Selling Off Privacy at Auctions
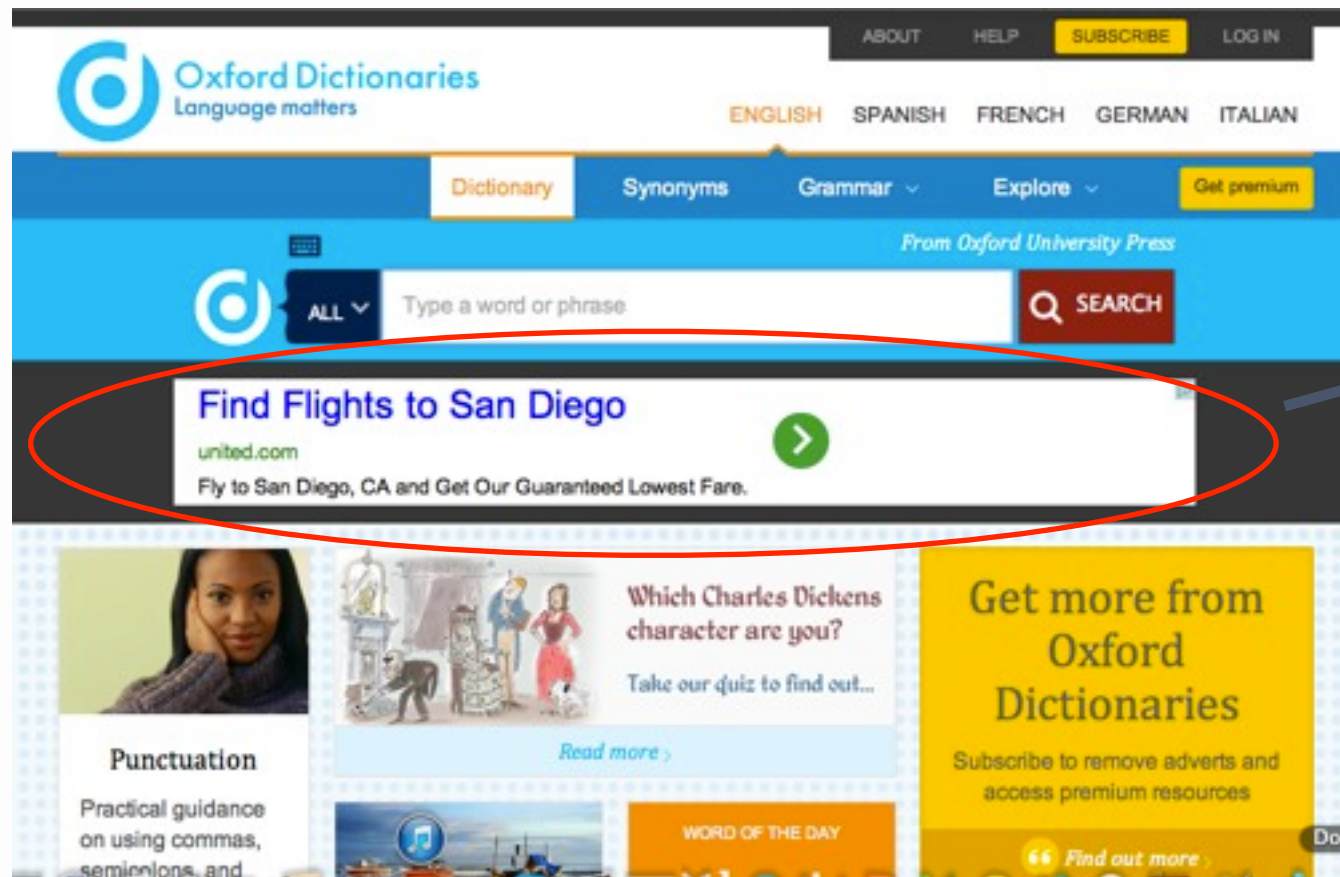
Lukasz OLEJNIK, **Minh-Dung TRAN**, Claude CASTELLUCCIA

Online Advertising

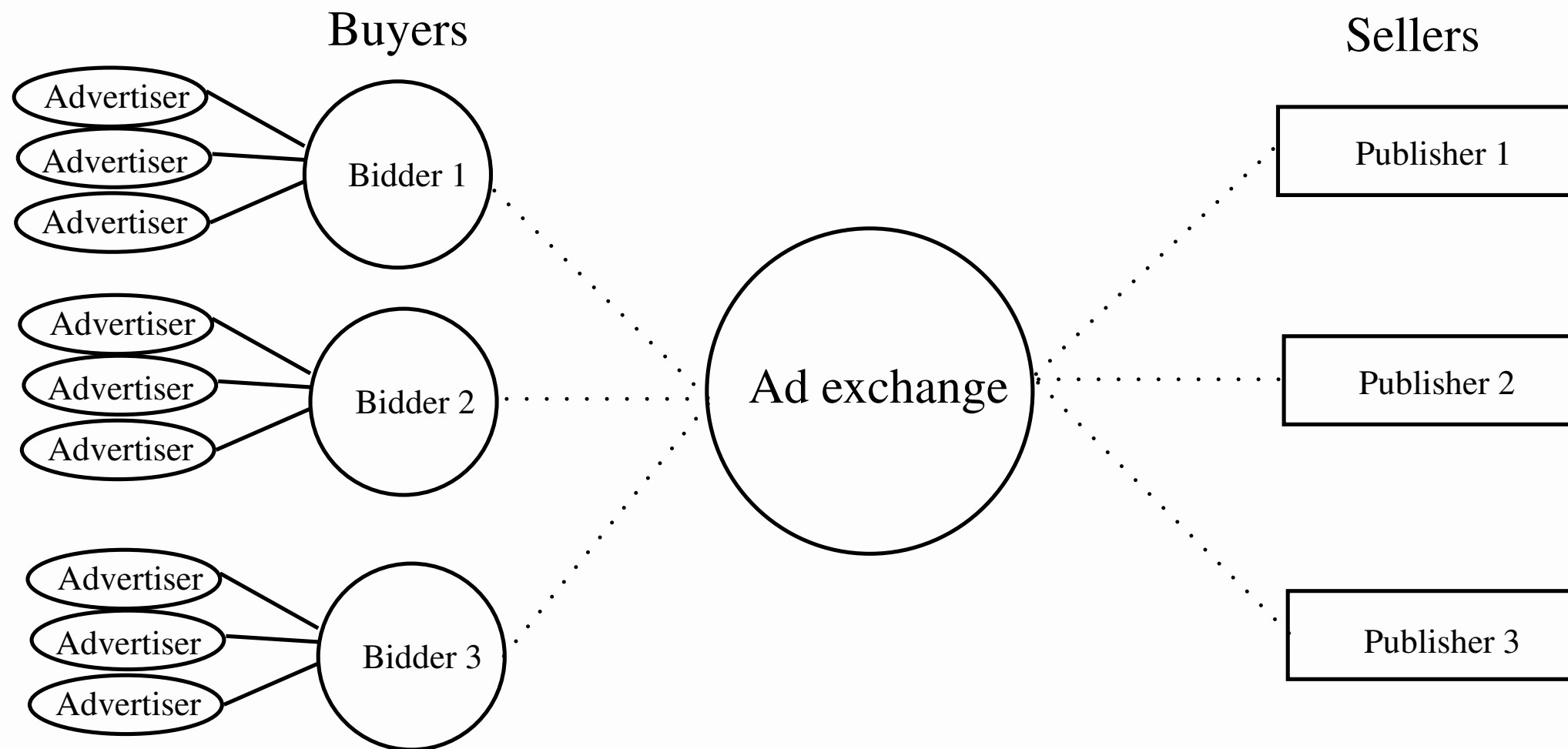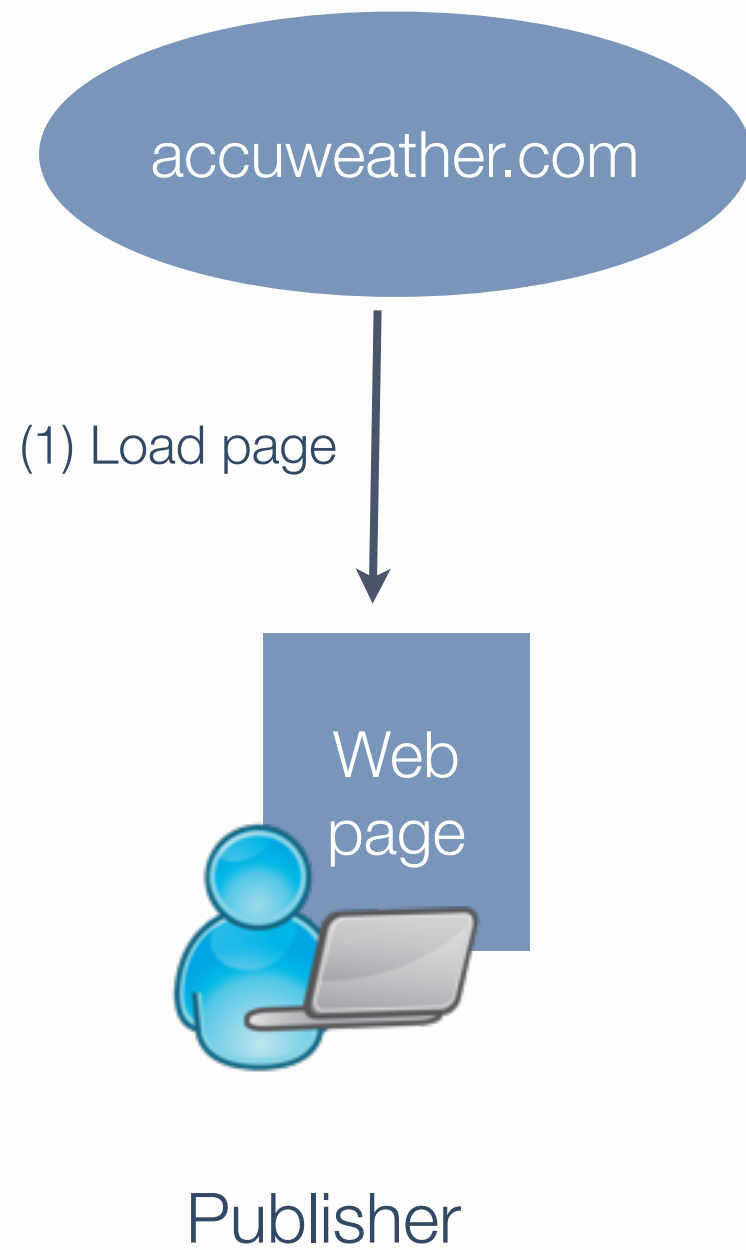Ad spaces

