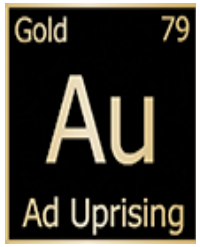
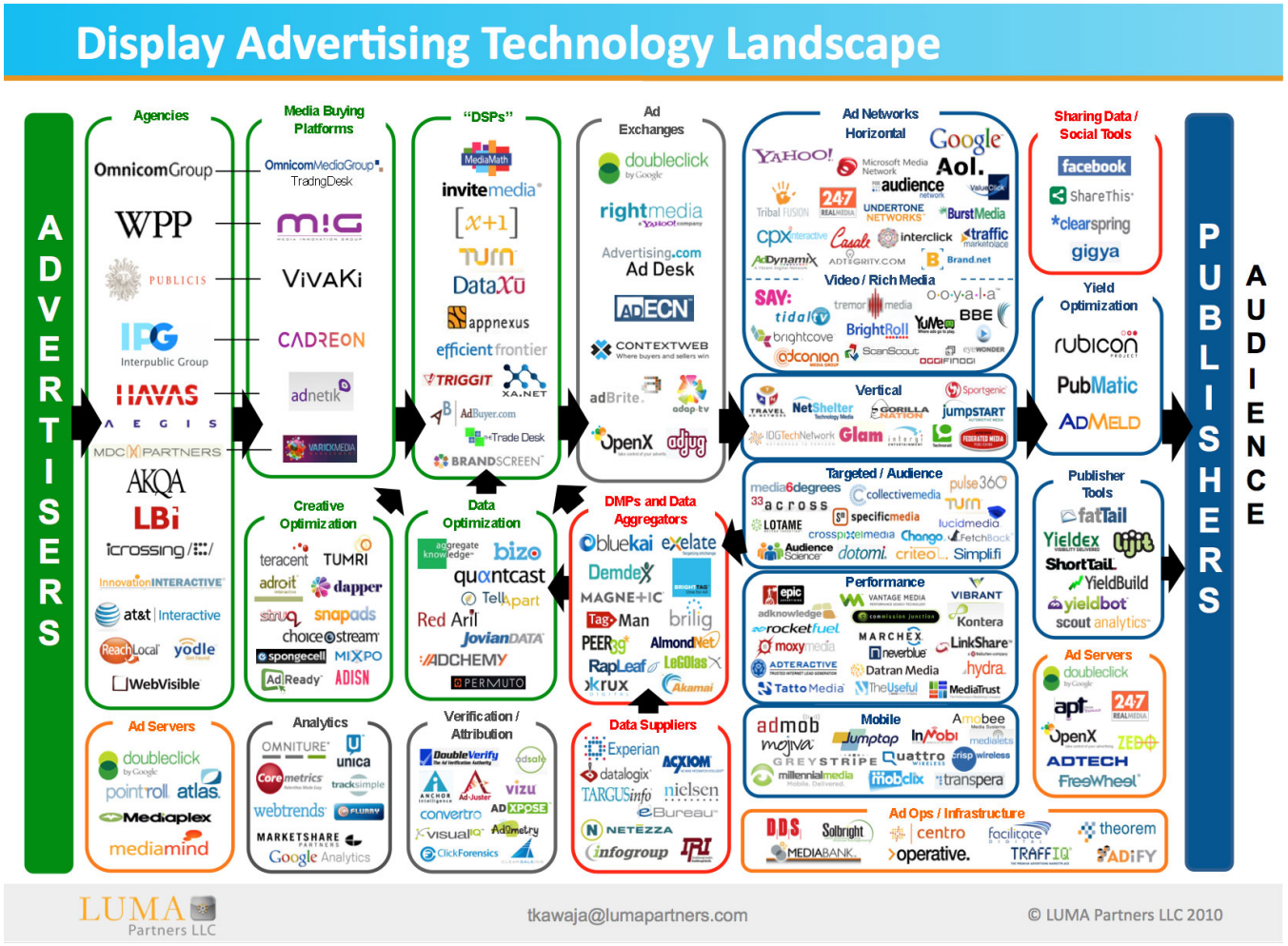
Perspectives on Measuring and Analyzing Online Ads

