Follow the Money: Understanding the Economics of Online Advertising

Presented by Andrew Kahn

Motivation

- Understanding the economic value of personal information
  - Highlight ways to increase ad revenues
  - Understand how users contribute to ad revenues
  - Model how changes in user behavior can affect ad revenues

Theory/Definitions

- Users: access of free content
- Publishers: host free content, sell page space to advertisers
- Aggregators: map advertisers to most effective ad placement based on publisher content and information about specific users

Aggregators

- Aggregator techniques:
  - Web bugs: “web bugs implemented through an embedded image include **tracking pixel**, **pixel tag**, **1×1 gif**, and **clear gif**.”
  - Cookies
  - Analytics

Aggregators

- Aggregators and publishers share revenues from impressions
  - Clicks left for future work
  - CPM (Cost per mile): price per 1000 impressions
- α: fraction of revenue earned by the aggregator (remaining fraction paid to the publisher)

Cost Per Mile

- $CPM(u, p, a) = RON_a \times TQM_p \times I_a(u)$

- RON_a = “Run-of-Network”, Base price in a certain ad network a
- TQM_p = “Traffic quality multiplier”, Multiplier based on value of publisher p
- $I_a(u)$ = “User Intent”, Intent of user u to purchase in ad network a

Modeling User Intent

- $II_a(u) <= EI(u)$

- Implicit Intent a of u: *Inferred* intent of u to purchase in an ad network a
- Explicit Intent of u: Actual intent of u to purchase given all of u's traffic

Revenue in Advertising

- $\sum u \sum p \sum a [u(p)/1000 * CPM(u, p\ a)]$

- "Sum over the users, publishers, and ad networks visits of a user to a page / 1000 * cost per 1000 impressions"

Data Analysis

- Datasets:

Table 1: Summary of data sets.

Trace	Setting	Country	Users	Sessions
HTTP	Neighborhood	A (4/2011)	~ 5K	40M
mHTTP	Country	B (8/2011)	~ 3M	1.5B
Univ	Campus	C (9/2010)	~ 8K	30M

- Definition:
  - Session: A page load.

Data Analysis

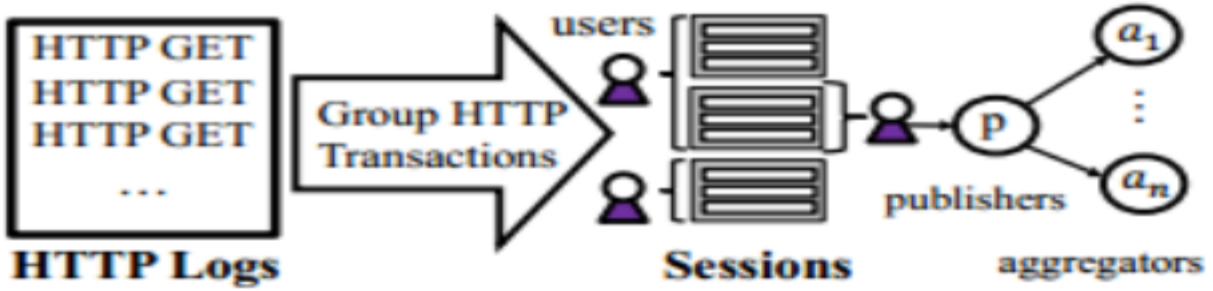


Figure 1: Data analysis pipeline to extract users, publishers and aggregators.

- Extract sessions from packet traces by (anonymized) user
- Classify aggregators per publisher by AS number

Data Analysis

- Session heuristics:
  - University dataset had Referer header set, used StreamStructure method to define session
  - HTTP/mHTTP used Content-Type header as “Text/HTML”. Must be more than 1s apart to join frames.
  - mHTTP used User-Agent to group application requests
  - Must contain more than 1 object, exclude known third-parties

Data Analysis

- Aggregator Identification:
  - First domain in session is the publisher
  - Domain from a different AS than the publisher is an aggregator
    - Unless they're a CDN
    - Or special cases that are ignored (Microsoft's msecd and nsatc)