$$$$

Ad sellers (Publishers)

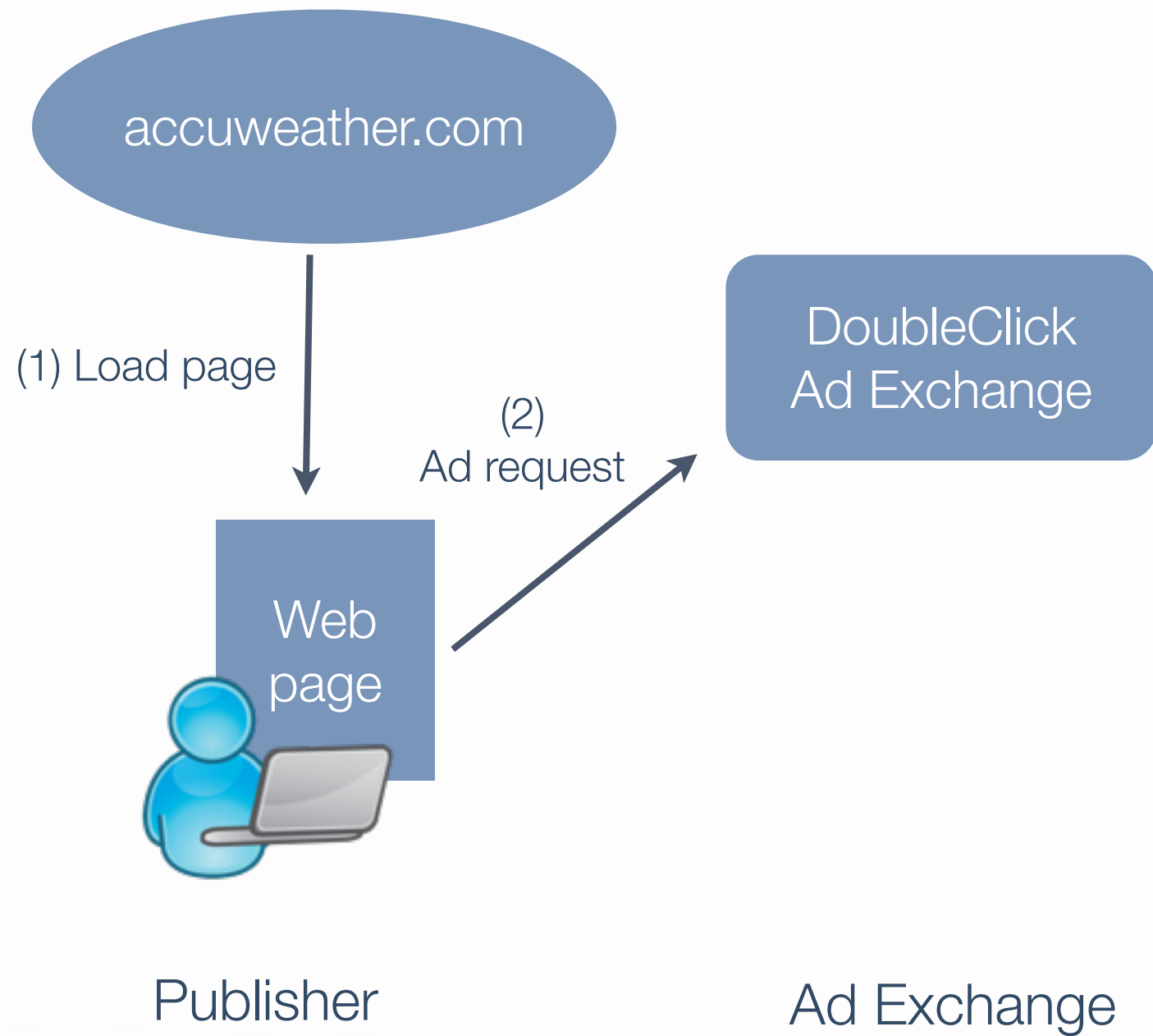Ad buyers (Advertisers)

Real-Time Bidding (RTB)

✦ RTB allows ad buyers (advertisers) and sellers (publishers) buy and sell ads in real time on a *per-ad-impression* basis through an Ad Exchange.

RTB Protocol

accuweather.com

(1) Load page

Web
page

Publisher

RTB Protocol

accuweather.com

(1) Load page

(2)
Ad request

Web
page

DoubleClick
Ad Exchange

Publisher

Ad Exchange

RTB Protocol

RTB Protocol

accuweather.com

(1) Load page

(2) Ad request

Web page

DoubleClick Ad Exchange

(3) Bid Request: User Information + Page Information

Criteo

(4) Bid response

Turn

...

Publisher

Ad Exchange

Bidders

RTB Protocol

accuweather.com

(1) Load page

(2)
Ad request

Web
page

DoubleClick
Ad Exchange

(3) Bid Request:
User Information +
Page Information

Criteo

(4) Bid response

Turn

...

Publisher

Ad Exchange

Bidders

RTB Protocol

accuweather.com

(1) Load page

(2)
Ad request

Web
page

(5) Ads

DoubleClick
Ad Exchange

(3) Bid Request:
User Information +
Page Information

Criteo

(4) Bid response

Turn

...