Paul Barford

Spring 2014



Motivation


Display Advertising Technology Landscape
ADVERTISERS
Agencies
OmnicomGroup
WPP
PUBLICIS
IPG
Interpublic Group
HAVAS
AEGIS
MDC PARTNERS
AKQA
LBi
icrossing
InnovationINTERACTIVE
at&t | Interactive
ReachLocal
yodle
WebVisible
Media Buying Platforms
OmnicomMediaGroup
TradingDesk
m!c
ViVAKi
CADREON
adnetik
VARICKMEDIA
"DSPs"
MediaMath
invitemedia
[x+1]
TURN
DataXu
appnexus
efficient frontier
TRIGGIT
XA.NET
AdBuyer.com
Trade Desk
BRANDSCREEN
Ad Exchanges
doubleclick by Google
rightmedia
Advertising.com
Ad Desk
ADECN
CONTEXTWEB
adBrite
adap.tv
OpenX
adjug
Ad Networks Horizontal
YAHOO!
Google
Aol.
Microsoft Media Network
audience network
Tribal FUSION
247 REALMEDIA
UNDERTONE NETWORKS
BurstMedia
cpx interactive
Casale
interclick
traffic marketplace
AdDynamix
ADTEGRITY.COM
Brand.net
Video / Rich Media
SAY:
tidal tv
tremor media
o·o·y·a·l·a
brightcove
BrightRoll
YuMe
BBE
adconion
ScanScout
eyeWONDER
OOYALA FINDIO
Vertical
TRAVEL
NetShelter
GORILLA NATION
jumpSTART
IDGTechNetwork
Glam
interest
FEDERATED MEDIA
Sportgenic
Targeted / Audience
media6degrees
collectivemedia
pulse360
33across
TURN
LOTAME
specificmedia
lucidmedia
crossPixelmedia
Chango
FetchBack
Audience Science
dotomi
criteo
Simplifi
Performance
epic
adknowledge
VANTAGE MEDIA
VIBRANT
rocketfuel
Kontera
moxymedia
MARCHEX
neverblue
LinkShare
ADTERACTIVE
Datran Media
hydra
Tatto Media
TheUseful
MediaTrust
Mobile
admob
mojiva
Jumptap
InMobi
Amobee
GREYSTRIPE
Quattro
millennialmedia
MobClix
transpera
Ad Ops / Infrastructure
DDS
Solbright
MEDIABANK
centro
facilitate
theorem
operative.
TRAFFIQ
ADIFY
Sharing Data / Social Tools
facebook
ShareThis
clearspring
gigya
Yield Optimization
rubicon
PubMatic
ADMELD
Publisher Tools
fatTail
Yieldex
ujjo
ShortTail
YieldBuild
yieldbot
scout analytics
Ad Servers
doubleclick by Google
apt
247 REALMEDIA
OpenX
ZEDO
ADTECH
FreeWheel
PUBLISHERS
AUDIENCE
Ad Servers
doubleclick by Google
pointroll
atlas
Mediaplex
mediamind
Analytics
OMNITURE
unica
Coremetrics
tracksimple
webtrends
FLURRY
MARKETSHARE PARTNERS
Google Analytics
Verification / Attribution
DoubleVerify
adsafe
ANCHOR Intelligence
Ad-Juster
vizu
convertro
ADXPOSE
visualIQ
Adometry
ClickForensics
Creative Optimization
teracent
TUMRI
adroit
dapper
struq
snapads
choicestream
spongecell
MIXPO
AdReady
ADISN
Data Optimization
aggregate knowledge
bizo
quantcast
TellApart
Red Aril
JovianDATA
ADCHEMY
PERMUTO
DMPs and Data Aggregators
bluekai
exelate
Demdex
MAGNETIC
BRIGHT TAG
Tag Man
brilig
PEER39
AlmondNet
RapLeaf
LeGoLas
krux
Akamai
Data Suppliers
Experian
ACXIOM
datalogix
TARGUSinfo
nielsen
eBureau
NETEZZA
infogroup
IRI
LUMA Partners LLC
tkawaja@lumapartners.com
© LUMA Partners LLC 2010

Outline

- **Overview**
  - **Background, objectives and challenges**
- **The publisher perspective**
  - **m.Labs & PPV nets**
- **The user perspective**
  - **Ad Uprising & adscape**

Objectives

- Investigate behavior and characteristics of the ad eco-system
  - Develop techniques for measuring aspects of the online ad eco-system
  - Compile diverse data repositories
- Develop new mechanisms that improve performance, yield, security and privacy
  - Many opportunities!
- Commercial impact
  - But it's a cluttered space!

