

ORIE 5355: People, Data, & Systems

Lecture 13: Experimentation complications: peeking and interference

Nikhil Garg

Course webpage: https://orie5355.github.io/Fall_2021/

Announcements

- Today 11/1 guest lecture virtually: Hannah Li on marketplace experimentation
 - 6:15pm – 7:30 pm on Zoom
 - Will be recorded; required to watch it (will be on quiz)
 - Live attendance not required but is appreciated
- HW4 released; due next week
- Quiz 4 next week
- Project details released in next week
 - Project partner form on EdStem



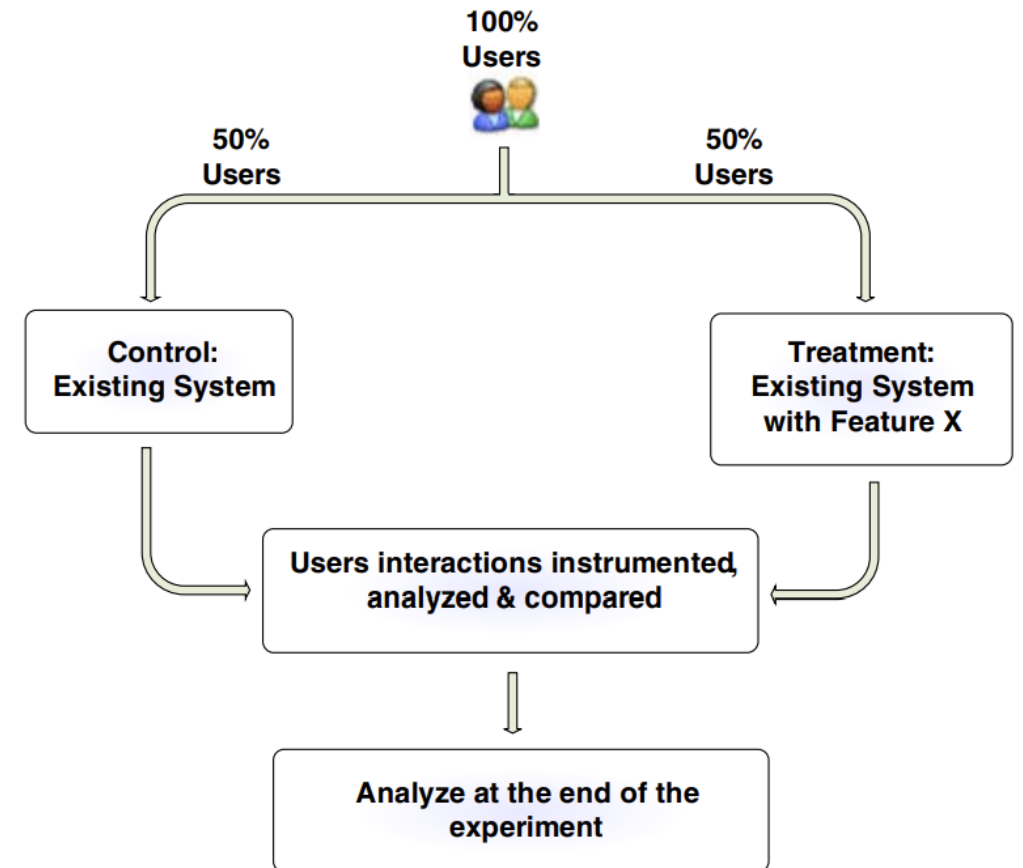
THE INVENTION OF CLINICAL TRIALS

[xkcd: Clinical Trials](#)

Peeking: a common mistake in running A/B tests in online marketplaces

Basics of basic A/B testing

- Have an idea for a system change
- Give $X\%$ of your users the changed system, everyone else the old system
 - Do this until you have N samples**
- Decide the *metric* you care about
- Check if your system improved the metric
- Launch your product if good things happened

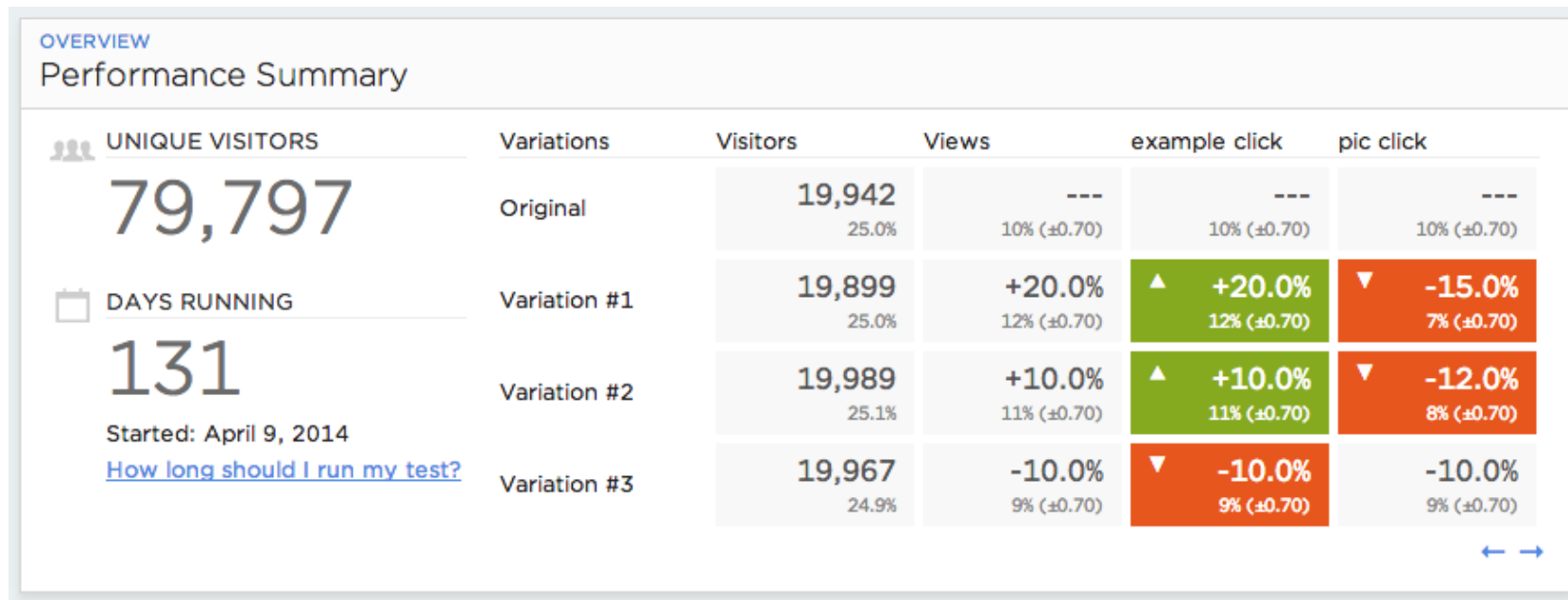


[Source: Controlled experiments on the web: survey and practical guide]

Experiment Dashboards

In modern internet experiments, it's easy to see experimental results *while they are happening*

Sample results dashboard:



[Image credit: Ramesh Johari (Stanford; also Optimizely at time of presentation)]

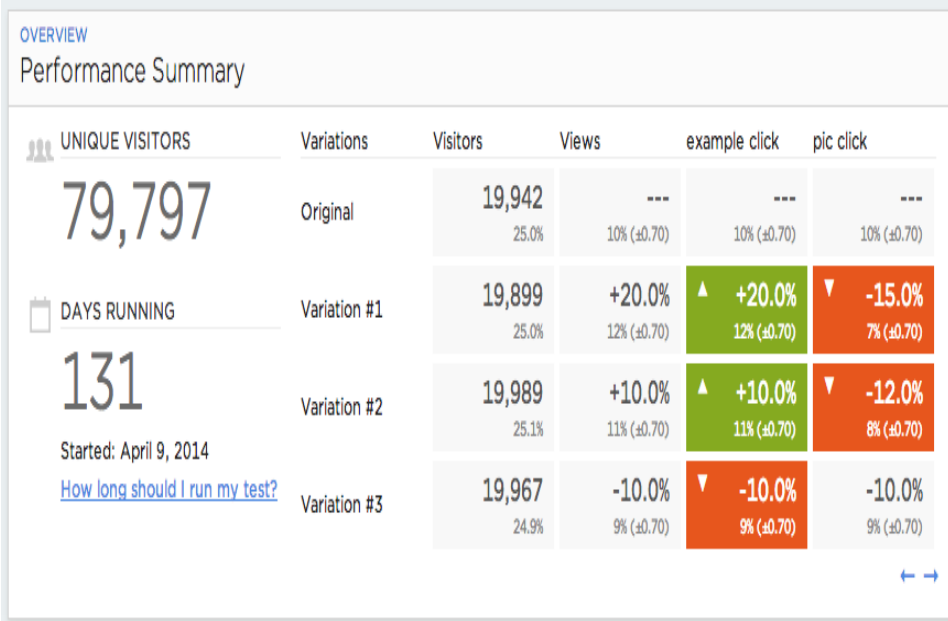
Peeking

In modern online setting, the approach I described above is wasteful

So you continuously monitor (stare at) the results dashboard.

