The Price of Free: Privacy Leakage in Personalized Mobile In-App Ads

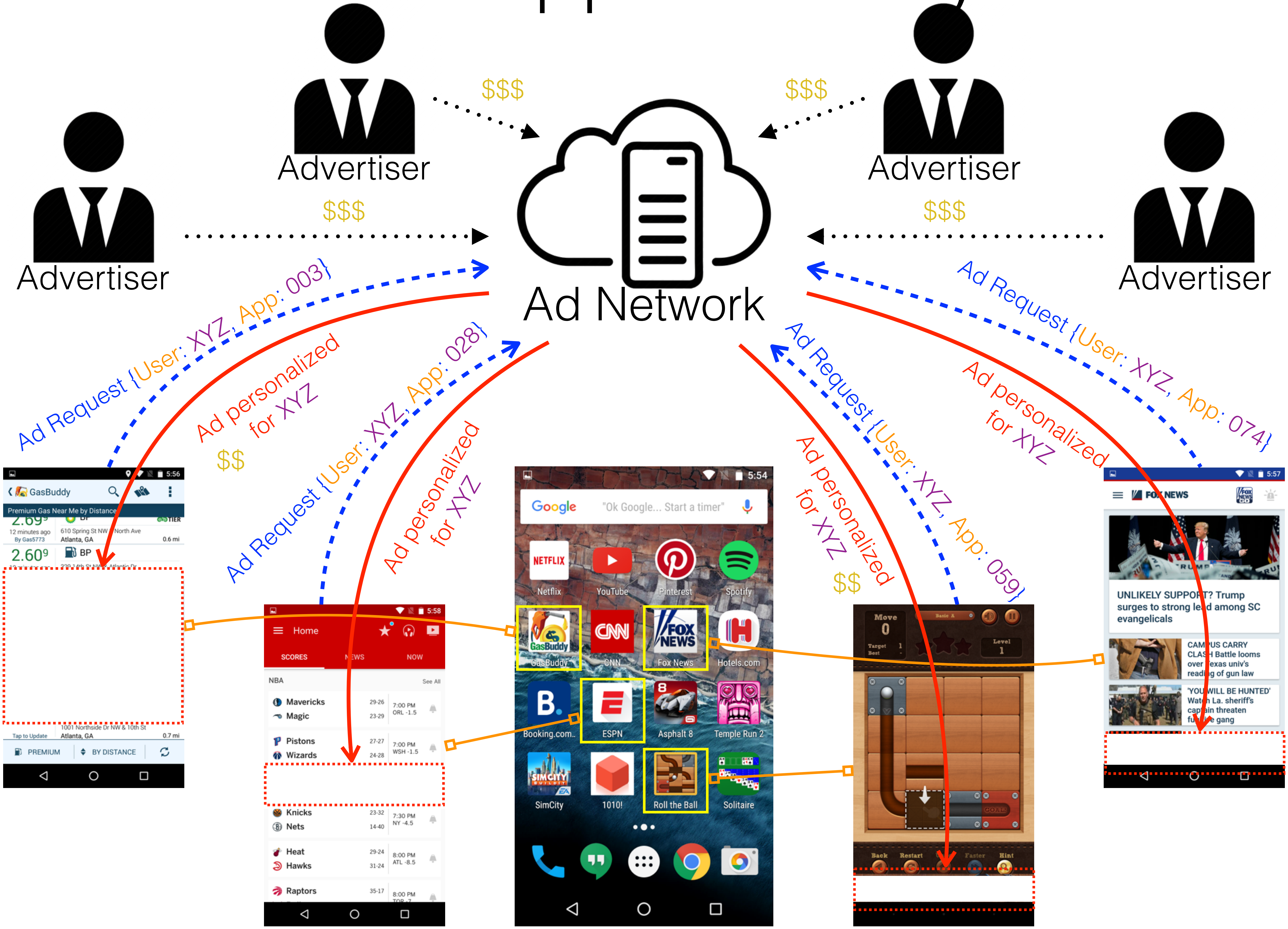
Wei Meng, Ren Ding, Simon P. Chung,
Steven Han, Wenke Lee

College of Computing
Georgia Institute of Technology

Outline

- Background & Motivation
- Methodology
- Characterization of Mobile Ad Personalization
- Privacy Leakage through Personalized Mobile Ads
- Discussion