Publisher revenue deletions

Invalid traffic includes both clicks and impressions that Google suspects to not be the result of genuine user interest

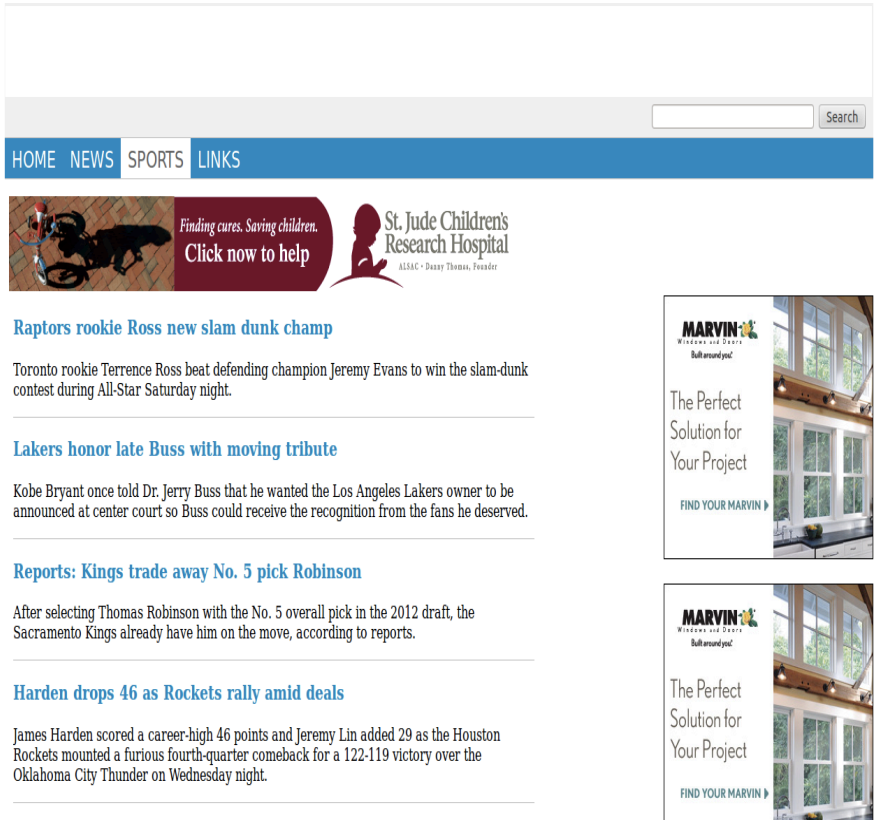
- **Standard means for *valid* traffic - AdWords**
- **Google simply notifies publishers that invalid traffic led to $XX deduction from you account**
  - **Large internal group that monitors traffic quality**
- **m.Labs – Web user & traffic quality analytics**
  - **Identify invalid impressions and clicks in real time**