Publisher

Ad Exchange

Bidders

Cookie Matching

Bid request
ADX's cookie
Visiting URL

ADX ——————▶ **Bidder**

User profile (at Bidder's DB)	
Bidder's cookie	URL1
	URL2
	URL3
	....

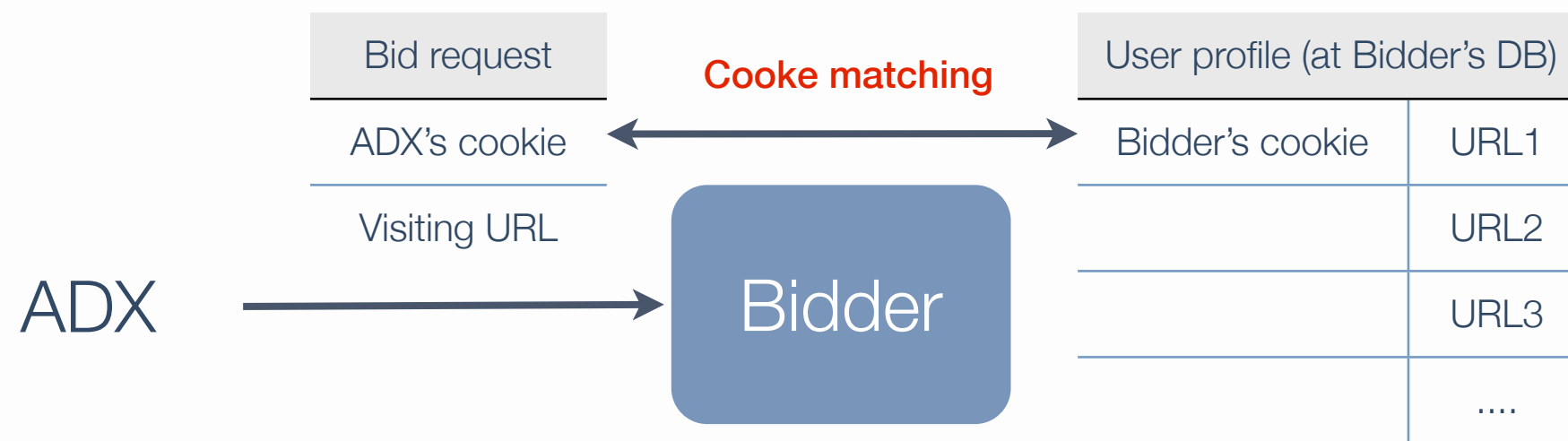# Cookie Matching

✤ CM helps bidders recognize users at bid requests:

RTB in Online Advertising

✤ RTB is rapidly growing, expectedly accounting for:

✤ 27% of total display advertising sales in the US, 25% in United Kingdom, 21% in France, 20% in Germany.

✤ 65% of indirect display ad sales revenue in the US and 55% in the most developed European markets.

...by 2015! (IDC, 2011)

✤ And privacy?

Product in RTB auction

User's presence
on a website

User X is now on website Y.
Are you interested?

Is it worth showing ads to X?
At what price?

Bid request

Bidder

Bid response
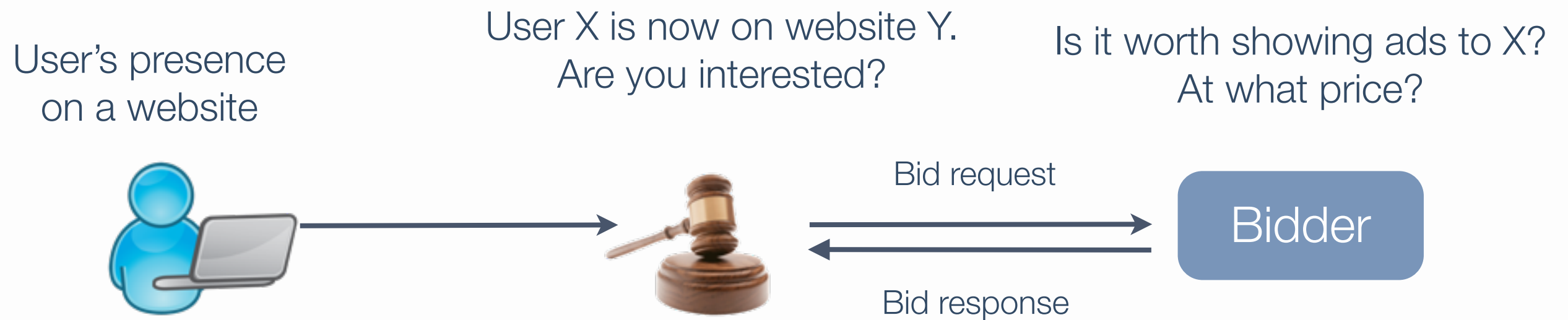
RTB and CM in Privacy Perspective

✤ Privacy problems in RTB:

✤ User data leakage? (**RTB Privacy Analysis**)

✤ User profiles can be linked among companies thanks to CM? (**CM Privacy Analysis**)

✤ Users are evaluated differently? (**Real-Time Price Analysis**)

"The first impression seen by a high-value person on the opening page of a major newspaper first thing in the morning has a different value than a user from China who is 12 and has been on the Web all day long playing games," says Frank Addante, the founder and chief executive of Rubicon.

CM and RTB Detection

CM Detection technique

✤ CM mechanism:

User's cookie/id: aaa

User's cookie: bbb

Ad Exchange
cm.adexchange.com

(3) Matches cookies
"aaa" with "bbb"

Bidder
cm.bidder.com

(1) Script or redirect instruction to load
"http://cm.bidder.com/cmservice?ExchangeUserId=aaa"

(2) "http://cm.bidder.com/cmservice?ExchangeUserId=aaa"
Cookie: UID = bbb
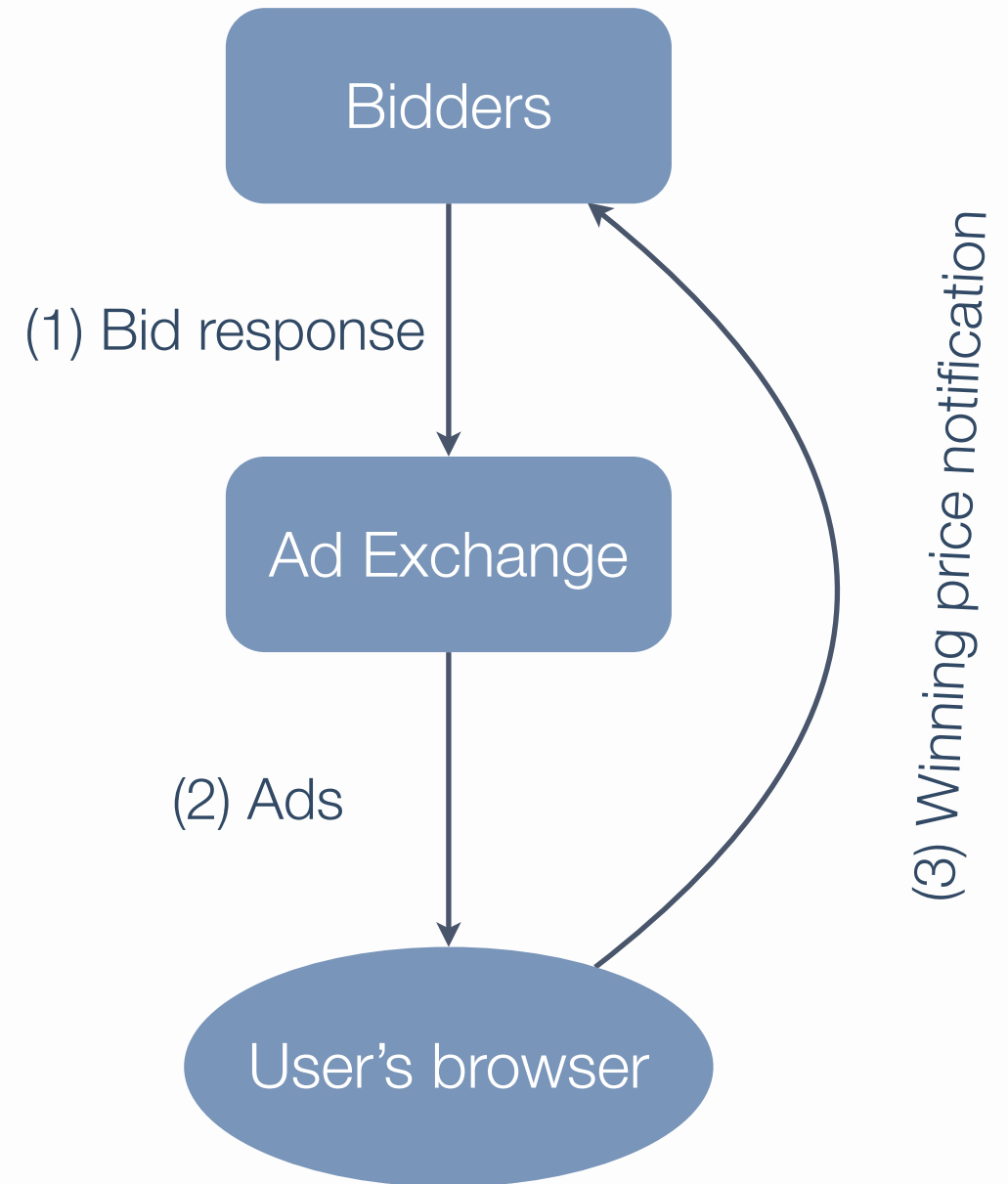
User's browser