You rely on the dashboard to tell you when your results are significant.

- As soon as results are significant, you end the test and declare victory
- This is called adaptive sample size testing:
 - You adjust the test length in real-time, based on the data coming in.
 - If difference $Y_1 - Y_0$ is *huge*, end the experiment early



OVERVIEW
Performance Summary

UNIQUE VISITORS	Variations	Visitors	Views	example click	pic click
79,797	Original	19,942 25.0%	--- 10% (±0.70)	--- 10% (±0.70)	--- 10% (±0.70)
DAYS RUNNING 131 Started: April 9, 2014 How long should I run my test?	Variation #1	19,899 25.0%	+20.0% 12% (±0.70)	▲ +20.0% 12% (±0.70)	▼ -15.0% 7% (±0.70)
	Variation #2	19,989 25.1%	+10.0% 11% (±0.70)	▲ +10.0% 11% (±0.70)	▼ -12.0% 8% (±0.70)
	Variation #3	19,967 24.9%	-10.0% 9% (±0.70)	▼ -10.0% 9% (±0.70)	-10.0% 9% (±0.70)

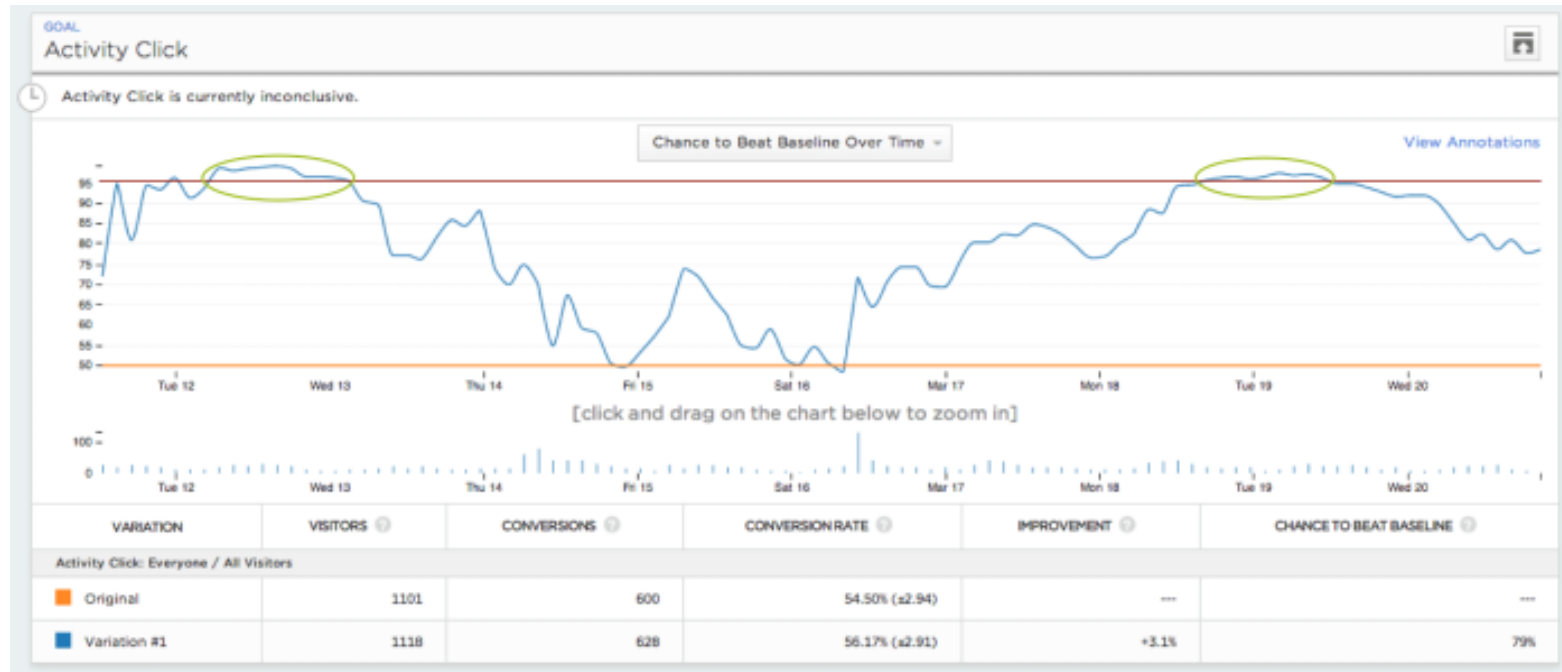
[Slide credit: Ramesh Johari (Stanford; also Optimizely at time of presentation)]

Effect of peeking

- Suppose 100 different individuals run A/A tests (same arm is treatment and control, so you know that $Y_1 - Y_0 = 0$)
- Each continuously monitors the dashboard, and waits for a significant result, i.e., p-value < 5% (up to a maximum of 10,000 visitors).
- *How many find a significant result and stop early?*
 - Remember, $\alpha = 0.05$ means that if there is no true difference ($Y_1 - Y_0 = 0$), then 5% of the time you will falsely declare that $\hat{Y}_1 - \hat{Y}_0 \neq 0$ in a statistically significant way (false positive)
- Answer: **Over HALF!** find a significant result if they peek
- In A/B testing, “peeking” can dramatically inflate false positives.

What went wrong?

A sample run of an A/A test (graph is of p-values over time)



If you wait long enough, there is a high chance of an eventually inconclusive result looking “significant” along the way!

[Slide credit: Ramesh Johari (Stanford; also Optimizely at time of presentation)]

Peeking: what to do about it

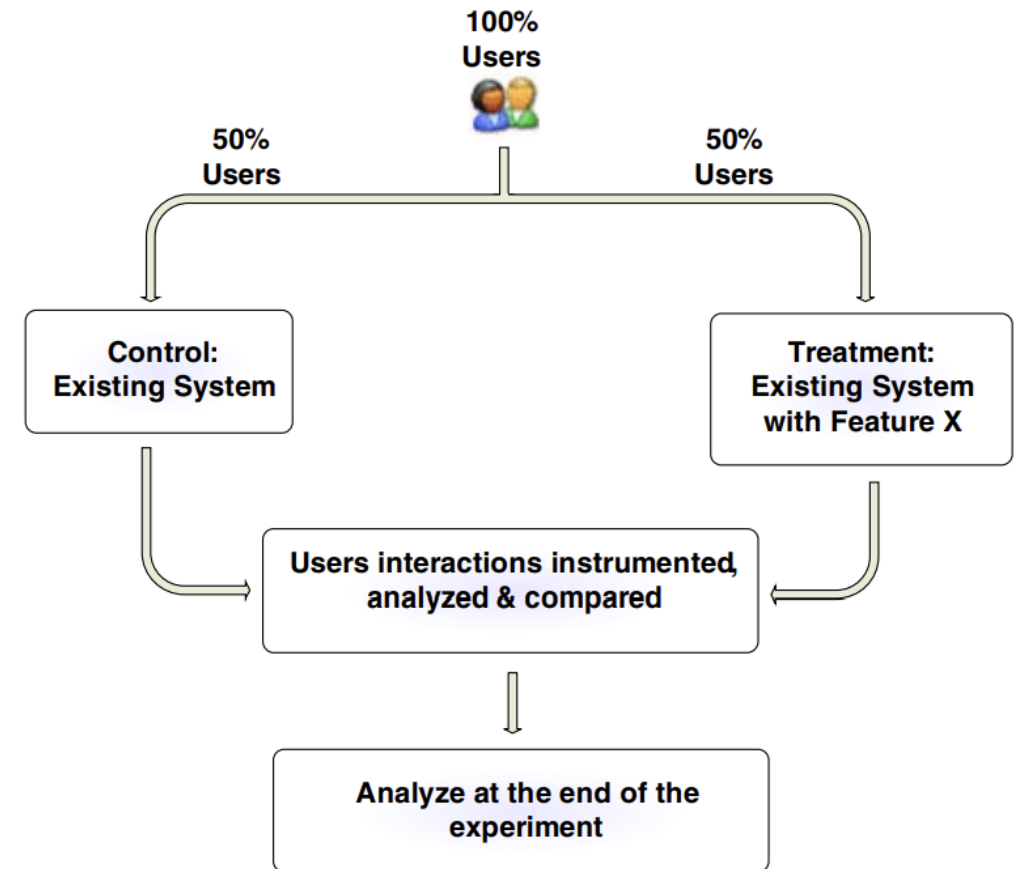
You have two options

- Don't peek: set a sample size N before the experiment starts, and don't end early no matter how large the effect is
 - Easy to do the statistics; no danger of inflating false positives
 - Could be **wasteful**: what if the effect is clearly huge?
 - Even medical trials have a procedure to end early if a drug is clearly fantastic
- Peek, but do **fancy statistics** to make sure your p-values are valid
 - This is the approach Optimizely implemented on their dashboards
 - If you're at a big company with an established experimental culture, they (hopefully) have a dashboard that does this