Mobile In-App Ad Ecosystem



Previous & Recent Work on Mobile Advertising

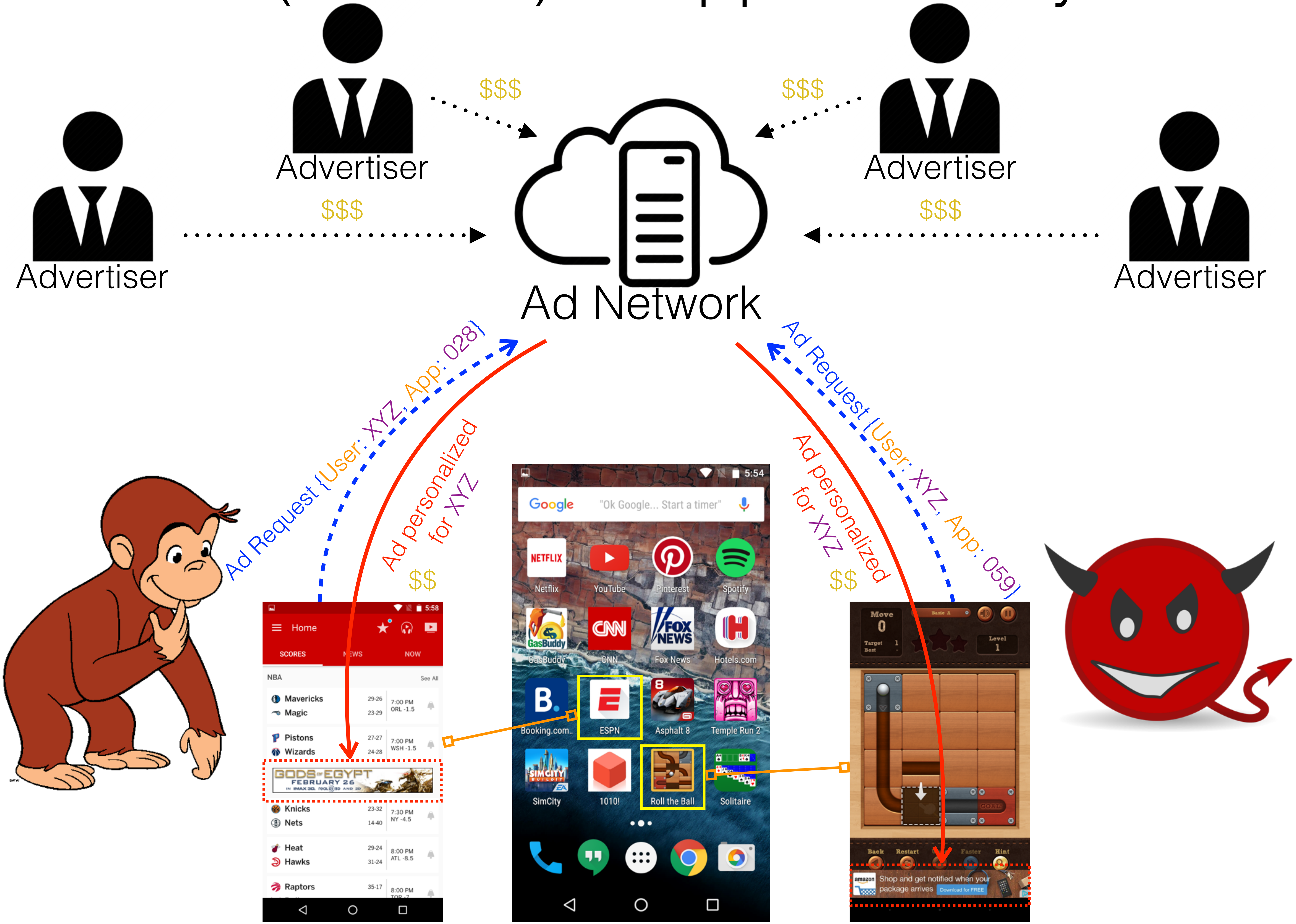
- Targeting & personalization
  [SmartAds (MobiSys'13), MAdScope (MobiSys'15)]

- Privilege abuse by mobile ad libraries
  [AdSplit (Security'12), AdDroid (ASIACCS'12), LayerCake (Security'13), …]

- Fraud in mobile advertising
  [AdSplit (Security'12), LayerCake (Security'13), DECAF (NSDI'14)]

- Privacy-Preserving mobile advertising
  [M. Götz, etc. (CCS'12)]

Mobile (Android) In-App Ad Ecosystem



Advertiser

Advertiser

Advertiser

Advertiser

$$$

$$$

$$$

$$$

Ad Network

Ad Request {User: XYZ, App: 028}

Ad personalized for XYZ

$$

Ad Request {User: XYZ, App: 059}

Ad personalized for XYZ

$$

This Work

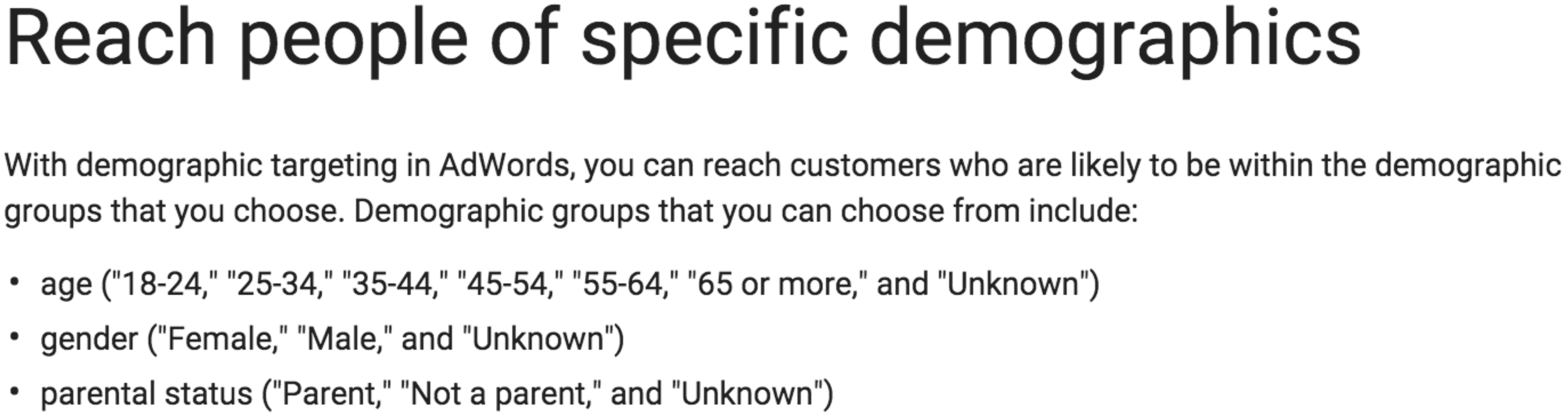
- Characterizing mobile in-app ad personalization for real people
  - What personal information about real end users a dominate ad network such as Google know and use in personalized mobile advertising?
- Estimating mobile app’s ability of learning about a user by observing personalized ads
  - Can an adversary with access to personalized mobile ads gain any information about real users?