ADX's cookie: aaa
Bidder's cookie: bbb
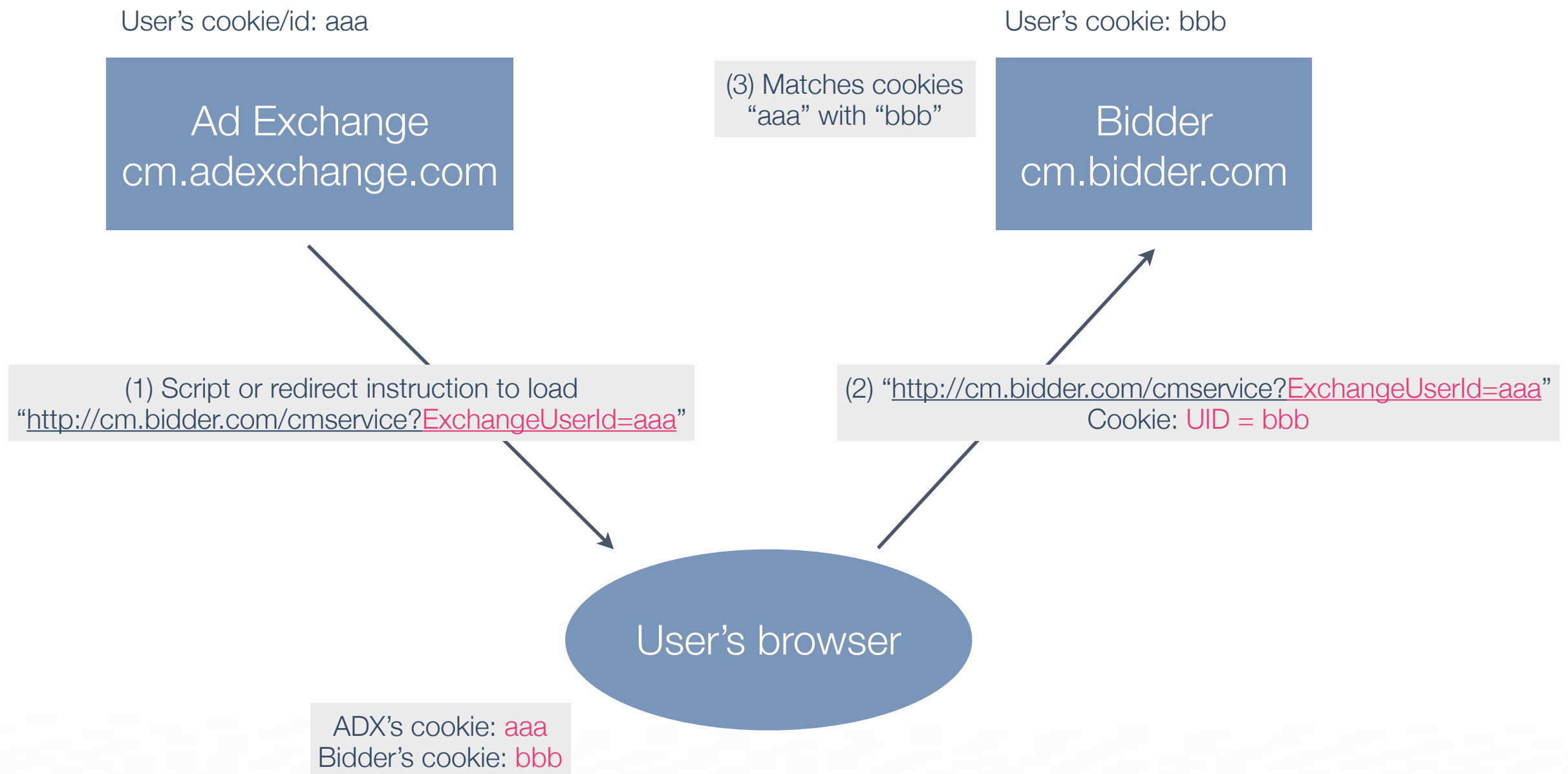
CM Detection Technique

* **Step 1:** Detect all casual relationships between HTTP requests, i.e. A -> B

  * Based on HTTP header's Referer & Location, Javascript...

* **Step 2:** Scan all cookies from A's response and all HTTP parameters included in B to detect cookie matching.

RTB Detection Technique

- Based on **winning price** detection:

  - Winning price is notified to the winning bidder by ADX

  - The winning price notifications happen through user browser in many cases. Example: *"http://invitemedia.com?cost=[WinningPrice]"*

  - The winning price is detectable in its encrypted form, which contains a timestamp **close to the real time** in the first 8 bytes.

Bidders

(1) Bid response

Ad Exchange

(2) Ads

User's browser

(3) Winning price notification

Clear-Text Price Detection

* Many companies use **clear-text prices** in winning price notification

  * http://invitemedia.com?cost=U2VuZCByZWluZm9yY2VtZW50cwHgW5uY5GtM4t&...

  * http://invitemedia.com?cost=0.5&.....

* The values are very often floating-point numbers or integers in micro formats.

* Many parameters are contextual and meaningful: "win_price", "cost", "price", "rtbwinprice".....