Interference in experimentation

Basics of basic A/B testing

- Have an idea for a system change
- Give X% of your users the changed system, everyone else the old system
 - Independently assign each user to treatment or control**
- Decide the *metric* you care about
- Check if your system improved the metric
- Launch your product if good things happened



[Source: Controlled experiments on the web: survey and practical guide]

Interference motivation

- Experimentation goal: ultimately, we want to measure – “what will happen if I launch this product for *everyone*, compared to if *everyone* gets the control”

“Global treatment effect”

- With A/B testing so far, we give some people the treatment and some people the control, and then calculate the treatment effect $Y_1 - Y_0$
- We implicitly assumed: if we give *some* people the treatment, individually that is equivalent to giving *everyone* the treatment:
 - Effect of giving someone a coupon doesn't depend on if their friend got a coupon
- This assumption is often violated in people-centric systems!
 - (Social) network effects, capacity constraints
- Different *units* (people) *interfere* with one another

Interference in experimentation

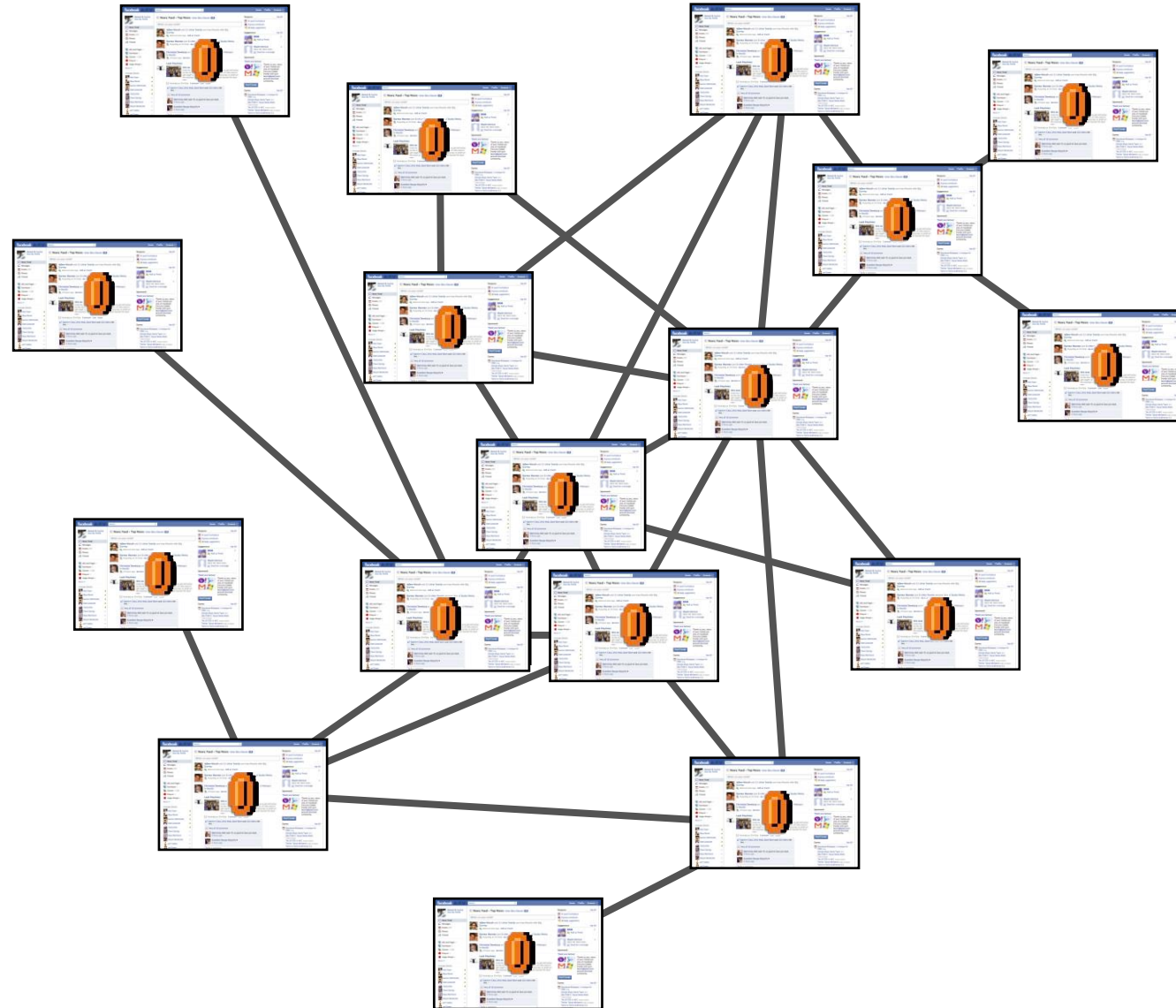
A/B testing in (social) networks