Traffic generation

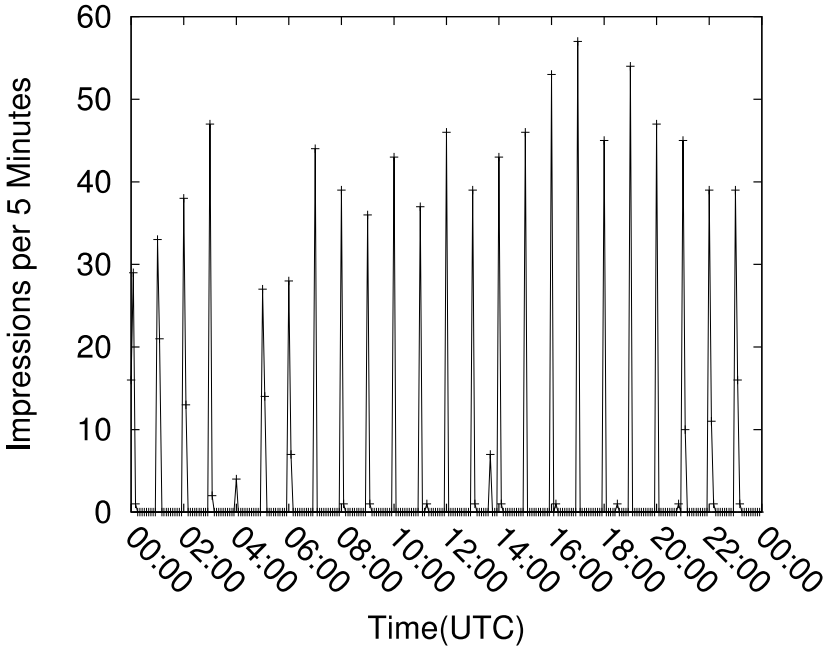
- Type “purchase web traffic” in Google
  - MANY traffic generation offerings
- Simple threats: script-based page retrieval
  - Ubiquitous - $12/10K impressions
- More complex threats: botnets
  - Geotargeting, clicks, and other characteristics
  - As much as $100/10K impressions
- Pay-per-view networks
  - Websites that load 3rd party pages in an obfuscated fashion when accessed by users

Honeypot websites

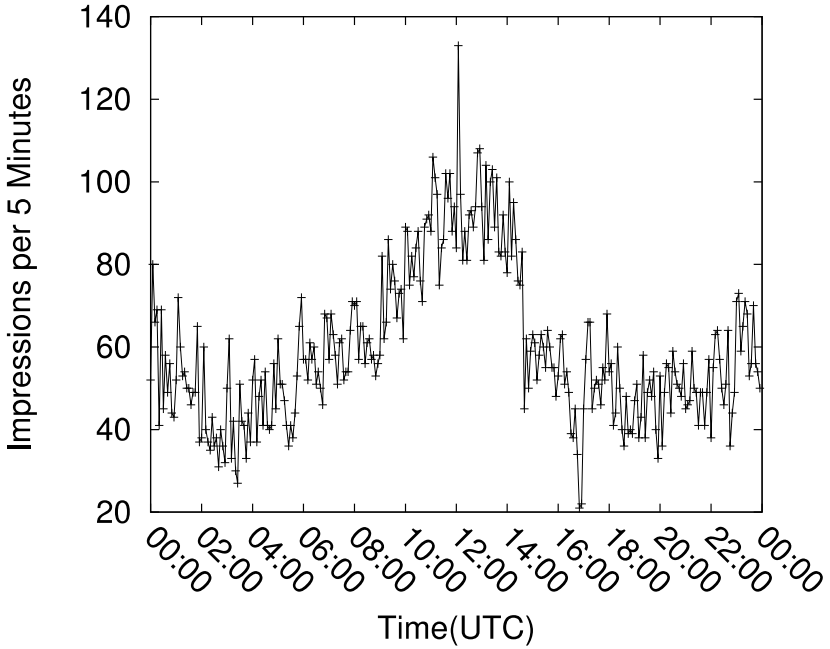
- Series of websites developed to be targets for traffic generation
  - Look and feel of a "real" site
- Instrumentation
  - Gather as much data per access as possible



Purchased traffic



BuildTraffic arrival process



AeTraffic arrival process

Deep dive into PPV nets

- When a user accesses a site running a PPV tag a pop-under window is generated
  - Typically requires a user action
- Pop-under calls PPV network server
  - Delivers details on user and site
- PPV network will deliver URL's of sites buying traffic
  - Often to 0 height frames
  - Frequent reloads

- K. Springborn and P. Barford, "Impression Fraud in On-line Advertising via Pay-Per-View Networks" To appear in the USENIX Security Symposium, 2013

Scope and impact of PPV nets

- Many PPV sites publish their volume
  - Average of 17.16M unique visitors and 6.29B page views per provider per day are claimed
- We searched Jan-June ‘12 Common Crawl DB for PPV tags from 10 providers
  - Over 4M PPV tags found on over 11K domains
- We used MuStats to estimate daily page views on identified pages
  - Over 168M daily page views
- Over $15M/month in wasted ad spend from 10 PPV networks alone!