Outline

- Background & Motivation
- Methodology
- Characterization of Mobile Ad Personalization
- Privacy Leakage through Personalized Mobile Ads
- Discussion

Personal Information of Interest

- Interest Profile
  - {Music, Games, Sports, …}
- Demographics
  - Age, Gender, Education, Income, Ethnicity, Political Affiliation, Religion, Marital Status, Parental Status



Reach people of specific demographics

With demographic targeting in AdWords, you can reach customers who are likely to be within the demographic groups that you choose. Demographic groups that you can choose from include:

- age ("18-24," "25-34," "35-44," "45-54," "55-64," "65 or more," and "Unknown")
- gender ("Female," "Male," and "Unknown")
- parental status ("Parent," "Not a parent," and "Unknown")

https://support.google.com/adwords/answer/2580383?hl=en

Challenges and Our Approaches

- Triggering personalization based on target attributes of our interest
  - Using synthetic user profile is circular
    - Does ad network know users’ gender? ->
    - (We do not know how ad network knows users’ gender ->)
    - Let us build profiles for male and female users ->
    - Observation: Ads are not correlated with “gender” ->
    - Ad network does not use / know users’ gender. Really???
  - Our approach: Using profiles of real users

Challenges and Our Approaches (cont.)

- Isolating personalization from other target attributes
  - Many attributes may affect ad personalization
    - App developers could provide target attributes through ad library APIs
    - Ads may be personalized based on user's geolocation
  - Our approach: Collecting data in an isolated app

Ad Collection

- Our “Mobile Ad Study” app
  - Connects user’s device to our VPN server (Isolating geolocation)
  - Serves Google AdMob ads only
  - Provides no target attributes through ad library API (Isolating other information, not including device information that ad library can access)
  - Collects the list of installed apps that include Google AdMob SDK

Subject Recruitment

- Human Intelligence Task on Amazon Mechanical Turk
    - Complete questionnaire regarding participant’s interests and demographic information
    - Use our data collection app to load 100 ads from Google AdMob
- We collected 217 valid responses from 284 participants

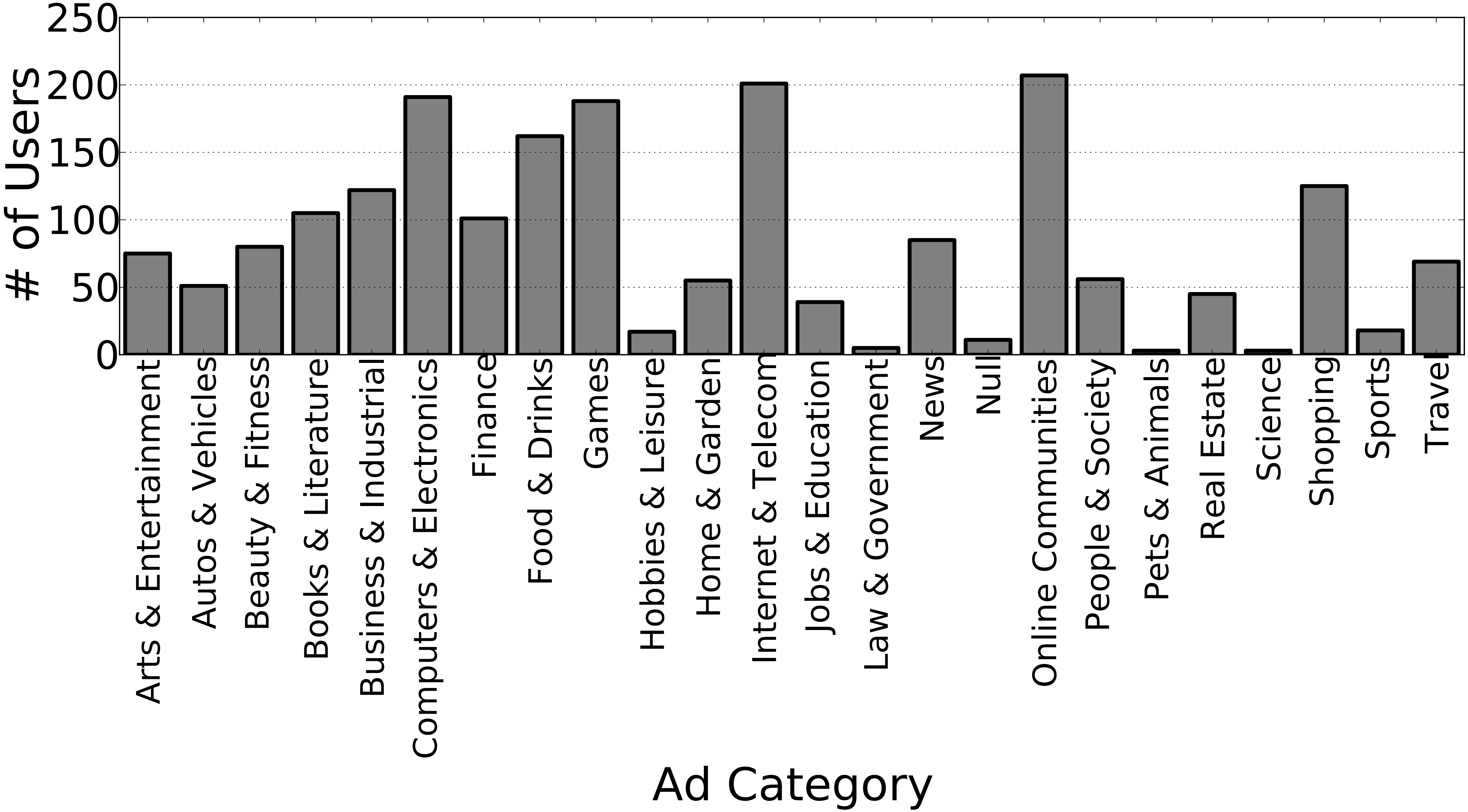
Subject Distribution

Gender		Political Affiliation			Parental Status		Income		
Female	Male	Independent	Democrat	Republican	Not a parent	Parent	< $30K	$30K-$60K	> $60K
95 43.78%	122 56.22%	108 49.77%	80 36.87%	29 13.36%	128 58.99%	89 41.01%	107 49.31%	67 30.87%	43 19.82%