A/B testing under network effects



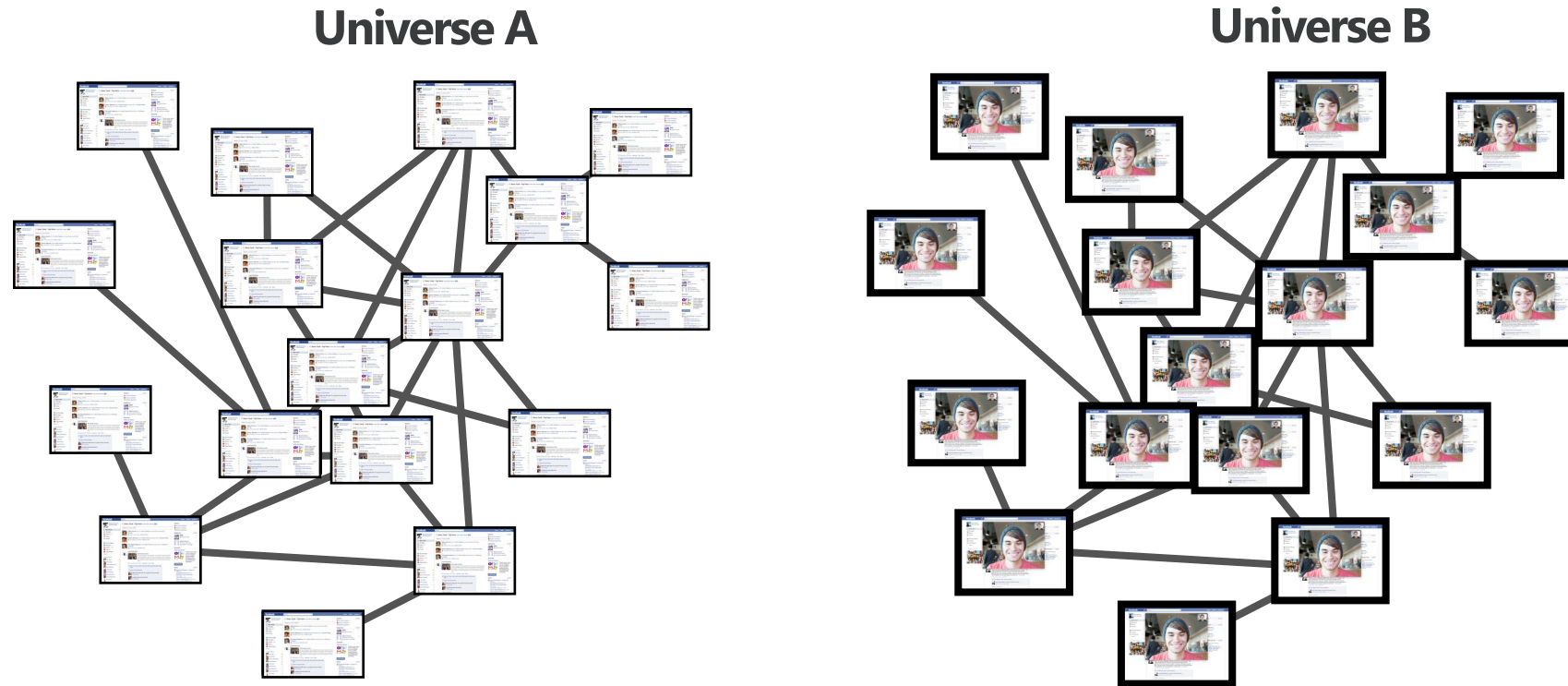
Slide credit: Johan Ugander, Stanford

A/B testing under network effects



Slide credit: Johan Ugander, Stanford

Causal inference & network effects

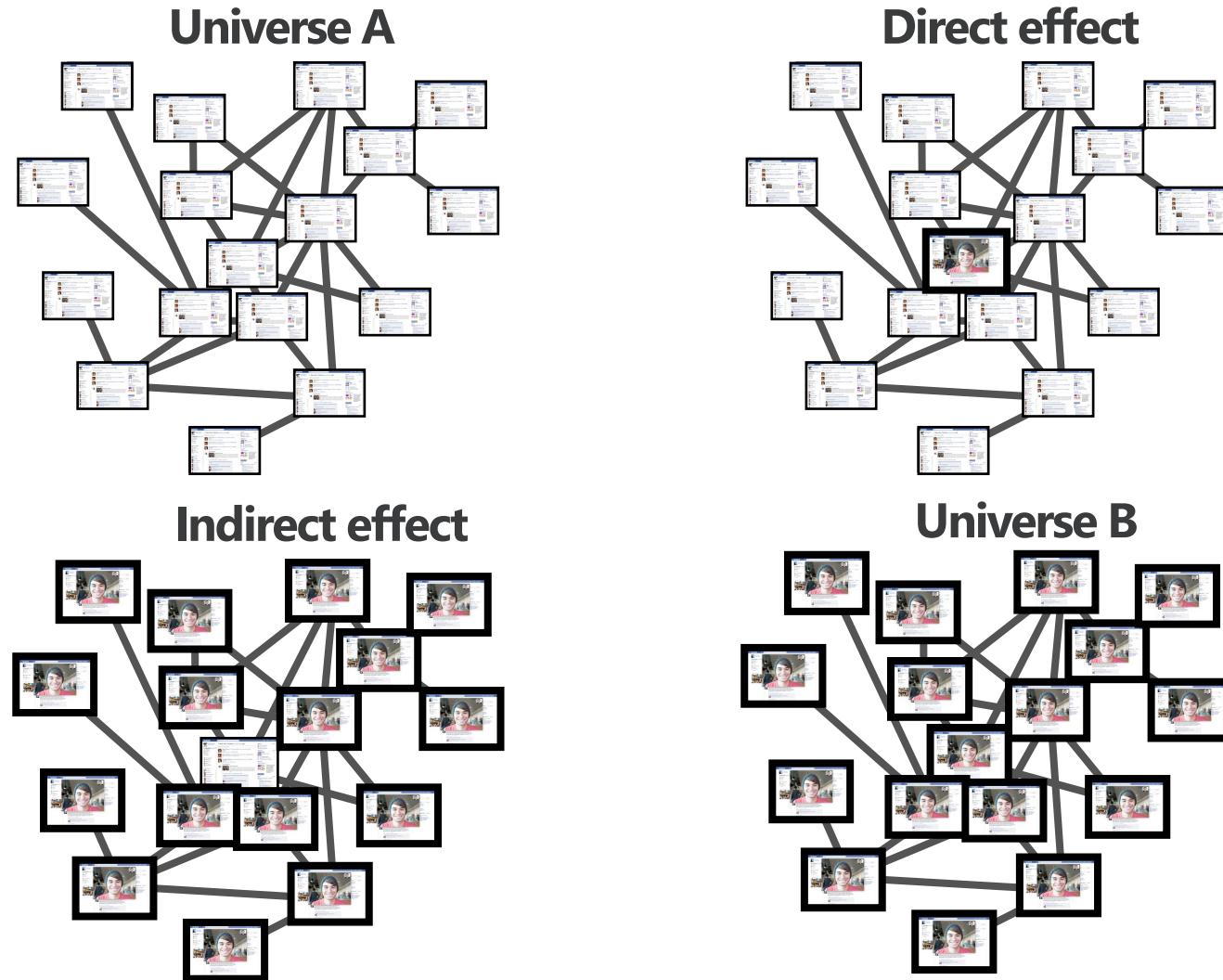


Fundamental problem: want to compare (average treatment effect, ATE), but can't observe network in both states at once.

- J Ugander, B Karrer, L Backstrom, J Kleinberg (2013) "Graph Cluster Randomization: Network Exposure to Multiple Universes," KDD.
- D Eckles, B Karrer, J Ugander (2014) "Design and analysis of experiments in networks: Reducing bias from interference," arXiv.
- S Athey, D Eckles, G Imbens (2015) "Exact P-values for Network Interference," arXiv.

Slide credit: Johan Ugander, Stanford

Direct vs. indirect effects

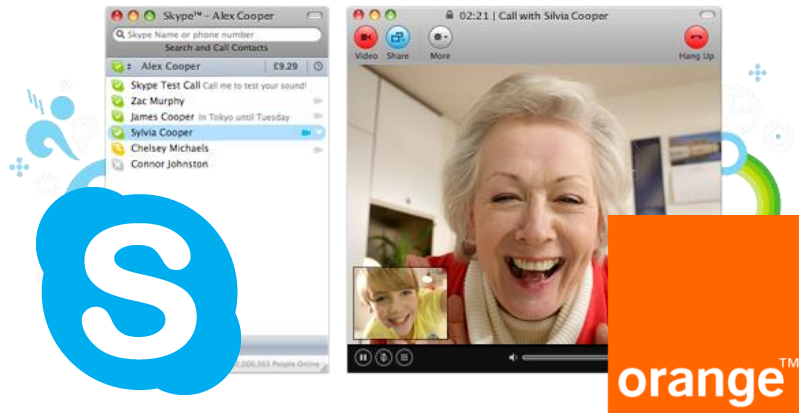


- P Aronow, C Samii (2013) "Estimating average causal effects under interference between units," arXiv.
- C Manski (2013) "Identification of treatment response with social interactions," The Econometrics Journal.

Slide credit: Johan Ugander, Stanford

Experiments with interference

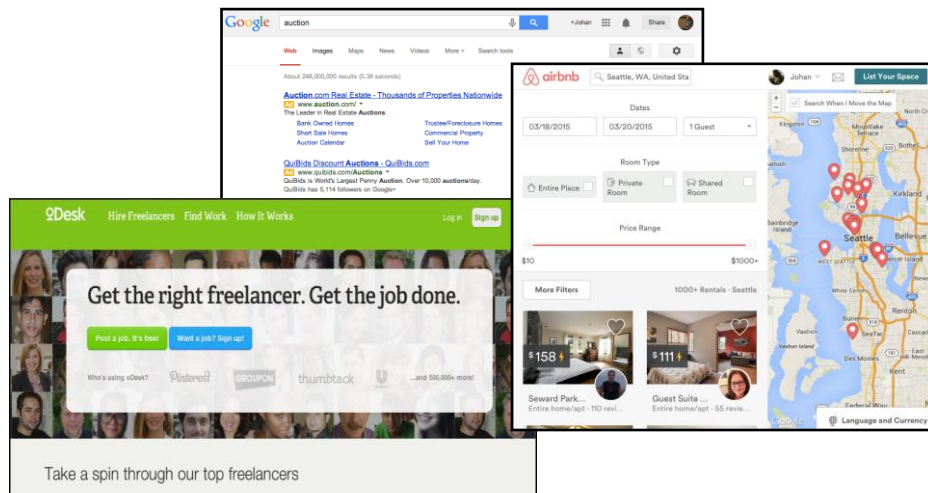
Chat/communication services



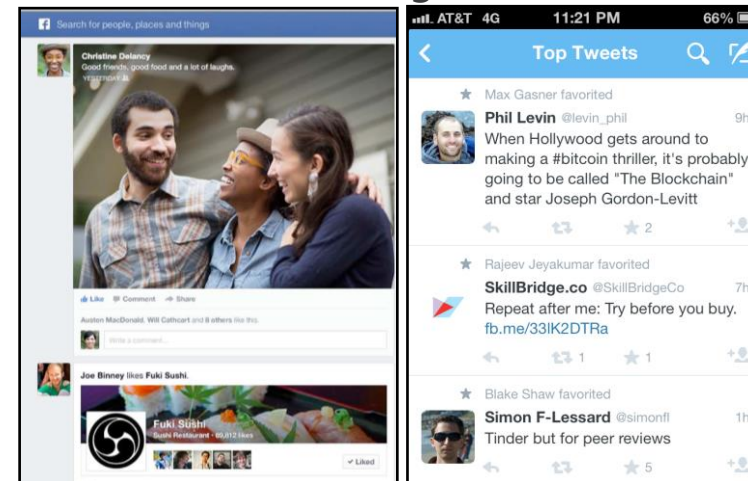
Social product design



Market Mechanisms (ads, labor, etc)