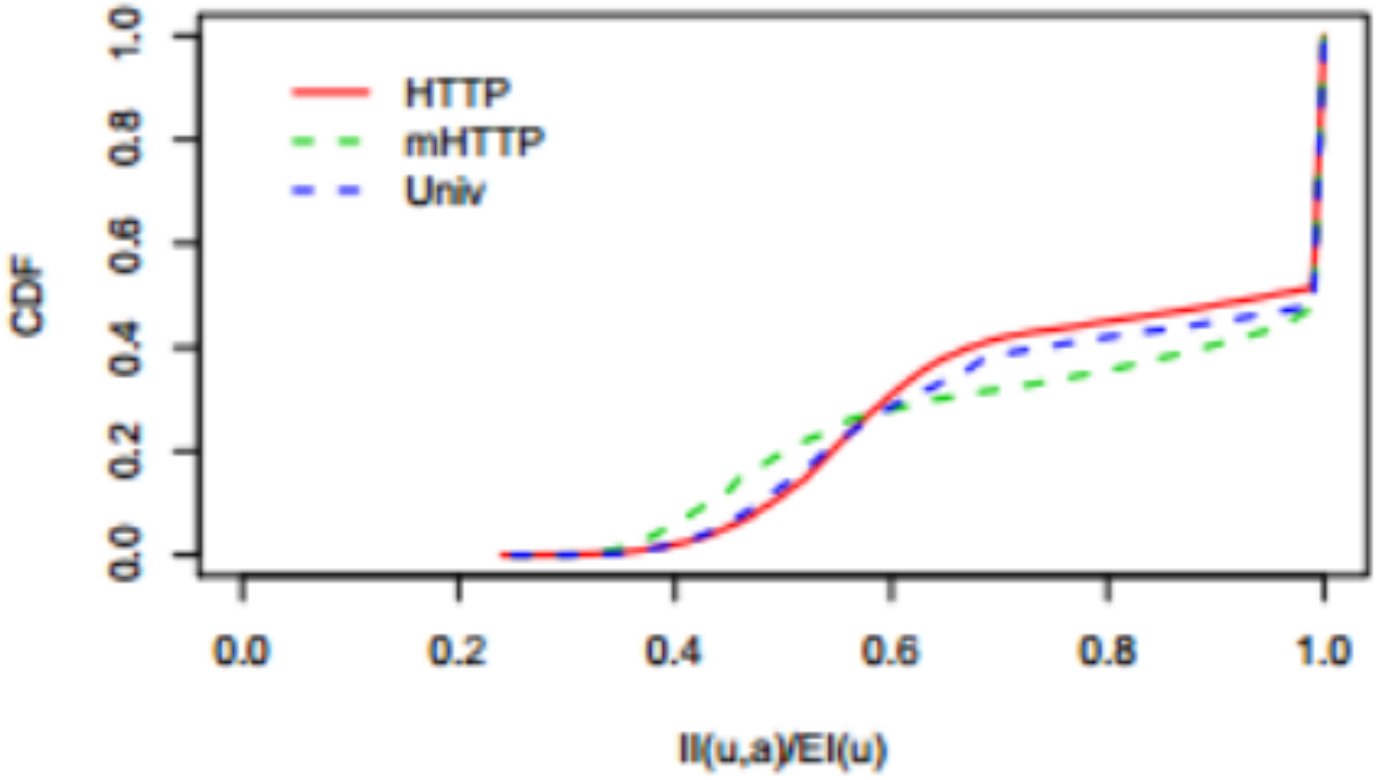
Data Analysis

- Calculating intent
  - Uses Alexa categorization of websites (espn.com: sports)
  - Google AdWords Contextual Advertising Tool converts category to “value” in range 2-10
  - Compute EI(u) and $II_a(u)$

