Challenges in Measuring Online Advertising Systems

Saikat Guha[1], Bin Cheng[2], Paul Francis[2]

[1]Microsoft Research India and [2]MPI-SWS

IMC 2010

Privacy-Preserving Advertising Systems

- Basic idea:
  - Agent on client profiles user, targets ads
  - Various techniques to ensure privacy
- Bunch of proposals: AdNostic, Privad, Nurikabe
- Will client-side profiling work?
  - Us: Let's build it, deploy it, try it
  - Skeptics: Why not just use what Google uses?

Google is God Fallacy

- No one knows what info Google is using
- So they extrapolate:
  1. Can Google use `X`?
  2. $\Rightarrow$ Let's assume Google uses `X`
  3. $\Rightarrow$ Hence you need to use `X`

Problem Statement

Figure out what information is actually used to target ads today.

- Make hypothesis: X is used today
- Create two profiles: one, with X; one without
- See if ads differ
  - Sounds simple. Anything but!

Problem Statement

Figure out what information is actually used to target ads today.

- Make hypothesis: `X` is used today
- Create two profiles: one, with `X`; one without
- See if ads differ
  - Sounds simple. Anything but!

Problem Statement

Figure out what information is actually used to target ads today.

- Make hypothesis: `X` is used today
- Create two profiles: one, with `X`; one without
- See if ads differ
  - Sounds simple. Anything but!

Problem Statement

Figure out what information is actually used to target ads today.

- Make hypothesis: `X` is used today
- Create two profiles: one, with `X`; one without
- See if ads differ
  - Sounds simple. Anything but!

Problem 1: Same Ad?

Red Prom Dresses	=	Red Prom Dresses
Win a Free Dress for Prom 2010.	≠	Beautiful Designer Prom Dresses
Find New Trends; Great Prices!	≠	to Fit Every Figure; Price Range.
DavidsProm.com	=	DavidsProm.com
MD5(RedirURL): 8ebc...45dc	≠	*MD5(RedirURL): 3646...85d3*
Dest: ...detail.jsp?i=2462	=	*Dest: ...detail.jsp?i=2462*

Solution 1

Use display URL to uniquely identify ads.

Problem 2: Got All Ads?

- All ads that could be shown, not always shown
  - Limits on number of ads per page
  - Frequency capping
- Reload many times. More the better? No.
  - Problem: Ad churn

Solution 2

For search ads, reload $\sim$ 10 times (50 seconds) to capture snapshot.

Problem 3: How to Compare?

Given two sets of ads, compare using:

1. Set overlap (Jaccard index)?
2. Cosine similarity on #impressions?
3. Cosine similarity on (rank based) "value"?
4. Cosine similarity on log of #impressions?

...Problem: Noise. Lot's of it.

Solution 3

Use extended Jaccard index with logarithmic weights to compare snapshots.

Problem 4: Systemic Artifacts

- Puzzle: Three identical browser instances. Same endhost. Two get same set of ads. One gets different ads.
  - Answer: DNS round-robin. Different datacenter.

Solution 4

- Use static hosts file. Try to use one endhost.
- Failsafe: Noise-level control (two identical profiles)

Problem 4: Systemic Artifacts

- Puzzle: Different HTTP headers, different ads. Different IP address, different ads.
  - Our guess: Load-balancer's hashing algorithm.

Solution 4

- Use static hosts file. Try to use one endhost.
- Failsafe: Noise-level control (two identical profiles)

Results Summary

Network	**Information Tested**
Google Search	Keywords, ~~*Behavioral Targeting*~~
Google Contextual	Location, ~~Recent Searches?~~ ~~Recent Product Clicks?~~
Facebook	Age, Gender, Education, Location, Interests, Relationship Status, Sexual Orientation

Tempest in a Teacup . . .

Tempest in a Teacup . . .

Facebook's Privacy Policy

"we may use [...] information you may have decided not to show to other users, such as your birth year or other sensitive personal information or preferences) to select the appropriate [. . .] advertisements."

Facebook Advertising Guidelines

"Ads must not be false, misleading, fraudulent, or deceptive."

"You may not use user data you receive from us or collect through running an ad, including information you derive from your targeting criteria, for any purpose off of Facebook, without user consent."