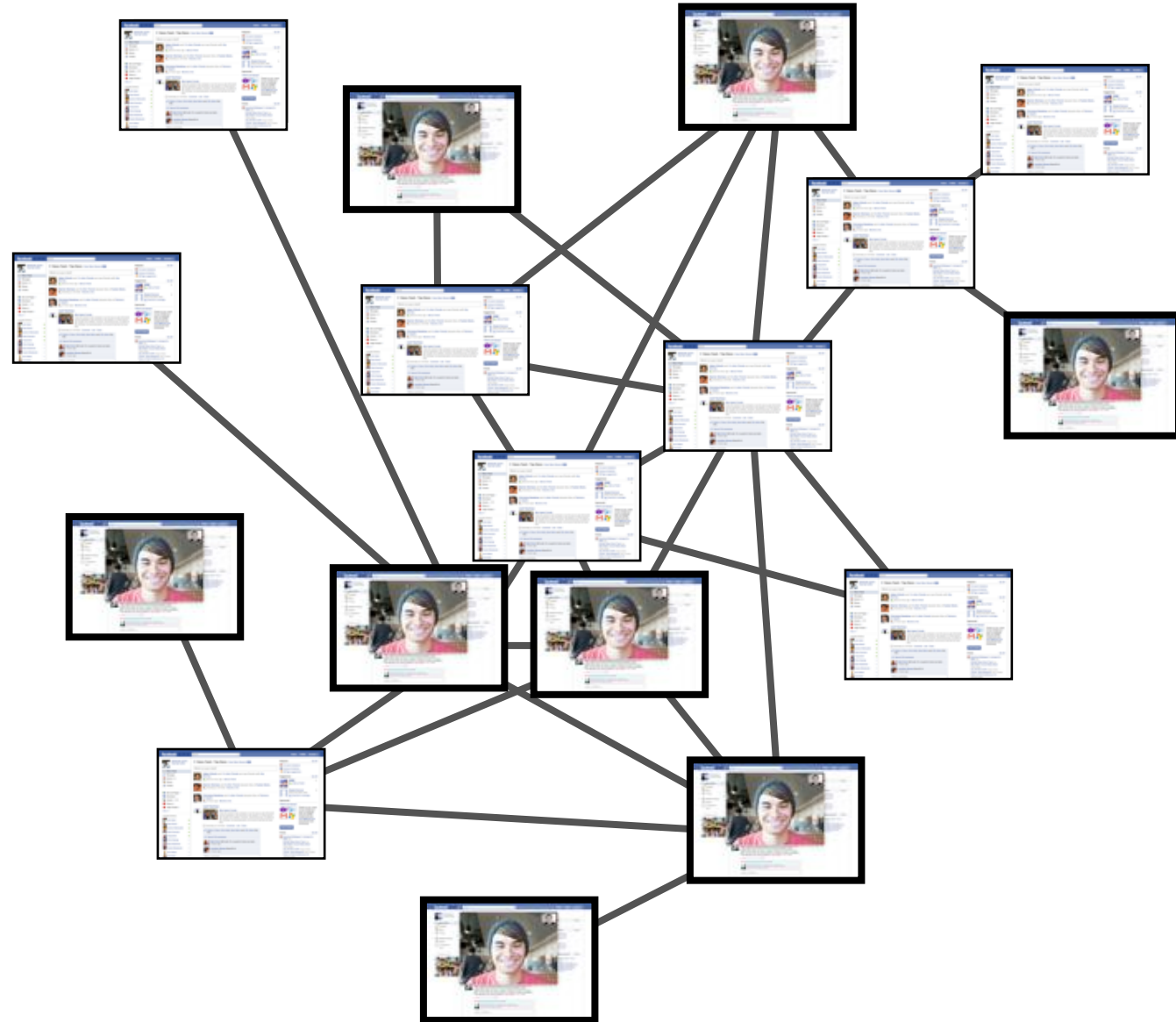


Content ranking models



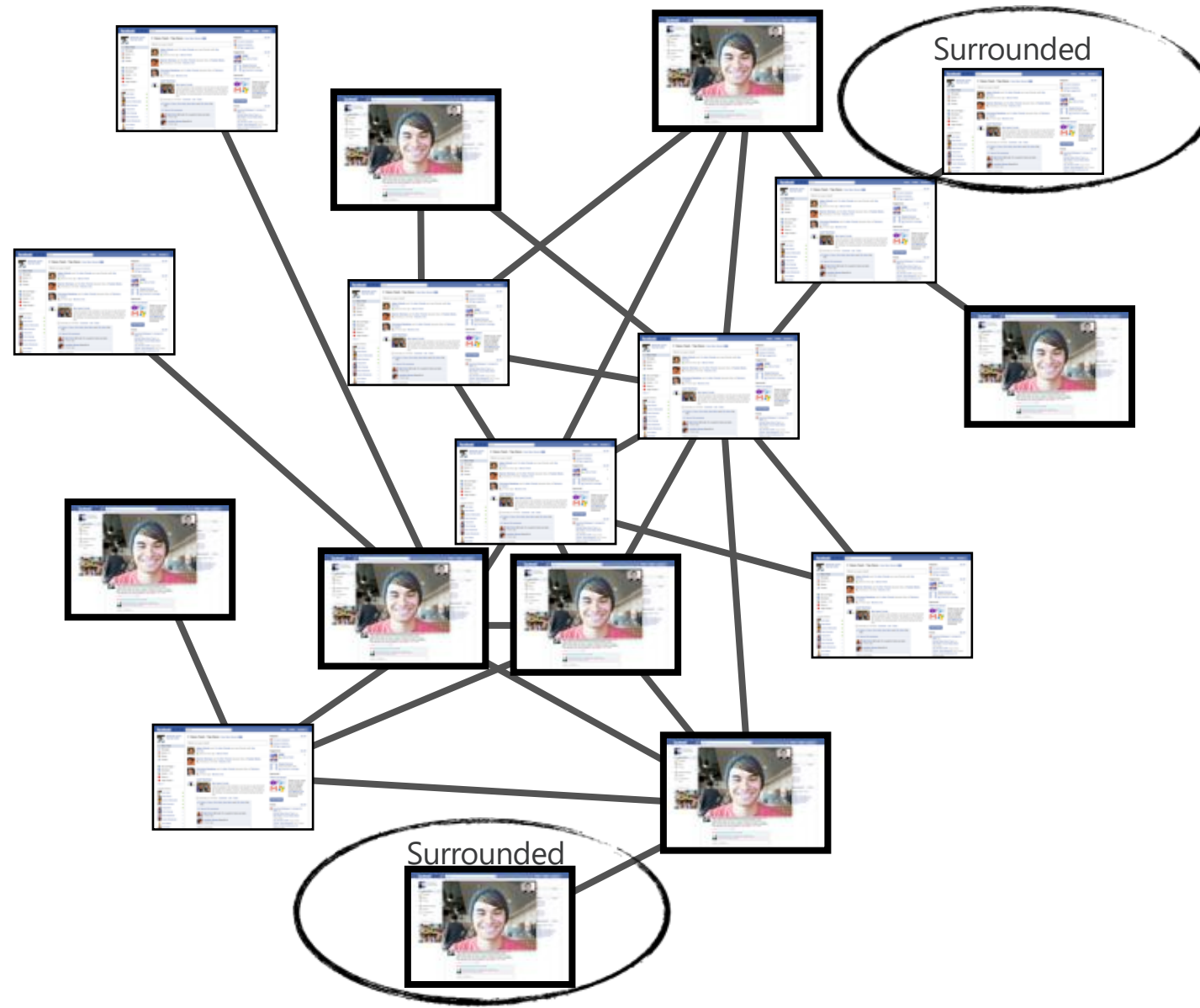
Slide credit:
Johan
Ugander,
Stanford

Design & Analysis



Slide credit:
Johan
Ugander,
Stanford

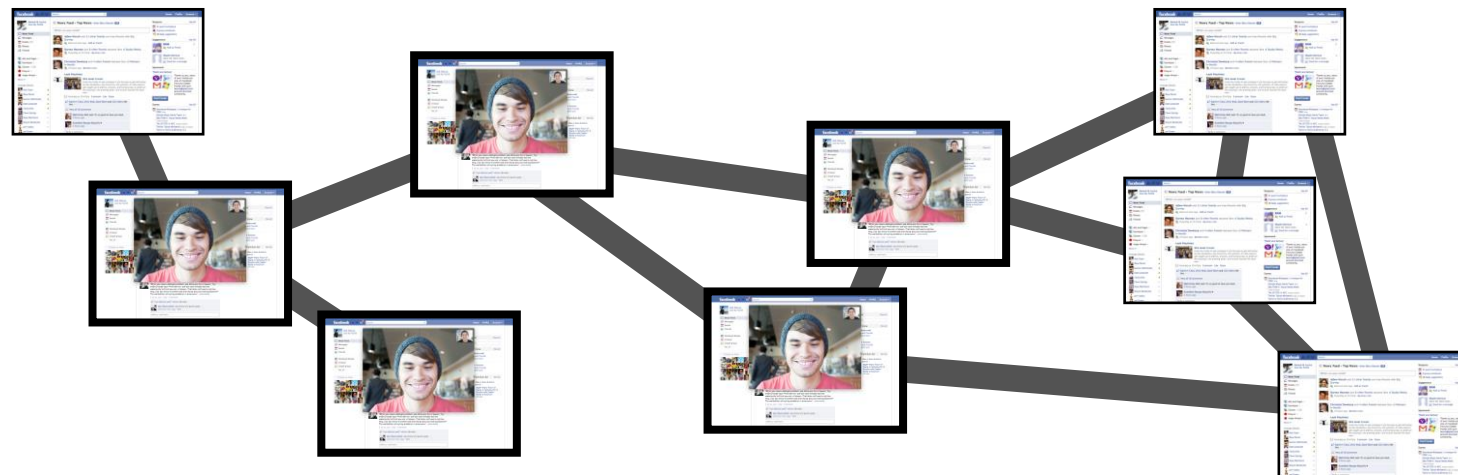
Design & Analysis



Slide credit:
Johan
Ugander,
Stanford

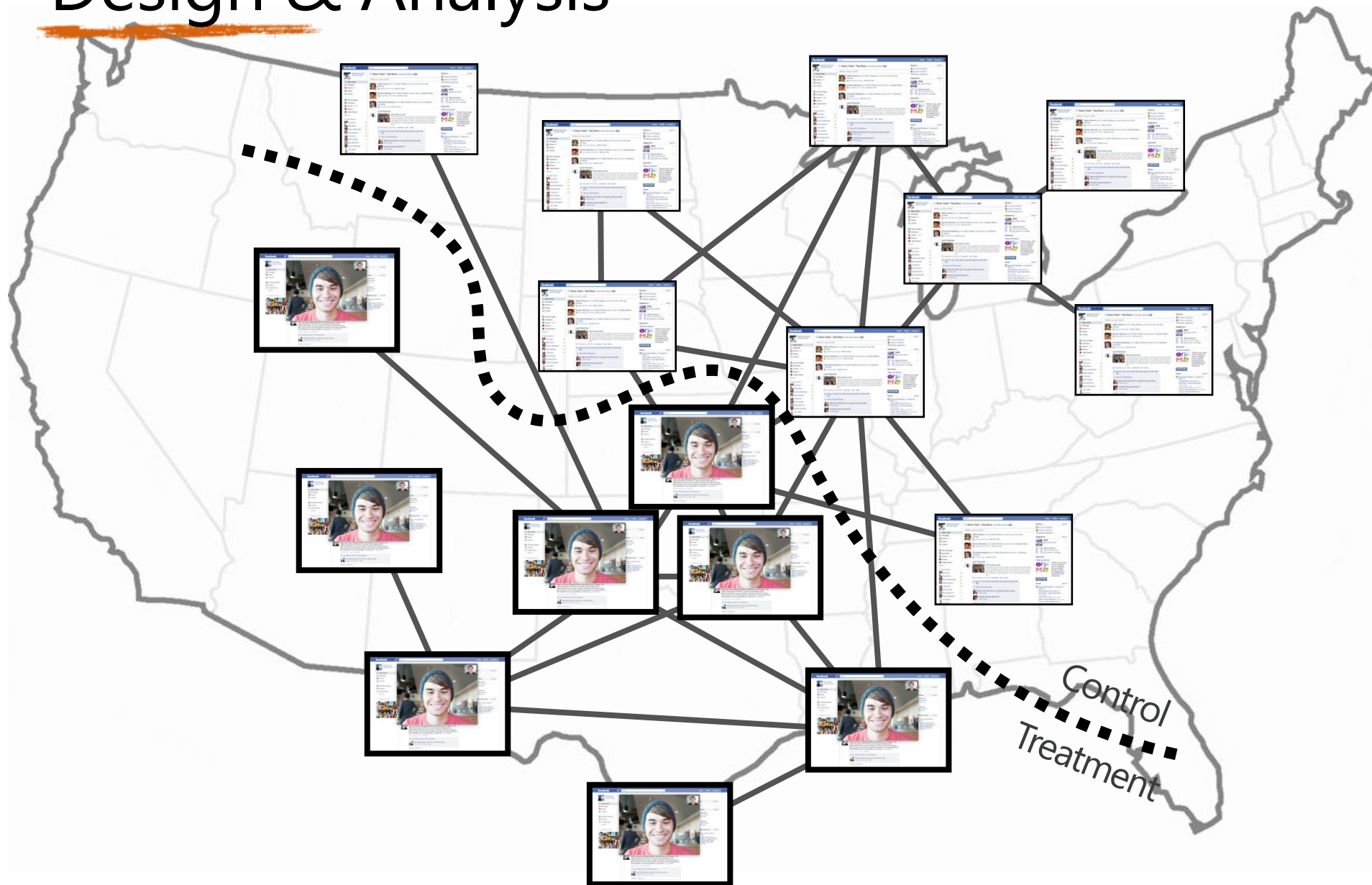
Analysis: "network exposure"

- Two treatment conditions: treatment/control.
- When are people network exposed to their treatment condition?
- Neighborhood exposure to treatment/control:
 - Full neighborhood exposure: you and all neighbors
 - Fractional neighborhood exposure: you and $\geq q\%$ neighbors
- Many more notions are plausible



Slide credit:
Johan
Ugander,
Stanford

Design & Analysis



Slide credit:
Johan
Ugander,
Stanford

New Zealand assignment



Image credit:
Johan
Ugander,
Stanford

Idea: Pick a region of the graph that is densely connected with each other, but less connected with other parts of the graph. Put treatment in region, control everywhere else

“Graph cluster” randomization

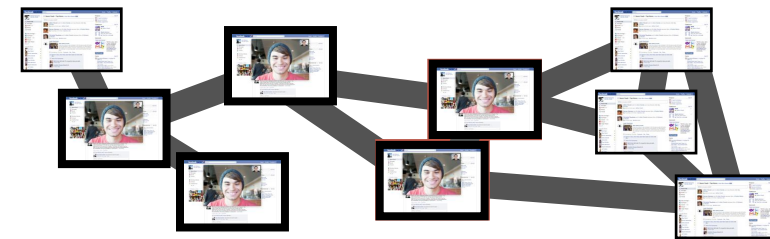