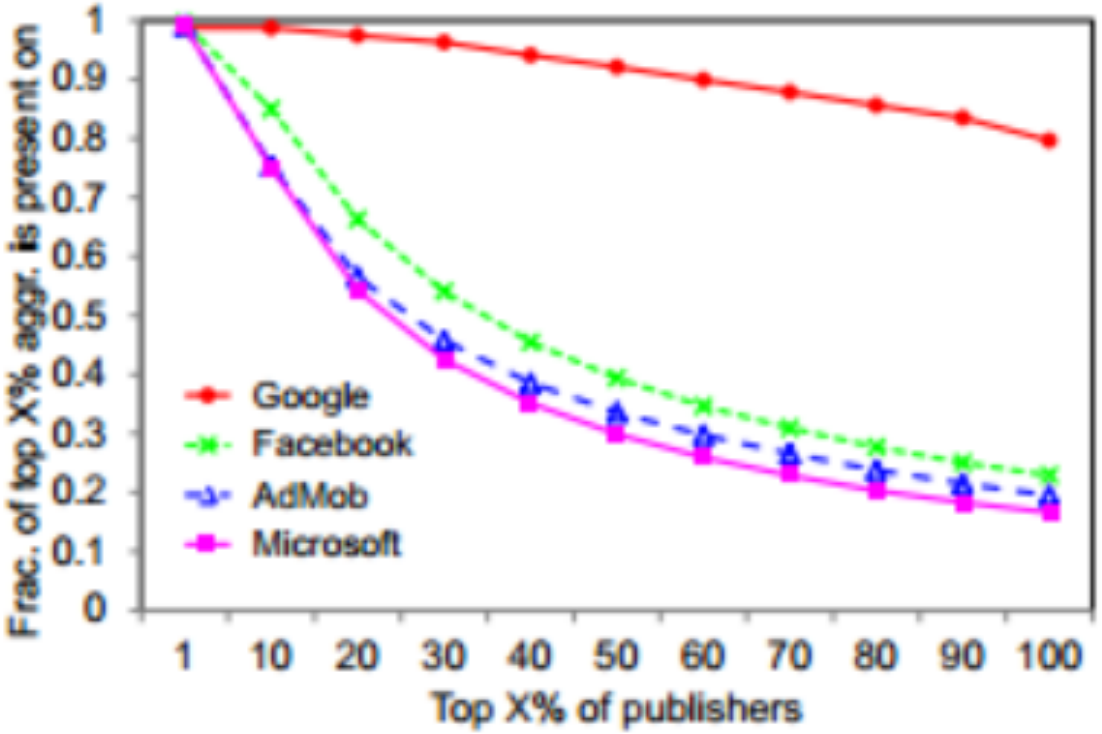
Data Analysis

- Calculating Traffic Quality Multiplier
  - 2 for Alexa top 500
  - .1 for DNS blacklisted sites
  - 1 otherwise
- Run-of-network
  - Set at $1.98
- Fraction earned by aggregators
  - α = .32 for Google Adsense

How much do advertisers know?



How much do advertisers know?



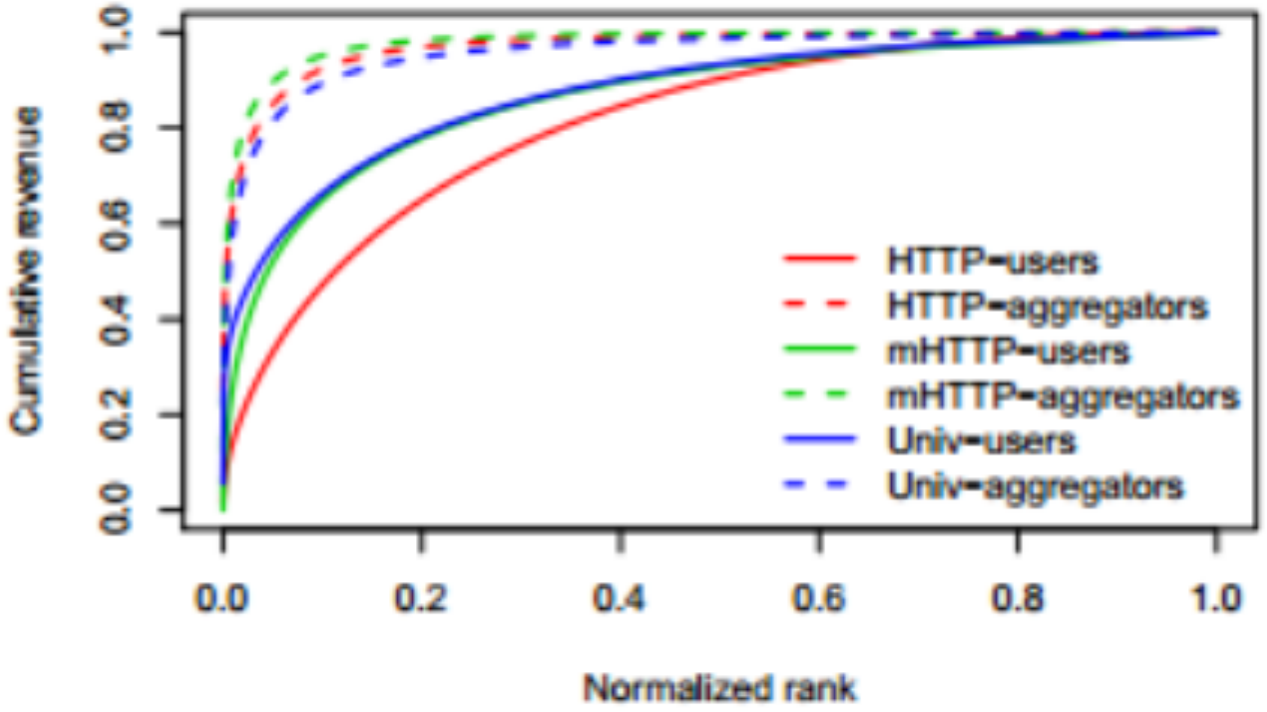
How valuable is this information?