Reset: the user perspective

- Research question: *what display ads are being delivered to users?*
- Targeting problem: select an ad to deliver to a user accessing a particular web page
  - Objectives: build awareness, click through
  - Context, geography, placement, behavior, etc.
  - Targeting mechanisms are intrinsic to online ad eco-system
- AdUprising – Identifying the ***Internet Adscape***
  - What is being shown where and to whom?

Challenges in Adscape identification

- **Scope – millions of publisher sites**
- **Complexities of ad campaigns**
  - **Demo, geo, site lists, caps, etc.**
- **Complexities of publisher ad placements**
  - **Premium, exchanges, backfill, etc.**
- **Complexities of ad delivery and targeting mechanisms**
  - **This is what we're seeking to understand**

Building a display ad crawler

- We seek to understand ad delivery by harvesting ads from a broad set of sites
- Ad crawler requirements
  - Distinguish and collect ads from other images
  - Collect related data
  - Accommodate gigantic scale and highly dynamic nature of ads in a gentle fashion
  - Personalization
- We developed a scalable, profile-based ad crawler based on Firefox/firefly

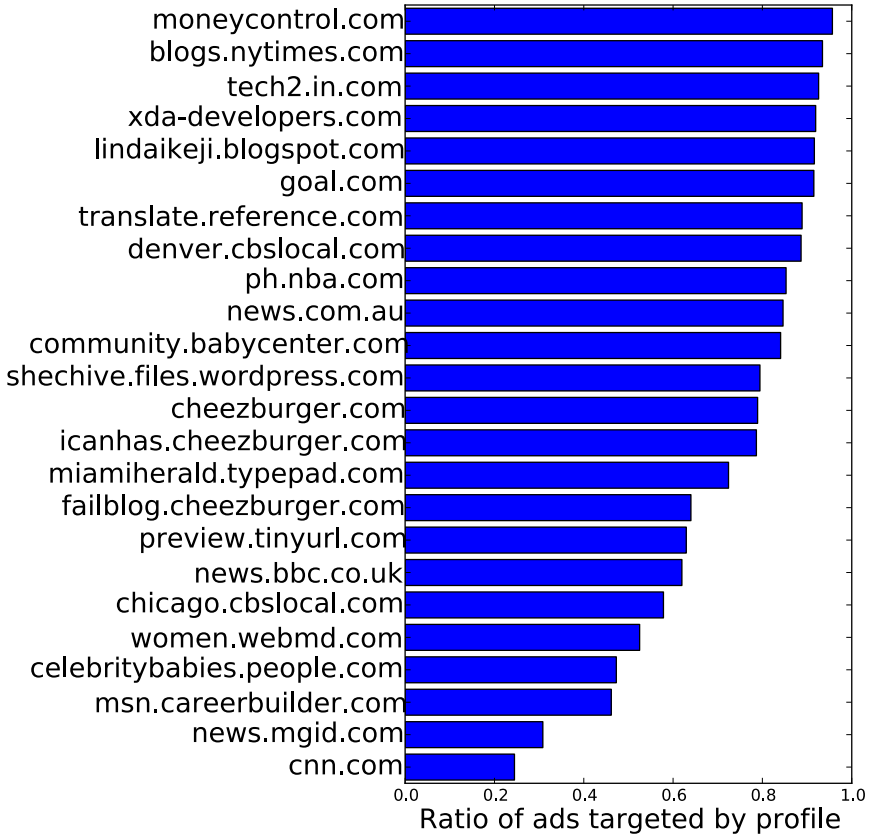
Building user profiles

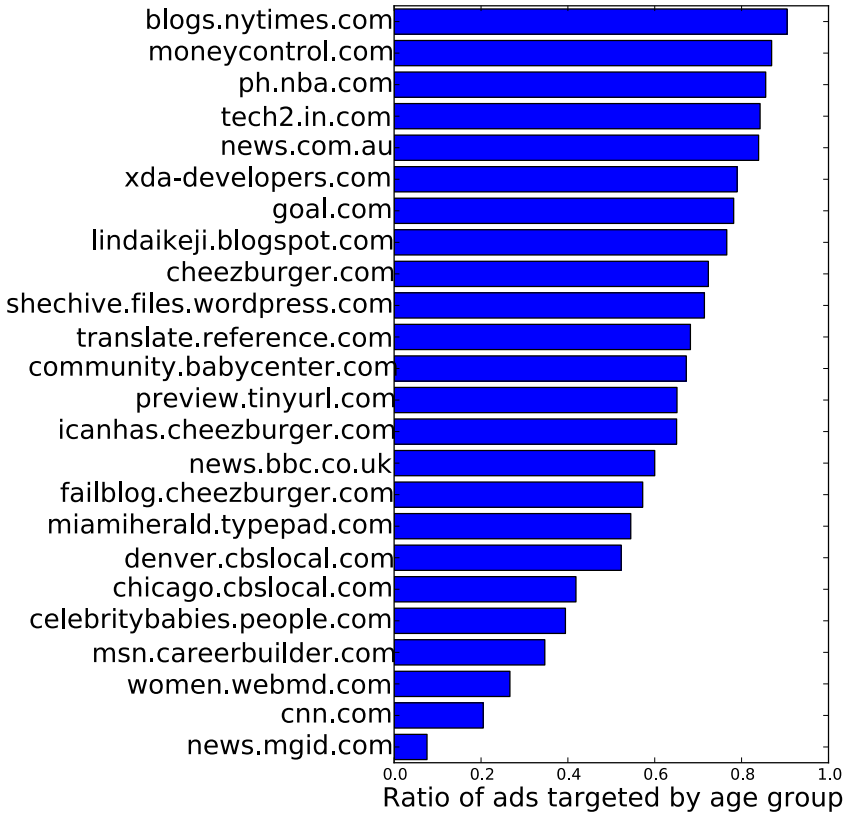
- We assume *profiles* can be built based on browsing careful selection of sites
  - Assume single interest users
- Site selection is base on Alexa categories
- Profiles are created by browsing top 100 websites from an Alexa category
  - Profiles are basis for ad collection
- Profile maintenance is a challenge when gathering ads from different sites
  - We find that profiles change significantly

Collecting ads

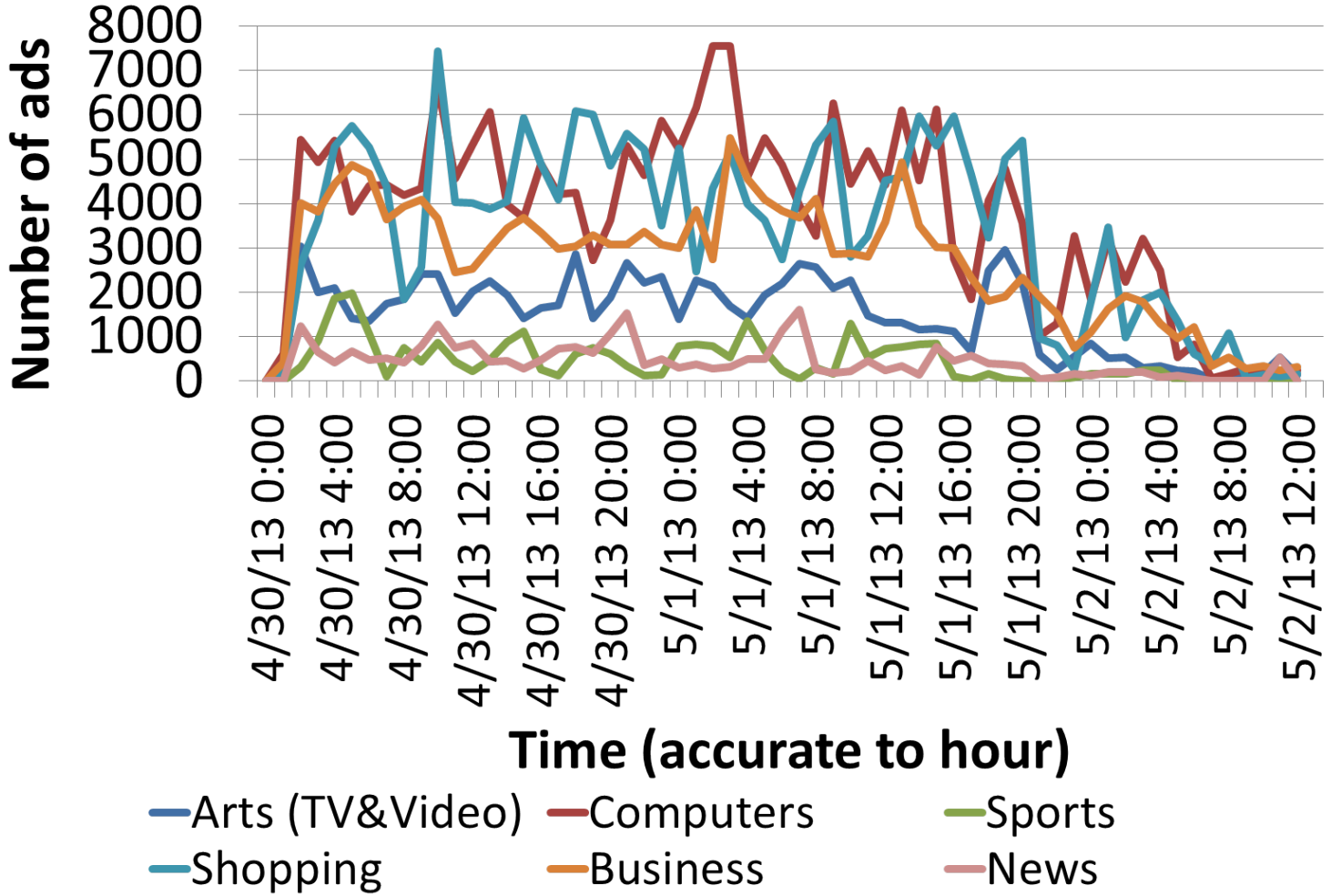
- Objective: gather as many distinct ads as possible
- Simplifications
  - Single collection site (no geo diversity)
  - No consideration of time varying effects
  - Limited crawl to fixed period and fixed sites
- 134 sites crawled using 60 different profiles
  - 462 [website, profile] pairs
- Over 175K unique ads gathered from ~3.5K entities