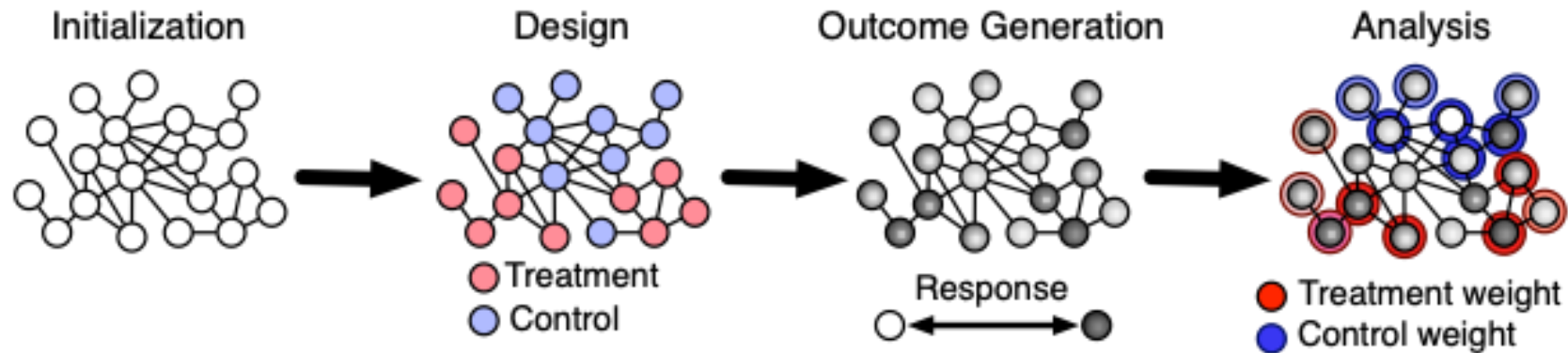


Image credit:
Johan
Ugander,
Stanford

Idea: Algorithmically find many such regions, and then assign half of them treatment, and the other half control

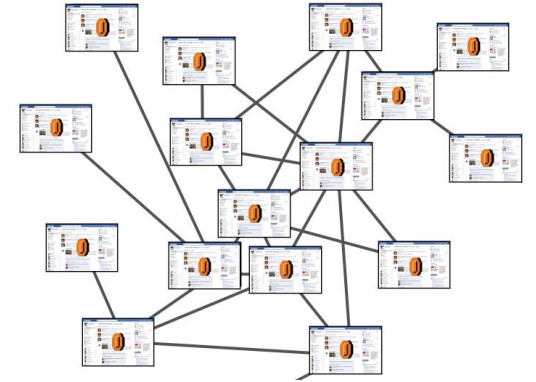
Network Experimentation summary



- Initialization: An empirical graph or graph model
- Design: Graph cluster randomization
- Outcome generation: Observe behavior (or observe model)
- Analysis: Discerning effective treatment

General lesson: “unit” of randomization

- If you randomize at the “individual” level (each individual is its own “unit”), then treatment and control units can interfere with each other
- Solution is often to change the *unit of randomization*: randomize “clusters” instead of individuals
 - Hope: clusters are *close to independent*
 - If independent, experiment is *unbiased*
- Downside: Experiment “variance” goes down with sample size of experiment
 - Before: Sample size is *millions* (of users)
 - Now: Sample size is *hundreds* (of clusters)
- Same bias-variance trade-off we’ve seen before!



