkedIn Connection Recommendation)

Candidate generation experiment: Popular candidate generation heuristic is number of shared edges. We tested a variant based on a normalized version of shared edges.

Ranking model experiment: The ranking stage scores all candidates based on the model assignment of the viewers. Ranking models may be composite models optimizing for viewer and/or viewee side outcomes. One such treatment model optimized for viewee side retention.

Statistically significant results:

Metrics	Delta % (candidate generation)	Delta % (ranking model)
Weekly Active Unique users	+0.51%	+0.13%
Sessions	+0.57%	+0.11%

- If underlying network is too dense the cluster based methods often struggle to measure experiments statistically
- With this technique we have been able to overcome the issue of having low power

Key Takeaways

1. **UniCoRn is effective**: its design provides a mechanism to unify multiple counterfactual rankings to facilitate producer side A/B testing

2. **UniCoRn is flexible**:
   a. It offers an explicit parameter to control tradeoff between cost and accuracy making it suitable for large scale real-world applications
   b. It is Agnostic to the underlying network density or structure and makes no assumptions on treatment effect propagation through the network

3. UnicoRn has shown real world impact

Thank you