Summary

- Systematic methodology to measure what information is actually used in online ad targeting today
- Analyzed three ad networks; slightly surprised Google doesn't personalize more than it appears to
- Served as filler for a few news cycles

Methodology 1: Same Ad?

- Datasets: Scraped Google search ads
  - All fashion related keywords
  - Dress related keywords
- Approach
  - Same redirect URL
  - Same display URL
  - Same title and display URL
  - Same title and summary (w/ keywords masked)
- Methodology: Manually classify false-positives and false-negatives

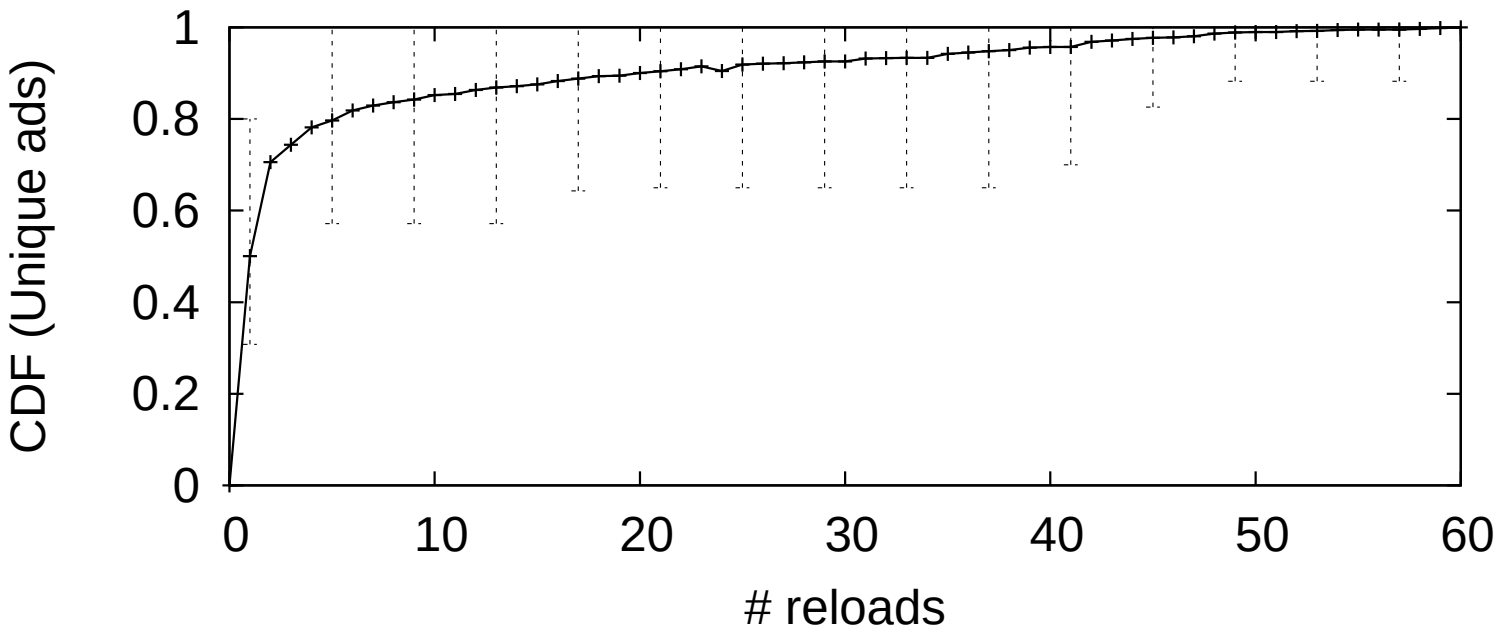
Methodology 1: Same Ad?

For 100 ad-pairs each:

Approach	All Fashion		Dresses only	
	% FP	% FN	% FP	% FN
RedirURL	0	38	1	52
DisplayURL	7	13	12	10
Title + DisplayURL	0	45	0	50
Title + Summary	0	68	0	69

Methodology 2: Got All Ads?