13

CM and RTB Privacy Analysis

Methodology
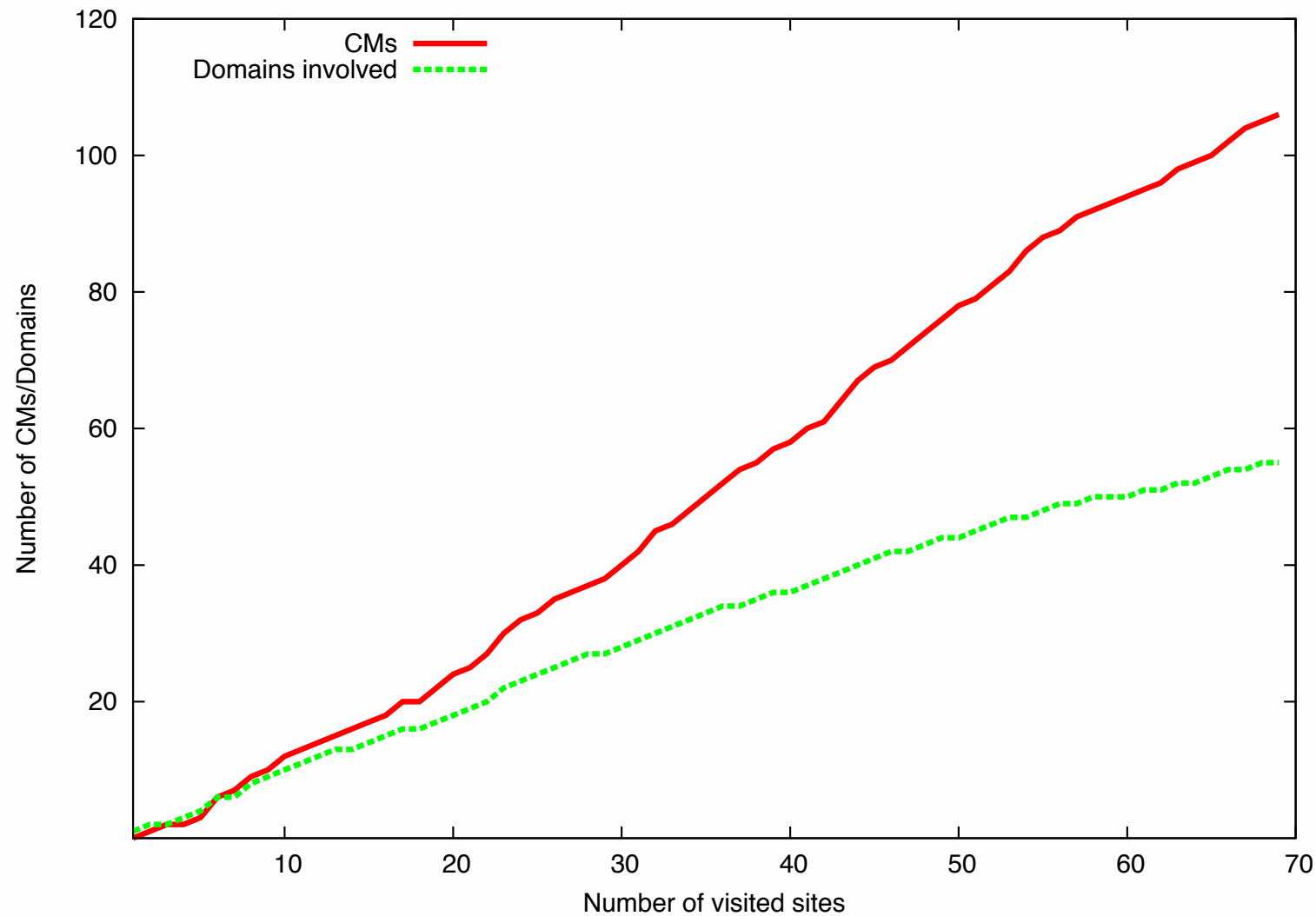
* Tools

  * RTBAnalyzer: Firefox plugin implementing CM and RTB detection techniques

* Dataset:

  * 100 volunteers using RTBAnalyzer during several days

  * **User data is anonymized**

CM Privacy Analysis

Company A   Company B

	CM		Combine	
Cookie 1	⟷	Cookie 2	→	Cookie 1 & 2
URL1		URL4		URL1
URL2		URL5		URL2
URL3		URL6		URL3
....		....		URL4
				URL5
				URL6
				.....
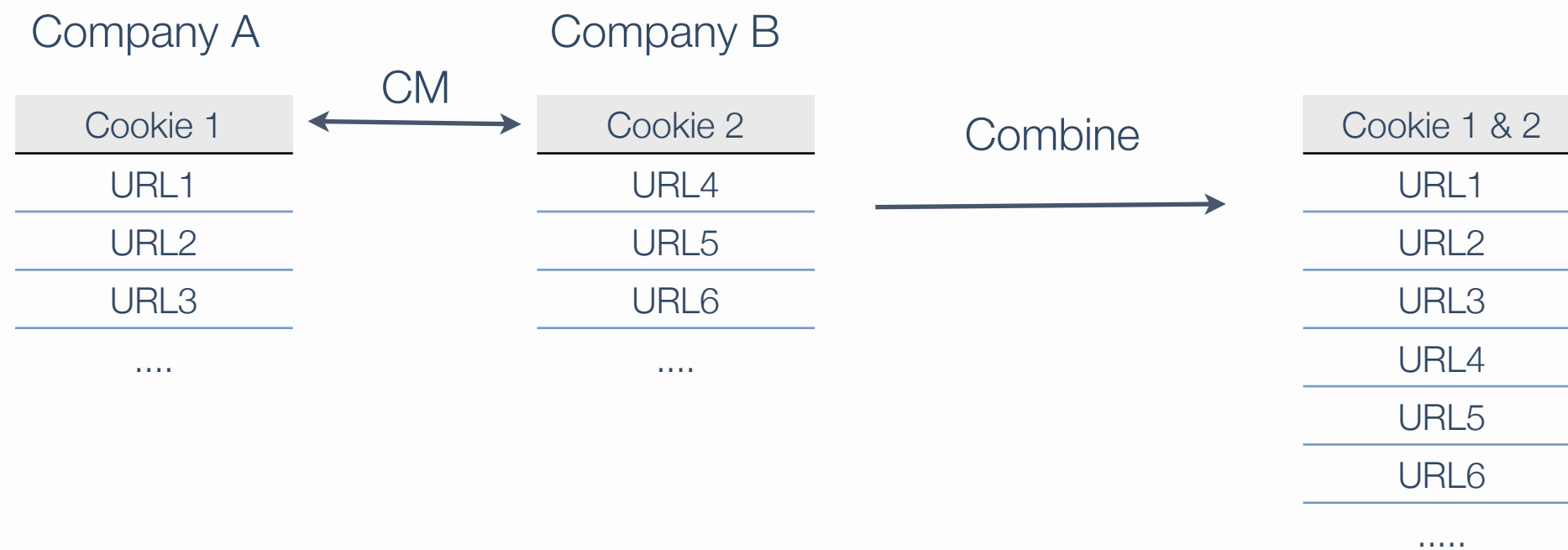
What is the extent of such potential combination?

CM Privacy Analysis

CM prevalence

| Pair of domains	CM frequency
facebook - adnxs.com	91%
turn.com - admeld.com	87%
doubleclick.net - rfihub.com	86%
doubleclick.net - adnxs.com	85%
doubleclick.net - mathtag.com	85%
adnxs.com - admeld.com	84%
doubleclick.net - turn.com	80%
doubleclick.net - yieldmanager.com	77%
invitemedia.com - admeld.com	73%
mathtag.com - admeld.com	71%
.....	......

Top pairs of domains
peforming CM the most

CM Privacy Analysis

User profile

Facebook | AppNexus
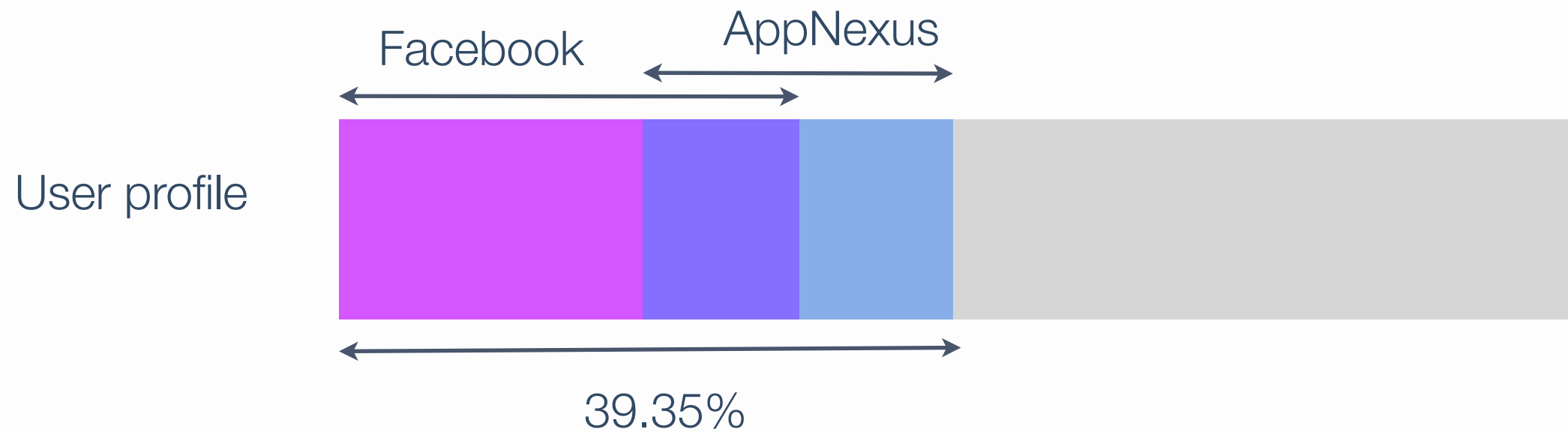
39.35%

| Pair of domains	Profile combination
doubleclick.net - adnxs.com	52.43%
doubleclick.net - yieldmanager.com	52.01%
facebook.com - adnxs.com	39.35%
....	....