- Aggregator statistics

Table 2: High revenue aggregators(mHTTP).

Aggregator	Frac. Rev.	Frac. Users	Frac. Pubs.
Google	0.18	0.17	0.80
Facebook	0.06	0.09	0.23
GlobalCrossing (AdMob)	0.04	0.11	0.19
AOL	0.03	0.04	0.07
Microsoft	0.03	0.04	0.17
Omniture	0.03	0.05	0.07
Yahoo! (AS42173)	0.03	0.04	0.07

How skewed are ad revenues?



What information vectors are most lucrative?

- Users are valuable that:
  - Have more sessions
  - Visit sites in popular categories
- The most popular publishers do not necessarily generate the most revenue
  - Correlated more closely with number of aggregators present on a publisher

Popular Publishers

- Revenue generating publishers

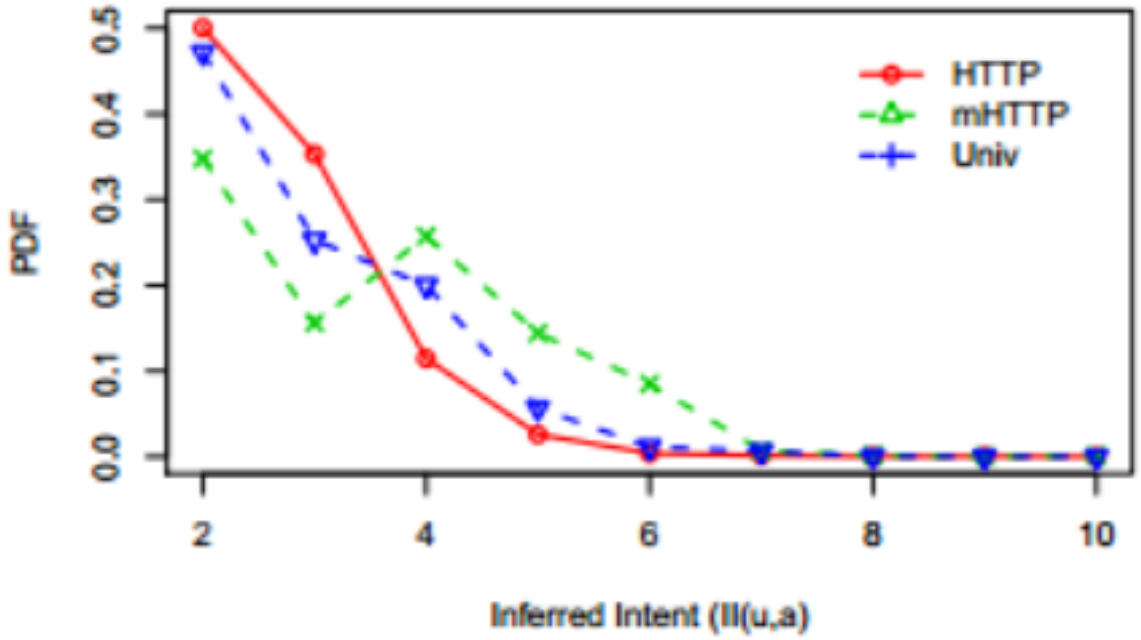
Table 3: High revenue publishers (mHTTP).

Publisher	Frac. Rev.	Frac. Users	Category
facebook.com	0.09	0.15	society
google.co.uk	0.04	0.11	computers
bbc.co.uk	0.03	0.07	arts
fbcdn.net	0.03	0.13	society
twitter.com	0.03	0.04	computers
yahoo.com	0.03	0.04	computers
google.com	0.02	0.18	computers

Quantifying Losses

- Blocking tracking prevents coverage of the implicit intent
  - Top 5% blocking decreases revenues by 35%-60%
  - All users blocking decreases revenues by a factor of 4.2
- Blocking
  - Limit Javscript execution, deny cookies, inject noise into searches

Importance of Inferring Intent



Conclusions

- A few aggregators dominate revenues through coverage of users and publishers
  - Top 5% of aggregators get 90% of revenue

Conclusions

- Aggregators have coverage of few publishers
  - Google at 80% coverage of publishers is the exception
  - Top aggregators cover 70% of top 10% of publishers
  - Only cover 10%-20% of publishers overall

Conclusions

- Users are not as heavily skewed and contribute more evenly to revenues
  - Top 35%-55% of users account for 90% of revenues

Conclusions

- Preventative measures impacting the tracking of aggregators would significantly impact advertising revenues
  - No prediction of intent, 300-400% drop in revenues