Religion			Marital Status			Education			
Atheist	Non-Christian	Christian	Single	Married	Separated	High school	Associates	Bachelor	Master & Doctoral
83 37.79%	47 21.66%	88 40.55%	124 57.14%	73 33.64%	20 9.22%	78 35.94%	50 23.04%	71 32.72%	18 8.30%

Age					Ethnicity				
18-24	25-34	35-44	45-54	55+	Other	Hispanic	Asian	African American	Caucasian
45 20.74%	106 48.85%	47 21.66%	14 6.45%	5 2.30%	8 3.69%	12 5.53%	12 5.53%	23 10.60%	162 74.65%

Subject distribution (cont.)

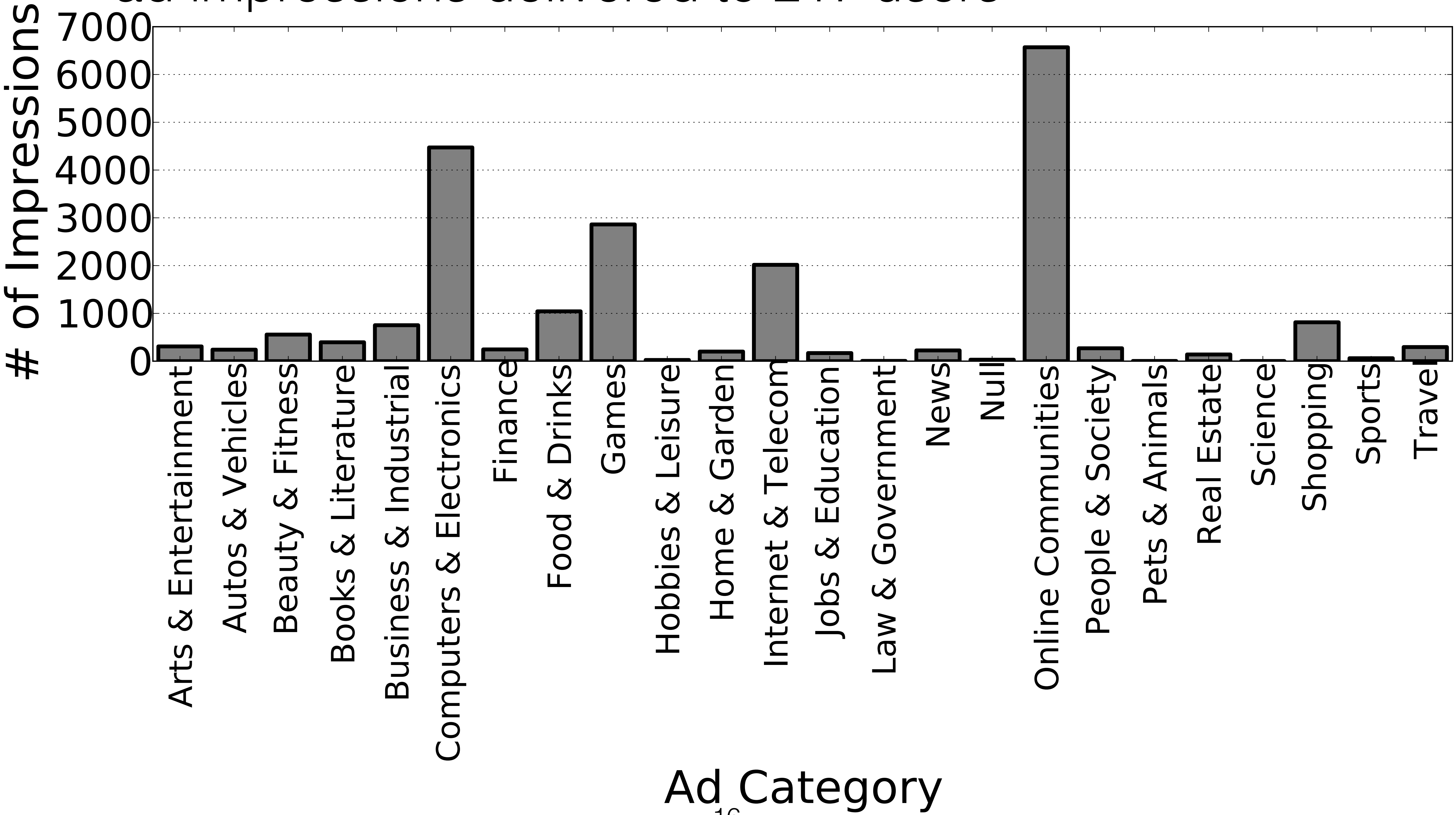


Outline

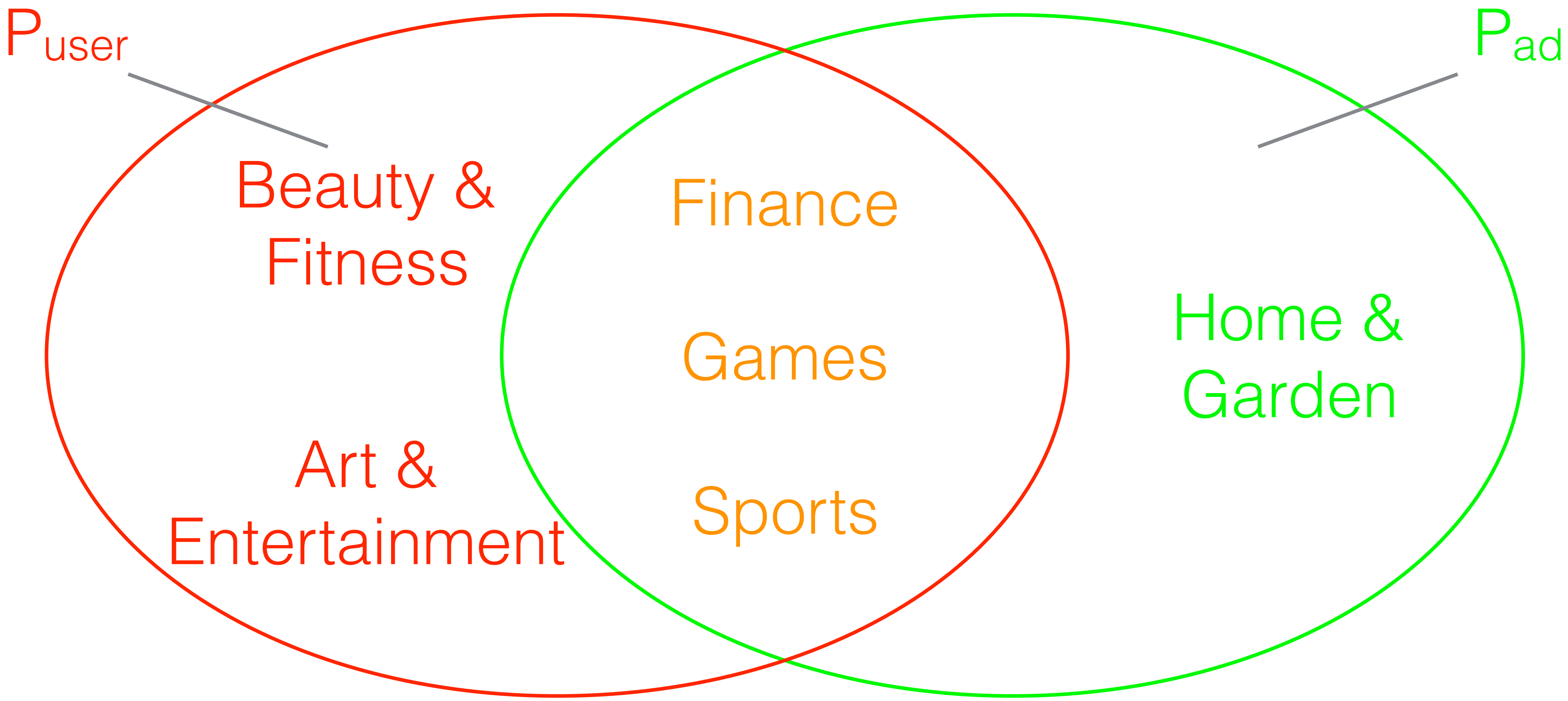
- Background & Motivation
- Methodology
- Characterization of Mobile Ad Personalization
- Privacy Leakage through Personalized Mobile Ads
- Discussion

Dataset

- We collected 695 unique ads which resulted in 39,671 ad impressions delivered to 217 users