Interference in marketplaces

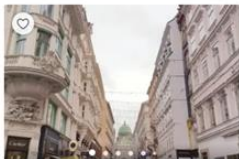




- Interference between treatment and control also arises in marketplaces
- In social networks: Interference because use case is *social* – me getting video messaging doesn't matter if none of my friends get it
- In markets, interference rises from *competition and capacity constraints*
- If I make half the products cheaper, customers will *increase* their purchases of the cheaper items...why?
 - *Decrease* their purchases of the more expensive items (**cannibalization**)
 - Go from not purchasing at all, to buying the now cheaper item (**new customer**)
- *Not* a good representation of what would happen if I make *all* my products cheaper
 - Cannibalization effect would not occur; only attraction of new customers*
- **Tonight and next time: experimentation in marketplaces under interference**

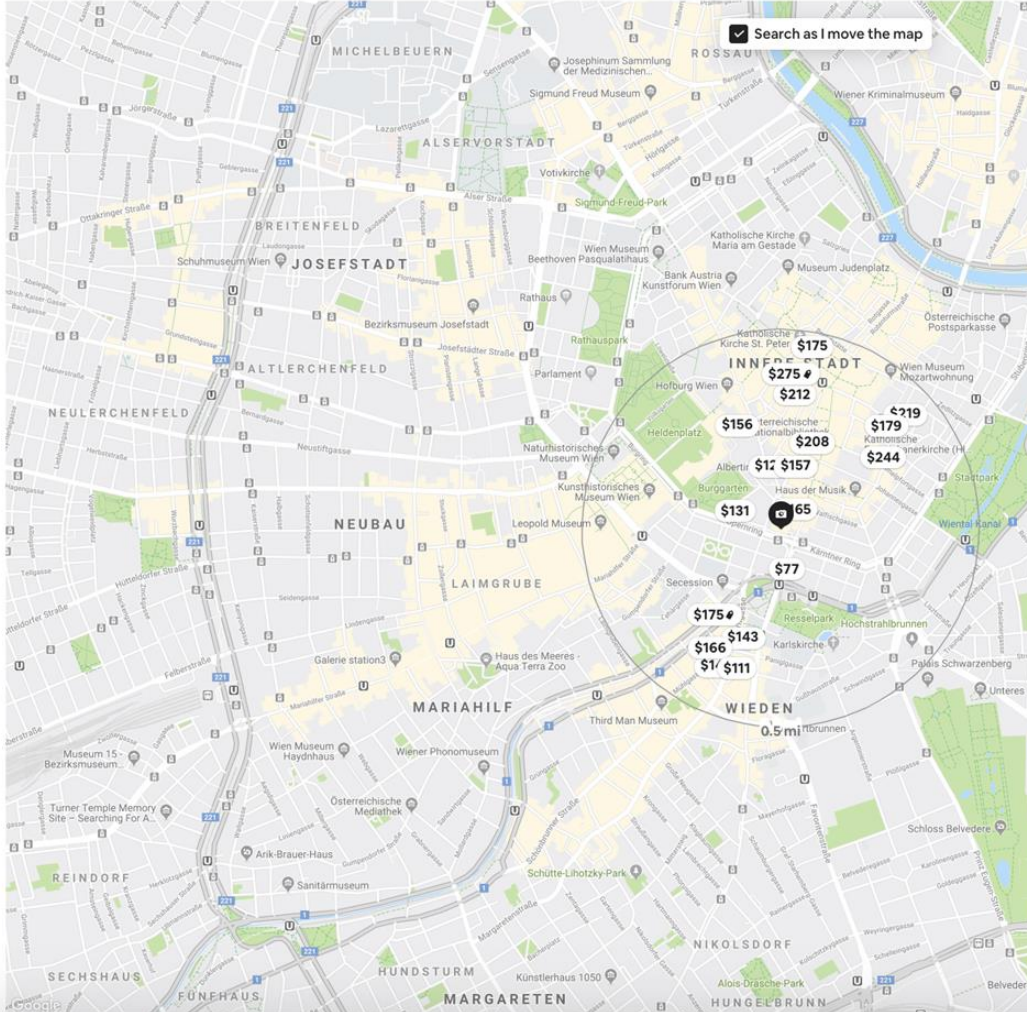
Example: price change experiment on Airbnb

Q Wien · Stays

Dec 20 - 23 2 guests Work trip Type of place Price Instant Book More filters

47 places to stay

-  **SUPERHOST** Entire apartment
Elegant modern flat in the heart of Vienna
2 guests · 1 bedroom · 1 bed · 1 bath
Wifi · Kitchen · Heating · Washer
★ 4.92 (37)
\$140 / night
\$505 total
-  Entire apartment
Living in a historic Apartment in the center
4 guests · 1 bedroom · 1 bed · 1 bath
Kitchen · Heating
★ 4.44 (59)
\$131 / night
\$492 total
-  Private room
Most Central modern Room in Historical Building
3 guests · 1 bedroom · 2 beds · 1 private bath
Wifi · Heating
★ 4.33 (9)
\$127 / night
\$462 total
-  Entire apartment
City Center Opera Apartment
3 guests · 1 bedroom · 1 bed · 1 bath
Wifi · Kitchen · Heating
★ 4.74 (114)
\$165 / night
\$610 total
-  Entire apartment
YOURS- quiet and sunny home at the heart of Vienna
4 guests · 1 bedroom · 1 bed · 1 bath
Wifi · Kitchen · Heating · Washer
★ 4.77 (44)
\$249 \$175 / night



Map of Vienna showing Airbnb listings with prices overlaid on various districts like Innere Stadt, Neubaun, and Mariahilf. The map includes a search bar and a 'Search as I move the map' checkbox.

Slide credit:
Dave Holtz,
UC Berkeley

Example: price change experiment on Airbnb

The screenshot shows an Airbnb search for 'Wien - Stays' for Dec 20-23 for 2 guests. The search results list 47 places to stay. The first four listings are highlighted with pink boxes:

- Listing 1:** **SUPERHOST** Entire apartment, Elegant modern flat in the heart of Vienna, 2 guests · 1 bedroom · 1 bed · 1 bath, Wifi · Kitchen · Heating · Washer. Price: \$140 / night, \$505 total.
- Listing 2:** Entire apartment, Living in a historic Apartment in the center, 4 guests · 1 bedroom · 1 bed · 1 bath, Kitchen · Heating. Price: \$131 / night, \$492 total.
- Listing 3:** Private room, Most Central modern Room in Historical Building, 3 guests · 1 bedroom · 2 beds · 1 private bath, Wifi · Heating. Price: \$127 / night, \$462 total.
- Listing 4:** Entire apartment, City Center Opera Apartment, 3 guests · 1 bedroom · 1 bed · 1 bath, Wifi · Kitchen · Heating. Price: \$165 / night, \$610 total.

The map on the right shows Vienna with price markers for various neighborhoods. A search radius of 0.5 miles is centered in the city center. Price markers include: \$175, \$275, \$212, \$156, \$208, \$179, \$244, \$112, \$157, \$131, \$65, \$77, \$175, \$166, \$143, \$111, \$249, and \$175 / night.

If lower fees on half of the listings, bookings for those listings ↑ 3% 😊

Slide credit: Dave Holtz, UC Berkeley

Example: price change experiment on Airbnb

The screenshot shows an Airbnb search for 'Wien - Stays' for Dec 20-23 for 2 guests. The search filters include 'Work trip', 'Type of place', 'Price', 'Instant Book', and 'More filters'. The results show 47 places to stay. The first five listings are highlighted with pink boxes:

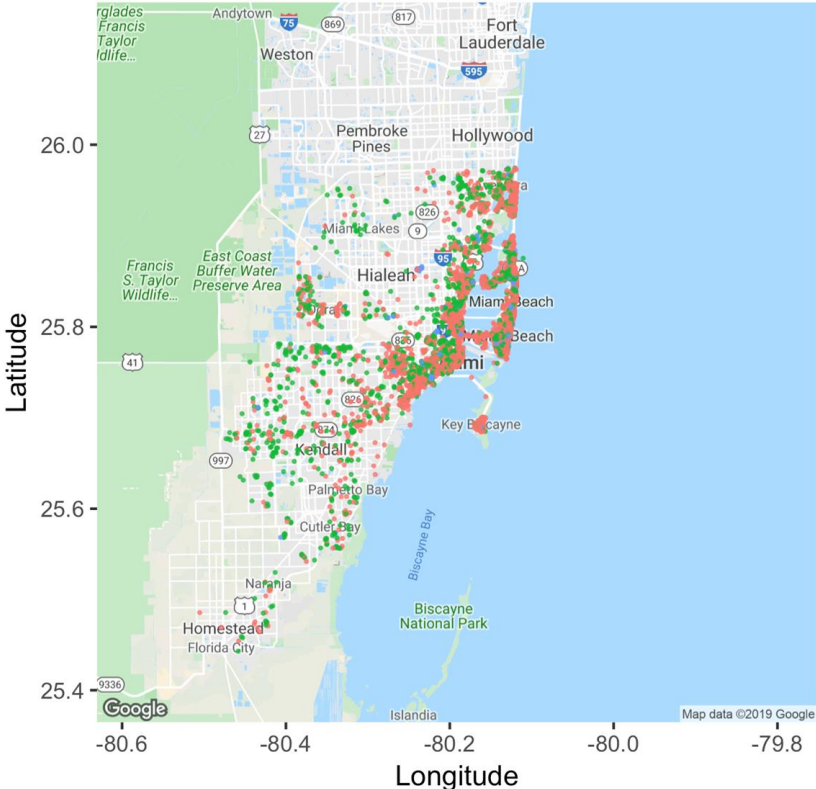
- Listing 1:** SUPERHOST Entire apartment, Elegant modern flat in the heart of Vienna, 2 guests - 1 bedroom - 1 bed - 1 bath, Wifi - Kitchen - Heating - Washer. Price: \$140 / night, \$505 total.
- Listing 2:** Entire apartment, Living in a historic Apartment in the center, 4 guests - 1 bedroom - 1 bed - 1 bath, Kitchen - Heating. Price: \$131 / night, \$492 total.
- Listing 3:** Private room, Most Central modern Room in Historical Building, 3 guests - 1 bedroom - 2 beds - 1 private bath, Wifi - Heating. Price: \$127 / night, \$462 total.
- Listing 4:** Entire apartment, City Center Opera Apartment, 3 guests - 1 bedroom - 1 bed - 1 bath, Wifi - Kitchen - Heating. Price: \$165 / night, \$610 total.
- Listing 5:** Entire apartment, YOURS- quiet and sunny home at the heart of Vienna, 4 guests - 1 bedroom - 1 bed - 1 bath, Wifi - Kitchen - Heating - Washer. Price: \$249 / night, \$175 total.