RTB Privacy Analysis
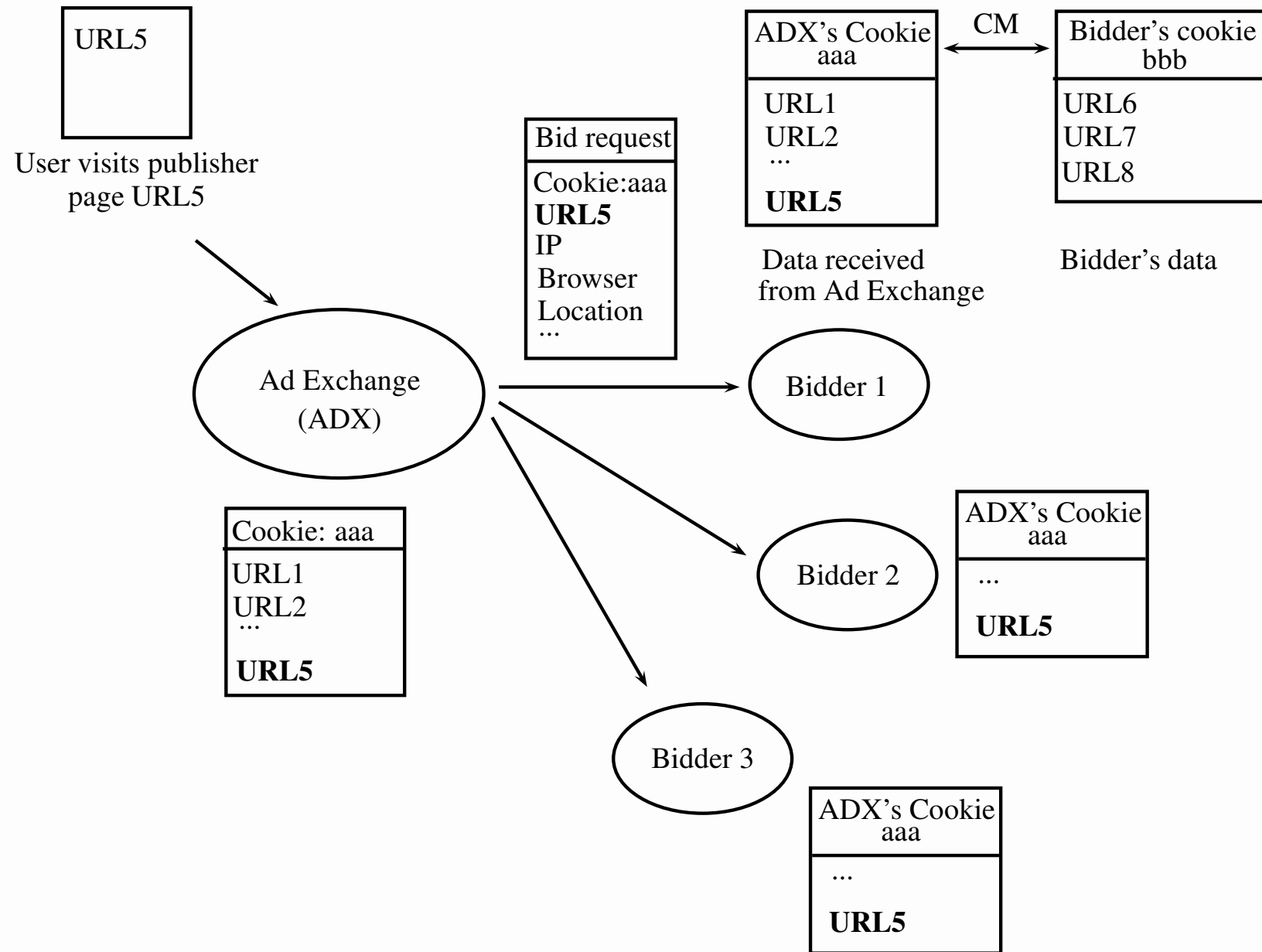
To which extent a user history could be leaked through RTB?

Example of Google's bid request

```
id: "Mv\2005\000\017.\001\n\345\177\307X\200M8"
ip: "\314j\310"
user_agent: "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US) AppleWebKit/534.13 (KHTML, like Gecko) Chrome/
9.0.597.107 Safari/534.13,gzip"
url: "http://www.example.com/"
detected_language: "en"
detected_vertical {
  id: 22
  weight: 0.67789277
}
detected_vertical {
  id: 355
  weight: 0.32210726
}
cookie_version: 1
google_user_id: "CAESEIcS1pC2TBvb-4SLDjMqsY9"
seller_network_id: 1
publisher_settings_list_id: "\357\237V\206)\231\3125%|$\032\"
vertical_dictionary_version: 2
timezone_offset: -300
cookie_age_seconds: 7685804
```