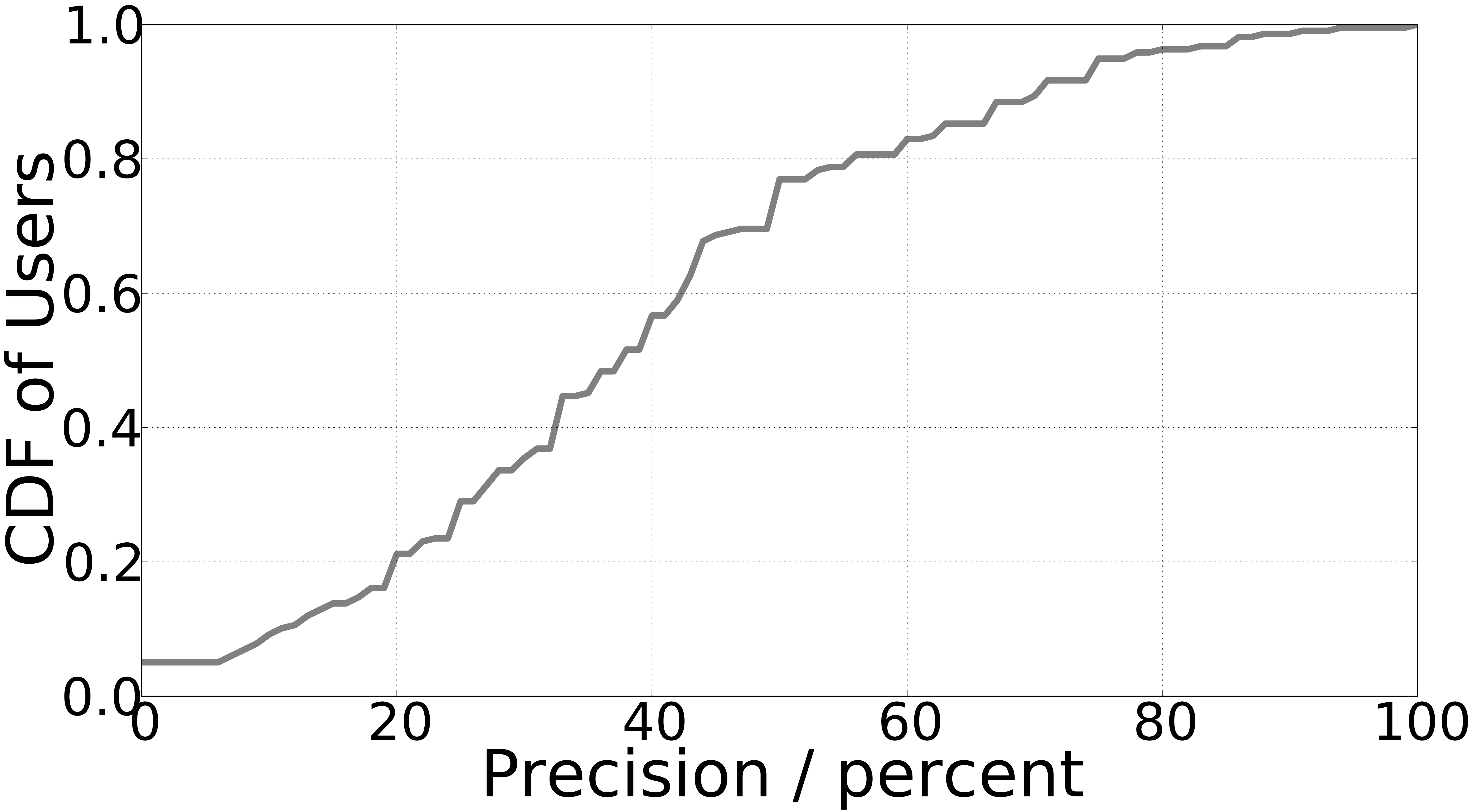


Interest Profile Based Personalization

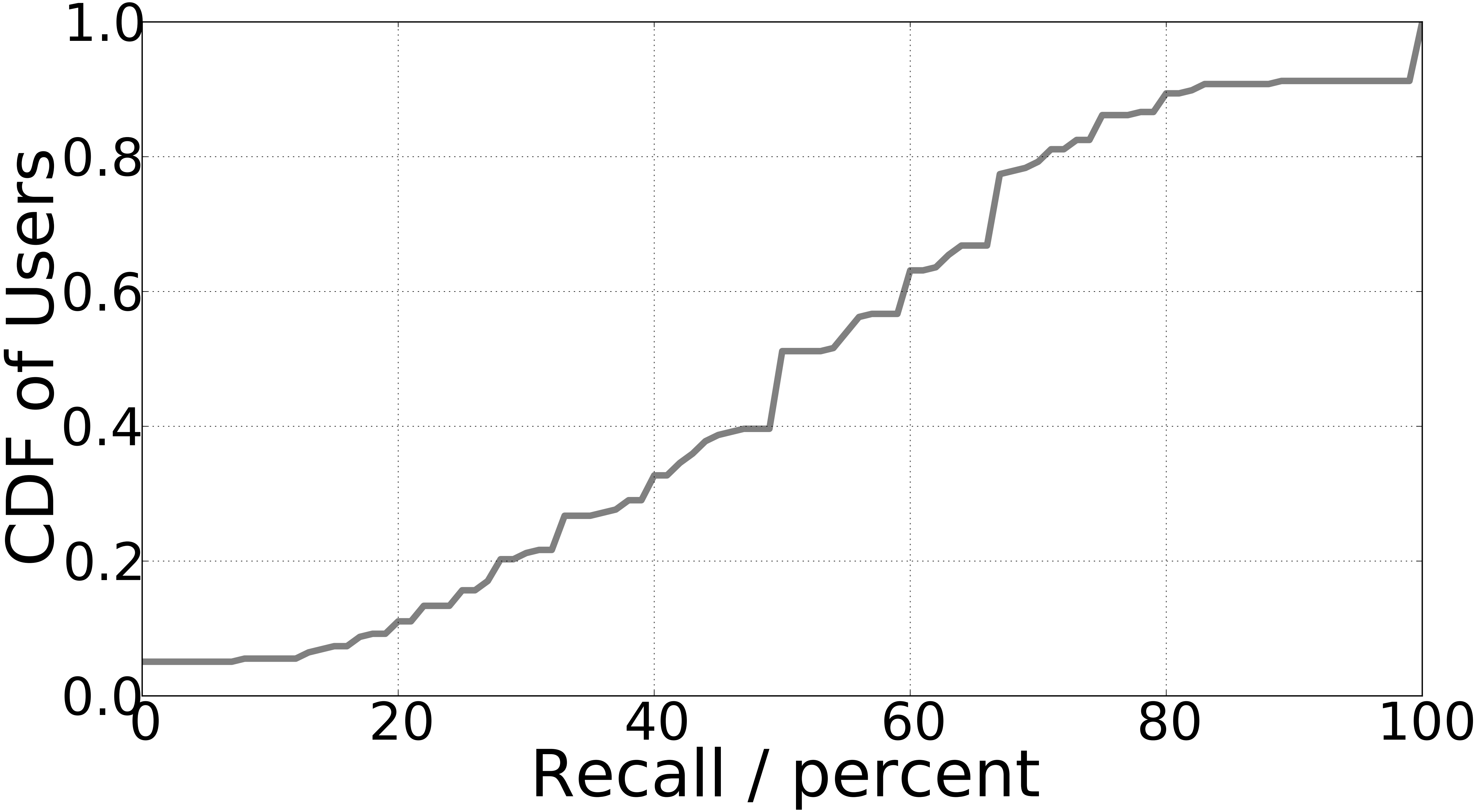


- Precision: $|P_{user} \cap P_{ad}| / |P_{ad}|$
- Recall: $|P_{user} \cap P_{ad}| / |P_{user}|$