The map on the right shows Vienna with price markers for various districts: MICHELBEUERN, ALSESVORSTADT, BREITENFELD, JOSEFSTADT, NEULERCHENFELD, ALTLECHENFELD, NEUBAU, LAIMGRUBE, MARIAHILF, WIENEN, REINDORF, SECHSHAUS, HUNDSTURM, MARGARETEN, NIKOLSDORF, and HUNGELBRUNN. Price markers range from \$77 to \$275.

If lower fees
on all the
listings,
**Overall
bookings flat**
☹️

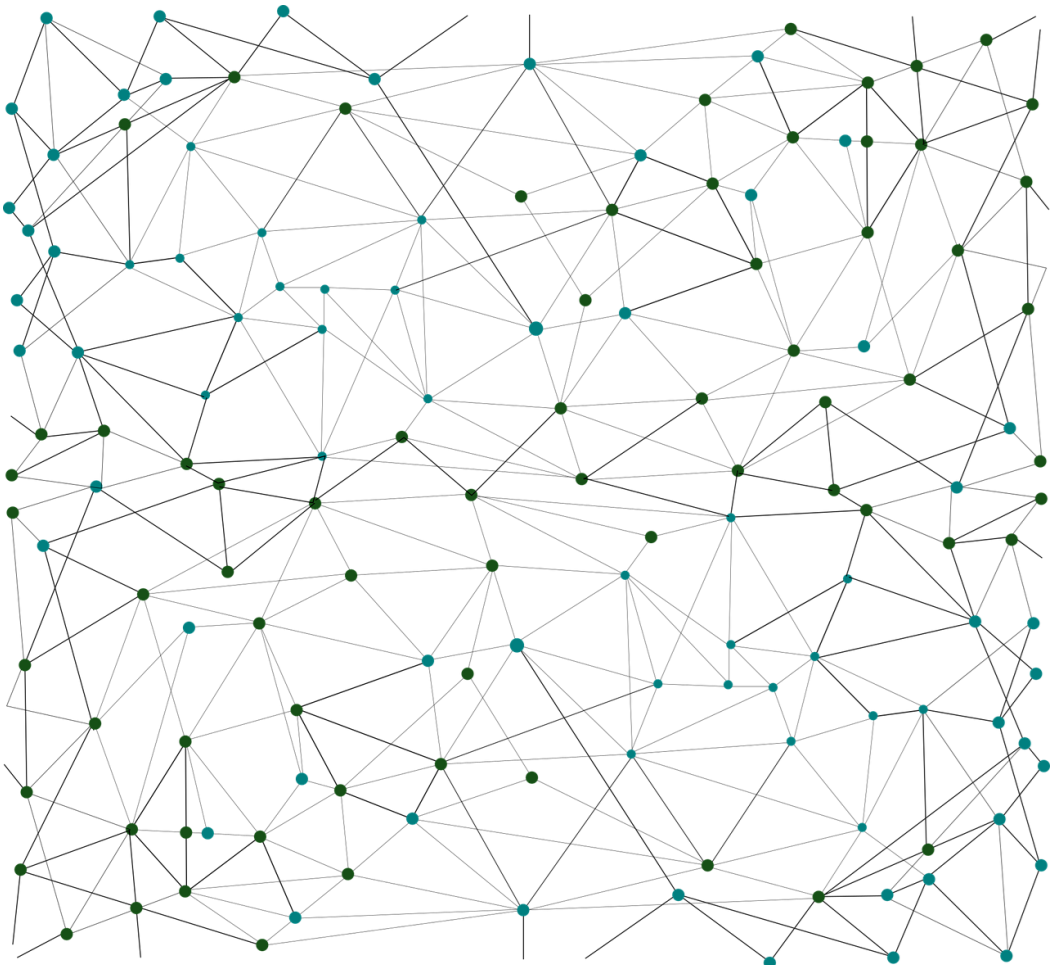
Slide credit:
Dave Holtz,
UC Berkeley

Approach 1: transform the marketplace into a network



Room Type

- Entire home/apt
- Private room
- Shared room



Network experiment designs + analysis techniques

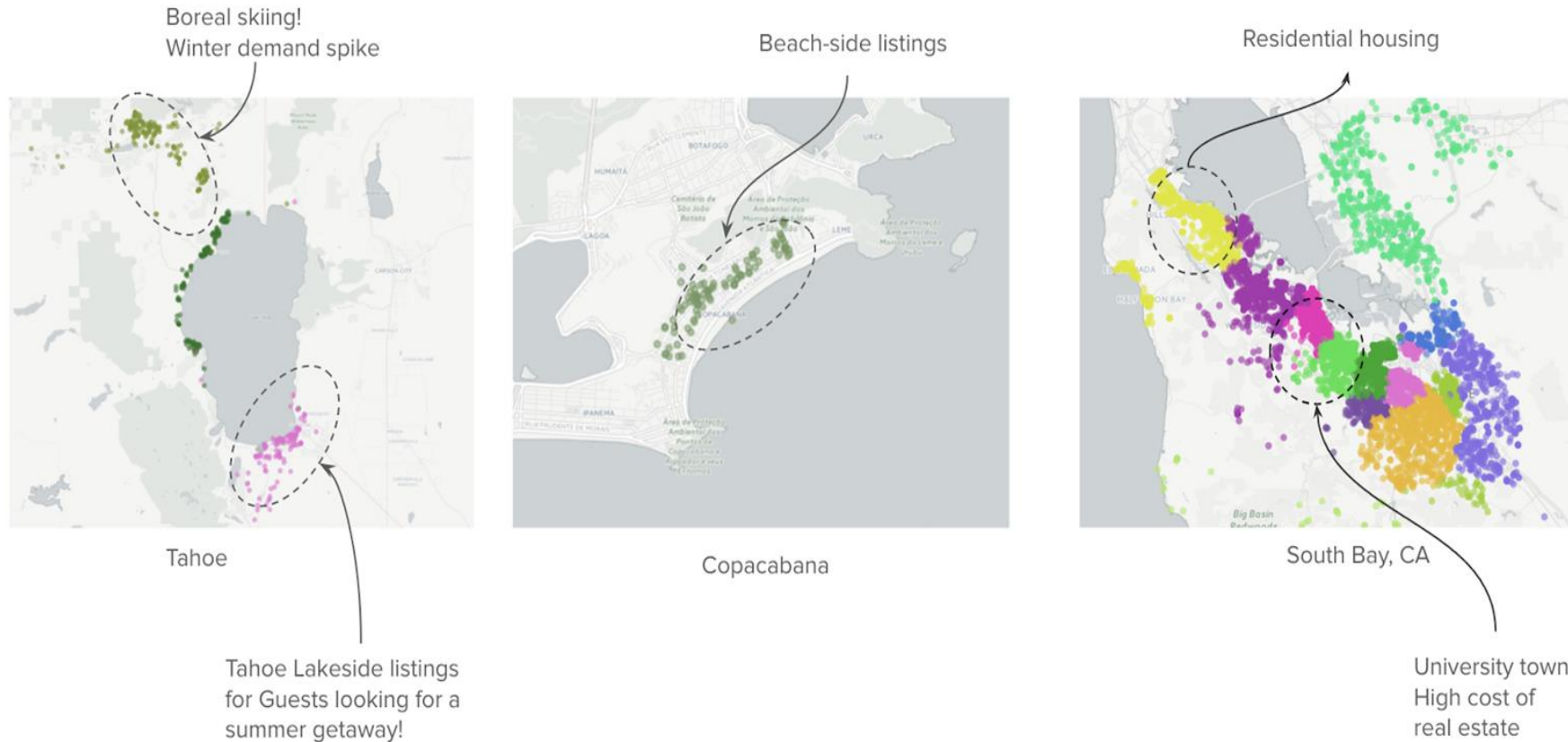


Image credit: Dave Holtz, UC Berkeley

- Now, listings are connected if they tend to be *substitutes*
- Much more complicated to learn the network structure
- Once have network structure, use cluster randomization techniques from above

Announcements

- Today 11/1 guest lecture virtually: Hannah Li on marketplace experimentation
 - 6:15pm – 7:30pm on Zoom
 - Will be recorded; required to watch it (will be on quiz)
 - Live attendance not required but is appreciated
- HW4 released; due next week
- Quiz 4 next week
- Project details released in next week
 - Project partner form on EdStem