Ad targeting





Ads impressions by category



Profiles vs. ad categories

	Arts	Business	Computers	Games	Health	Home	Kids_and_Teens	News	Recreation	Reference	Science	Shopping	Society	Sports
Arts	0.098	0.239	0.199	0.037	0.031	0.016	0.048	0.026	0.022	0.038	0.000	0.187	0.029	0.030
Business	0.099	0.299	0.145	0.015	0.018	0.011	0.024	0.044	0.015	0.083	0.000	0.173	0.038	0.036
Computers	0.125	0.099	0.415	0.021	0.012	0.008	0.031	0.007	0.012	0.030	0.000	0.193	0.025	0.022
Health	0.070	0.164	0.181	0.020	0.083	0.014	0.017	0.033	0.027	0.101	0.001	0.199	0.065	0.025
Home	0.102	0.230	0.287	0.002	0.011	0.030	0.003	0.017	0.034	0.100	0.010	0.123	0.032	0.019
News	0.070	0.313	0.197	0.014	0.038	0.008	0.003	0.050	0.037	0.060	0.002	0.116	0.037	0.056
Recreation	0.135	0.204	0.199	0.006	0.014	0.012	0.005	0.037	0.086	0.039	0.001	0.154	0.039	0.070
Reference	0.058	0.171	0.235	0.009	0.021	0.024	0.007	0.007	0.030	0.163	0.002	0.168	0.102	0.003
Shopping	0.106	0.132	0.231	0.004	0.012	0.005	0.004	0.012	0.013	0.036	0.000	0.409	0.021	0.015
Sports	0.087	0.207	0.249	0.001	0.025	0.012	0.012	0.017	0.032	0.108	0.001	0.158	0.029	0.063

Alexa profiles in rows, ad categories in columns, percentage of ads shown in cells

Barford et al. "Adscape: Harvesting and Analyzing Online Display Ads" Under submission to ACM IMC, 2013.

Summary and status

- Fraud is a gigantic problem in online advertising
  - From simple scripts to sophisticated bots to PPV networks
- m.Labs has developed filters to identify fraud
- m.Labs open experimental platform
  - Data repository and API for ad fraud detection

- The Internet Adscape is huge and diverse
- Initial study to characterize the Adscape
- Ad Uprising: ongoing data collection and analysis

Thank you!

- **Igor Canadi**
- **Darja Krushevskaja**
- **Qiang Ma**
- **Muthu**
- **Kevin Springborn**
- **Charles Thomas**