Interest Profile Based Personalization - Precision



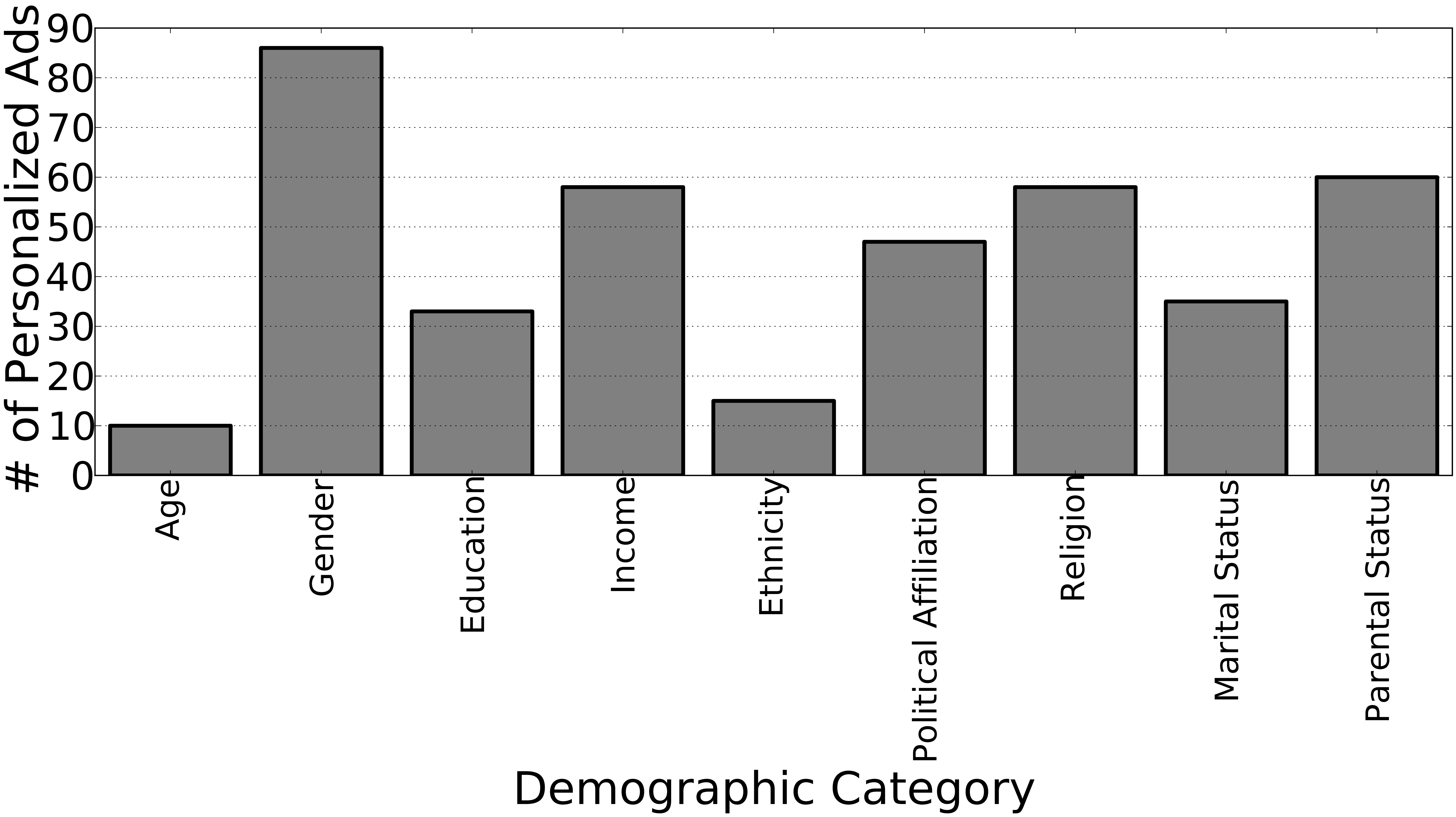
Interest Profile Based Personalization - Recall