RTB Privacy Analysis

The number of RTB events is about 10% of the number of visited sites.
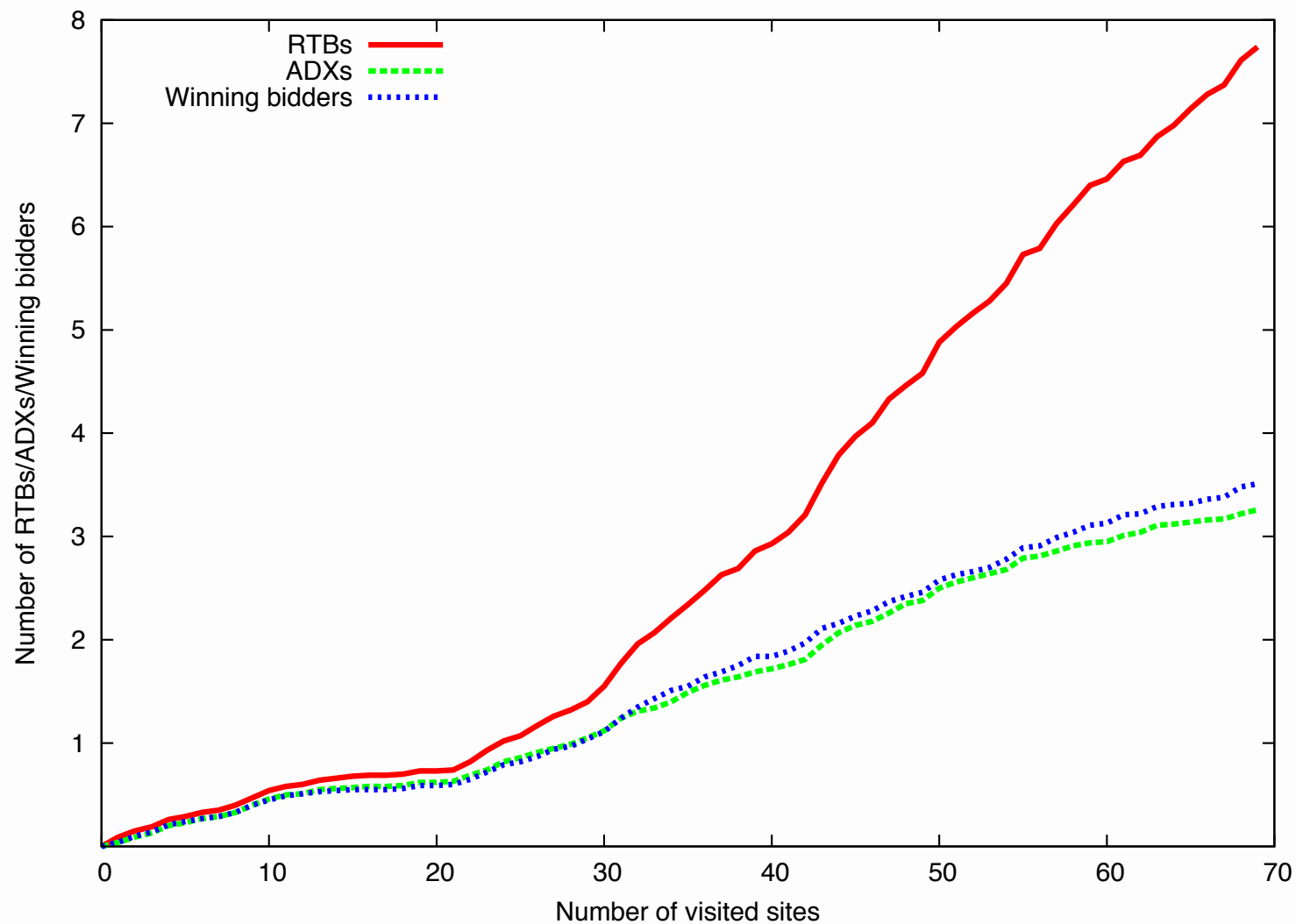
RTB Privacy Analysis
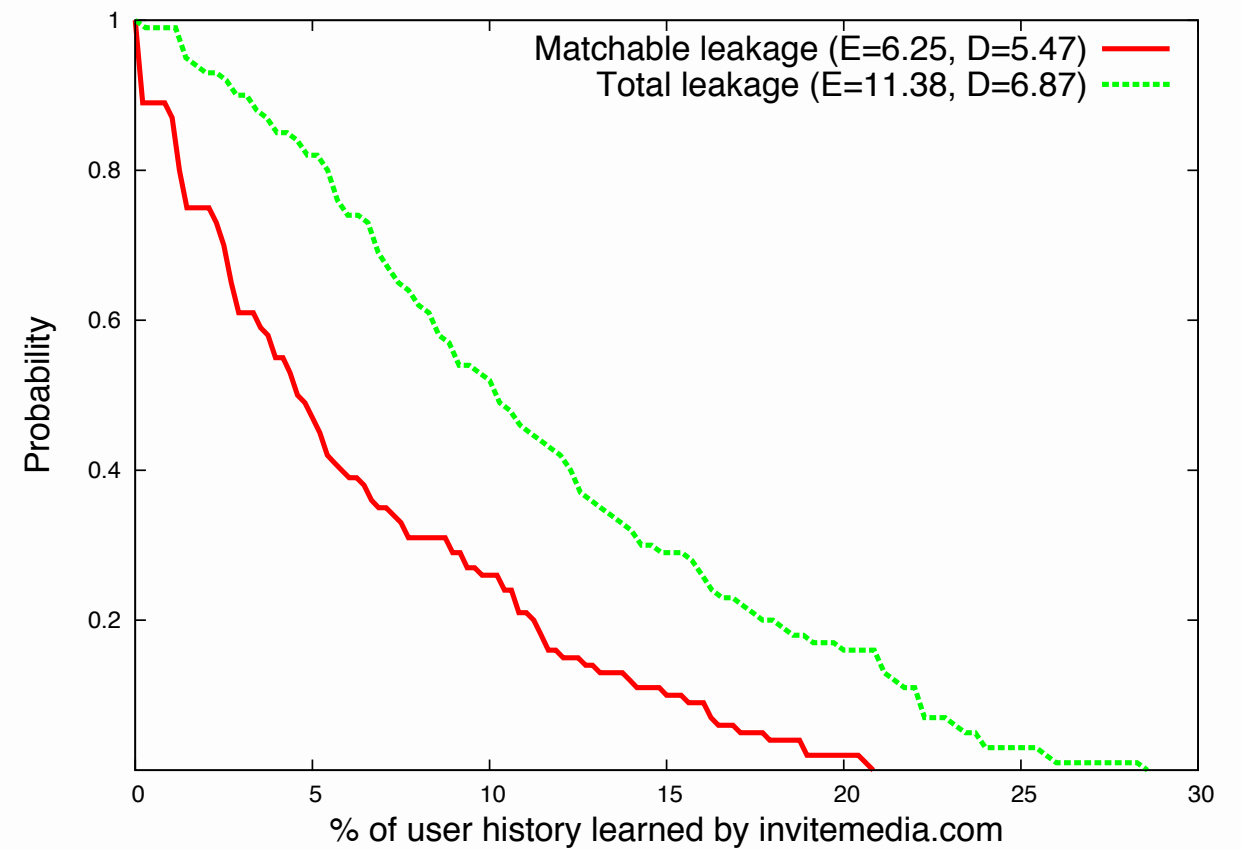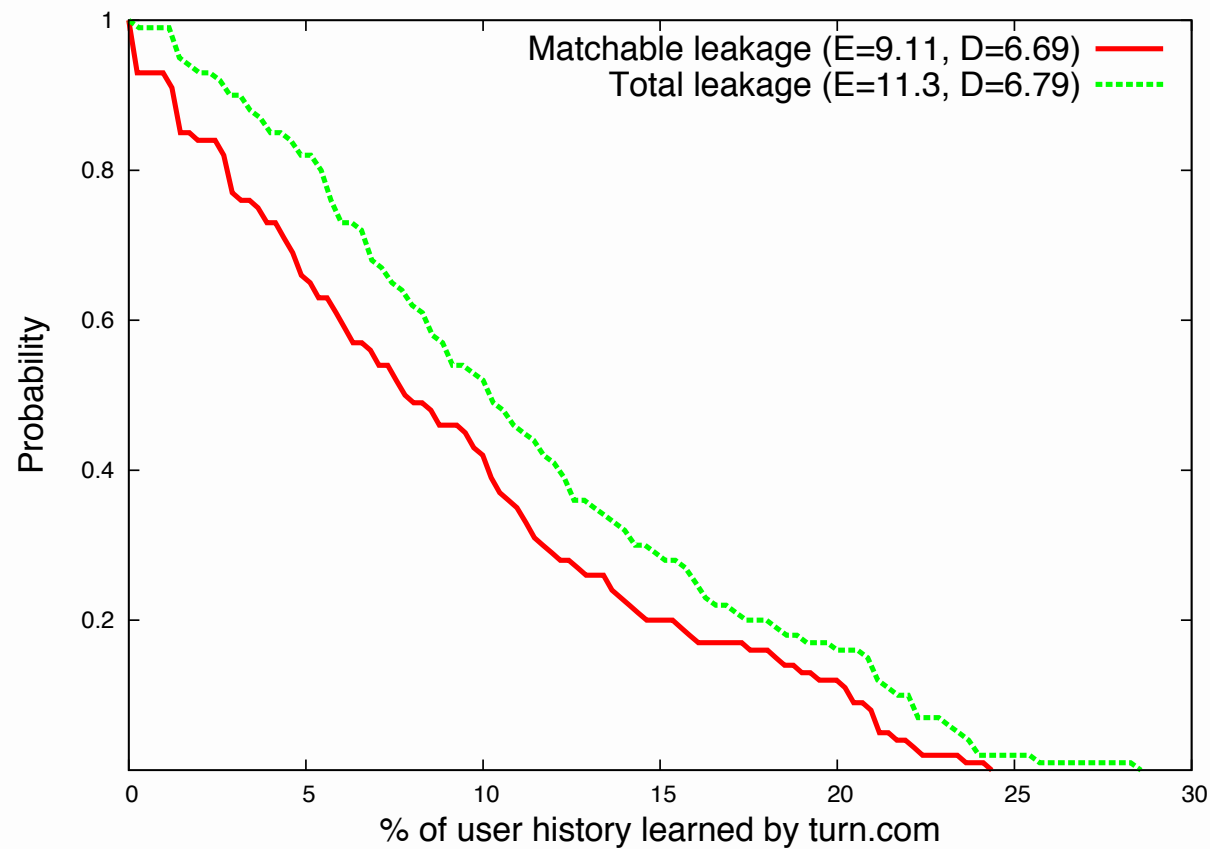
The leakages in case of these companies are about **11%** of user browsing history on average, but can be as high as **27%** with some user profiles!