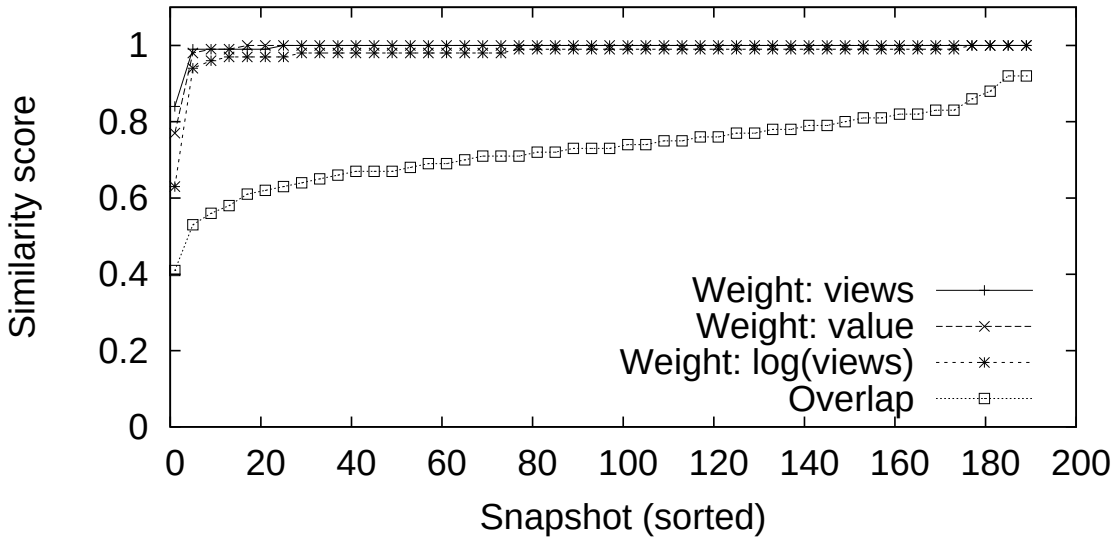
Google search ads; 200 product keywords; each keyword queried for 5m at 5s intervals:



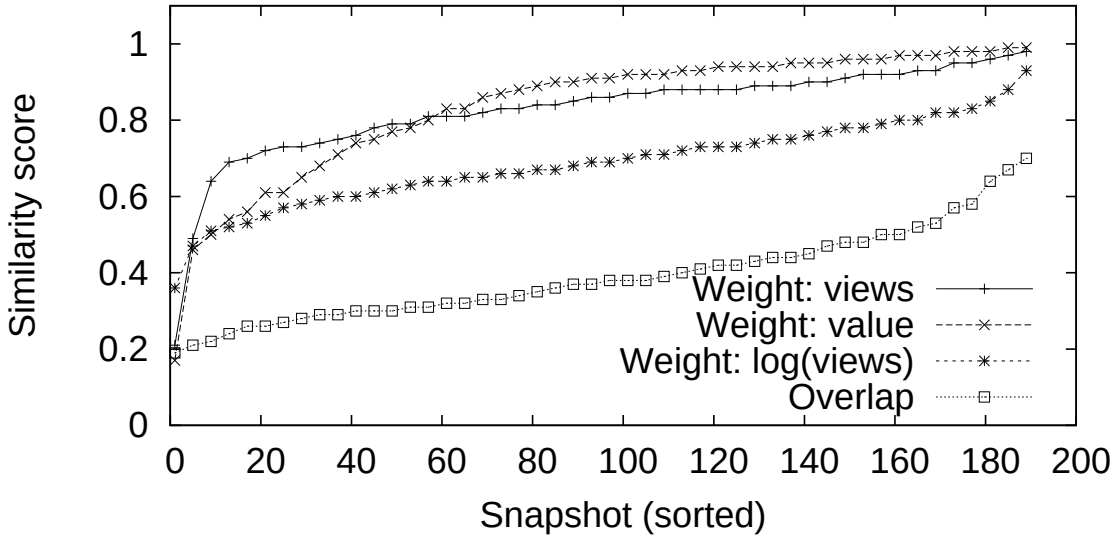
Methodology 3: How to Compare?

- Datasets: 15 queries; snapshots collected for 8d at 5m intervals
- Approach
  - Set overlap (Jaccard index)?
  - Cosine similarity on #impressions?
  - Cosine similarity on (rank based) "value"?
  - Cosine similarity on log of #impressions?
- Methodology: Vary client; check expectations
  - Identical clients $\longrightarrow$ expect significant similarity
  - Non-co-located clients $\longrightarrow$ expect some dissimilarity

Methodology 3: How to Compare?



Higher
=
Better



Lower
=
Better