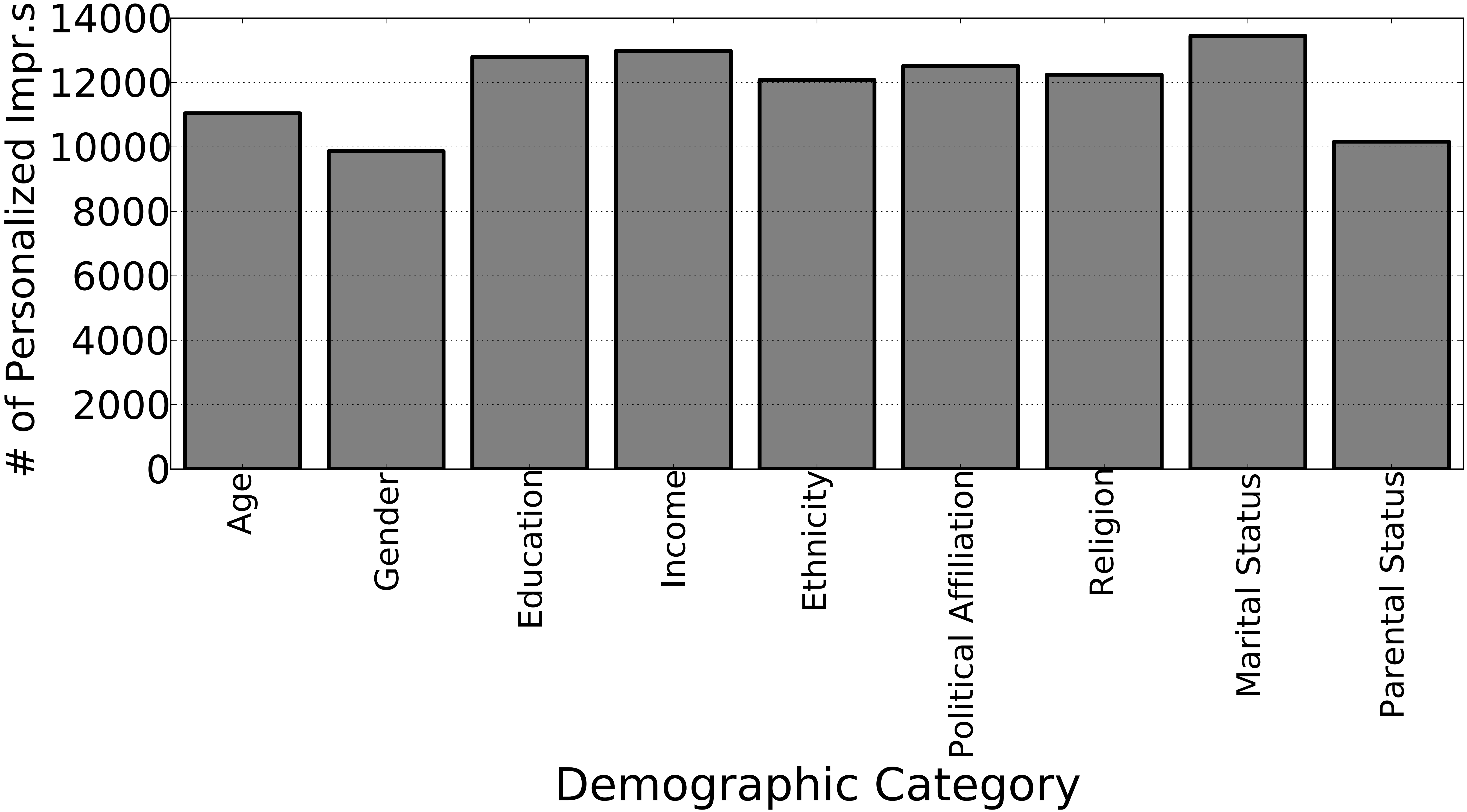
Demographics Based Personalization

- We clustered users into different demographic groups
- We tested the independence of ads and each demographic category
    - Pearson’s chi-squared test of independence
    - Null hypothesis: ad is independent of a demographic category
    - Significance level (P-value): 0.005
    - An ad is “personalized” based on the demographic category under test if null hypothesis is rejected

Demographics Based Personalization - Unique Ads



Demographics Based Personalization - Ad Impressions



Summary

- Both interest profile based personalization and demographics based personalization were prevalent in mobile in-app advertising

Outline

- Background & Motivation
- Methodology
- Characterization of Mobile Ad Personalization
- Privacy Leakage through Personalized Mobile Ads
- Discussion

Classification Models of Demographic Information

- Features
  - Number of impressions of ads that are correlated with each demographic category
  - List of installed app that include Google AdMob SDK
- Evaluation
  - 217 samples were randomly divided into 5 sets for 5-fold cross validation
- Metric for evaluating severity of privacy leakage
  - Cross validated accuracy (mean of accuracies of the 5 validations)
  - Adversary cannot have significant better accuracy than that obtained from tossing coins in a perfectly privacy-preserving system

Baseline Classifiers

- Dummy
  - Assumption: samples are evenly distributed across labels
  - Predicts any possible label with same probability
- Augmented Dummy
  - Assumption: samples are not evenly distributed
  - Knows the population distribution in prior
  - Always predicts the most popular label

Regrouping Subjects

- Observation: Samples were not evenly distributed across all labels

Gender		Political Affiliation		Parental Status		Income	
Female	Male	Independent	**Non-Independent**	Not a parent	Parent	< $30K	**> $30K**
95 43.78%	122 56.22%	108 49.77%	109 50.23%	128 58.99%	89 41.01%	107 49.31%	110 50.69%

Religion			Marital Status		Education		
Atheist	Non-Christian	Christian	Single	**Not Single**	High school	Associates	**Bachelor or higher**
83 37.79%	47 21.66%	88 40.55%	124 57.14%	93 42.86%	78 35.94%	50 23.04%	89 41.02%

Age			Ethnicity				
18-27	**28-33**	**34+**	Other	Hispanic	Asian	African American	Caucasian
71 32.72%	71 32.72%	75 34.56%	8 3.69%	12 5.53%	12 5.53%	23 10.60%	162 74.65%

Evaluation Result

	Age	Education	Ethnicity	Gender	Income
Best	0.54	0.40	0.76	0.74	0.62
Dummy	0.33	0.33	0.20	0.50	0.50
Augmented Dummy	0.35	0.41	0.75	0.56	0.51

	Marital Status	Parental Status	Political Affiliation	Religion
Best	0.63	0.66	0.59	0.43
Dummy	0.50	0.50	0.50	0.33
Augmented Dummy	0.57	0.59	0.50	0.41

Outline

- Background & Motivation
- Methodology
- Characterization of Mobile Ad Personalization
- Privacy Leakage through Personalized Mobile Ads
- Discussion

Privacy Implication

- In Android, host app can observe all personalized ads
- Ad network may be inadvertently leaking some of its collected user information (Age, Gender, Parental Status) to the app developer
- Adversary also has non-trivial advantage in predicting other aspects of the user’s demographics
    - These aspects may be correlated with those collected and used by ad networks

Limitation

- The size of our dataset is small

- More aggressive adversaries may achieve significant better result

    - They can invest more resources to obtain better ground truth data

    - They can observe ads received by users for a longer period of time

Countermeasures

- Root cause of the privacy leakage problem: lack of isolation between ads and host apps
    - Adopting HTTPS will not stop the problem
- We really need isolation between ads and host apps
- What can ad networks do?
    - Adding noise into personalized results
    - Providing coarser-grained targeting options

Summary

- We collected both the profile and observed mobile ad traffic from 217 real users
- We studied ad personalization based on real users' interest profiles and demographics
- We demonstrated that personalized in-app advertising can leak potentially sensitive information to any app that hosts ads

Thank you!

Q & A