Real-Time Price Analysis

Real-Time Price Analysis

* Context dependence:

  * Site, Category, Time, Location...?

* Profile dependence:

  * History category (the category of user visited sites)?

  * Intents (looking for a commercial products)?

Context Analysis

* Dataset: 5K RTB-enabled Alexa sites

* Experiments:

    * Visit sites with empty user profile

    * Different locations: France, the US, and Japan (using Planet Lab)

Context Analysis

Prices tend to be highest in the early morning.

Time division	The US	France	Japan
0 - 8h	0.75 (3,246)	0.39 (10,621)	0.28 (729)
8 - 16h	0.68 (2,772)	0.36 (11,375)	0.22 (732)
16 - 24h	0.62 (2,520)	0.31 (7,675)	0.19 (516)

Prices in the US are highest in the three analyzed countries.

Country	Average	Q1	Q2	Q3	Count
The US	0.69	0.15	0.33	1.00	8,538
France	0.36	0.11	0.24	0.47	29,671
Japan	0.24	0.04	0.07	0.22	1,977

Tuesday, February 25, 14

Context Analysis

Some site categories are more "worthy" than others.

Category	Avg. price	Std	Median
Adult / Mature Content	0.25	0.15	0.22
Humor	0.25	0.19	0.20
Sports	0.29	0.18	0.36
Games	0.32	0.16	0.42
Blogs / Web Communications	0.33	0.25	0.32
Entertainment	0.33	0.23	0.35
Streaming Media / MP3	0.36	0.27	0.42
Computers / Internet	0.38	0.24	0.38
News / Media	0.38	0.26	0.43
Society / Lifestyle	0.38	0.27	0.46
Vehicles	0.41	0.34	0.37
Reference	0.48	0.21	0.61
Restaurants / Food	0.59	0.31	0.73
Shopping	0.68	0.38	1.10

Tuesday, February 25, 14

Profile Analysis

* Dataset:
  * Top Alexa sites belonging to 14 categories (**to build user history category**)
  * Commercial products on sites maty.com, hotels.com, fnac.com (**to build user intents**)
* Experiment:
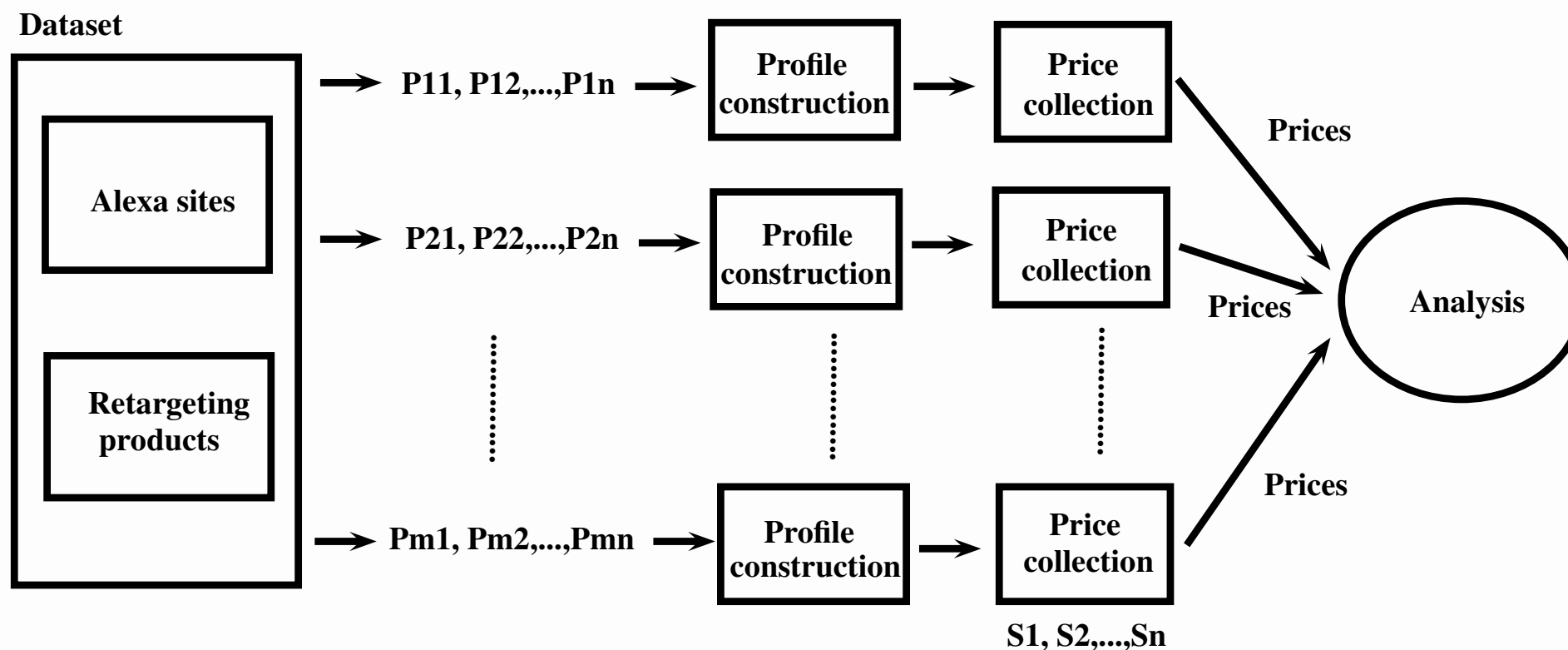
Pij is a site chosen from Alexa sites or retargeting products

S1, S2,..., Sk are the sites with prices. These sites are the same in all Price collection processes.

Profile Analysis

✤ **Users with history category** are evaluated higher than **new users**

✤ **Users with specific intents** are evaluated higher than **users with history category**

	Average	Standard Deviation
New user	0.41	0.10
Only history category	0.58	0.26
History category + products from fnac.com	0.64	0.29
History category + products from hotels.com	0.69	0.26
History category + products from maty.com	1.20	0.25
Only products from maty.com	1.17	0.26

Profile Analysis

✤ Among different "Only category" profiles, some are more valuable than others: *Games, Sport, Health, Kids and Teens*

Category	Avg. price	Standard Deviation
Adult	0.44	0.20
Arts	0.51	0.17
Business	0.55	0.22
Business - Financial Services	0.59	0.20
Computers	0.48	0.21
Games	0.80	0.35
Health	0.67	0.47
Home	0.58	0.21
Kids and Teens	0.64	0.33
News	0.50	0.12
Recreation	0.55	0.21
Science	0.50	0.19
Shopping	0.53	0.22
Sports	0.71	0.47

Tuesday, February 25, 14

Conclusion

✤ CM is prevalent and and might help companies significantly increase the size of their tracked user profiles.

✤ RTB can leak up to 27% of a user history to a bidder in auction

✤ Each user is being evaluated differently by bidders depending on their profile and visiting contexts.

✤ User's presence on a website is often sold off for less than $0.0005 (an equivalence of $0.5 CPM), much less than that from users' perspective (EUR 7)

Thank you for your attention!