

Tracing Information Flows Between Ad Exchanges Using Retargeted Ads

Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, Christo Wilson

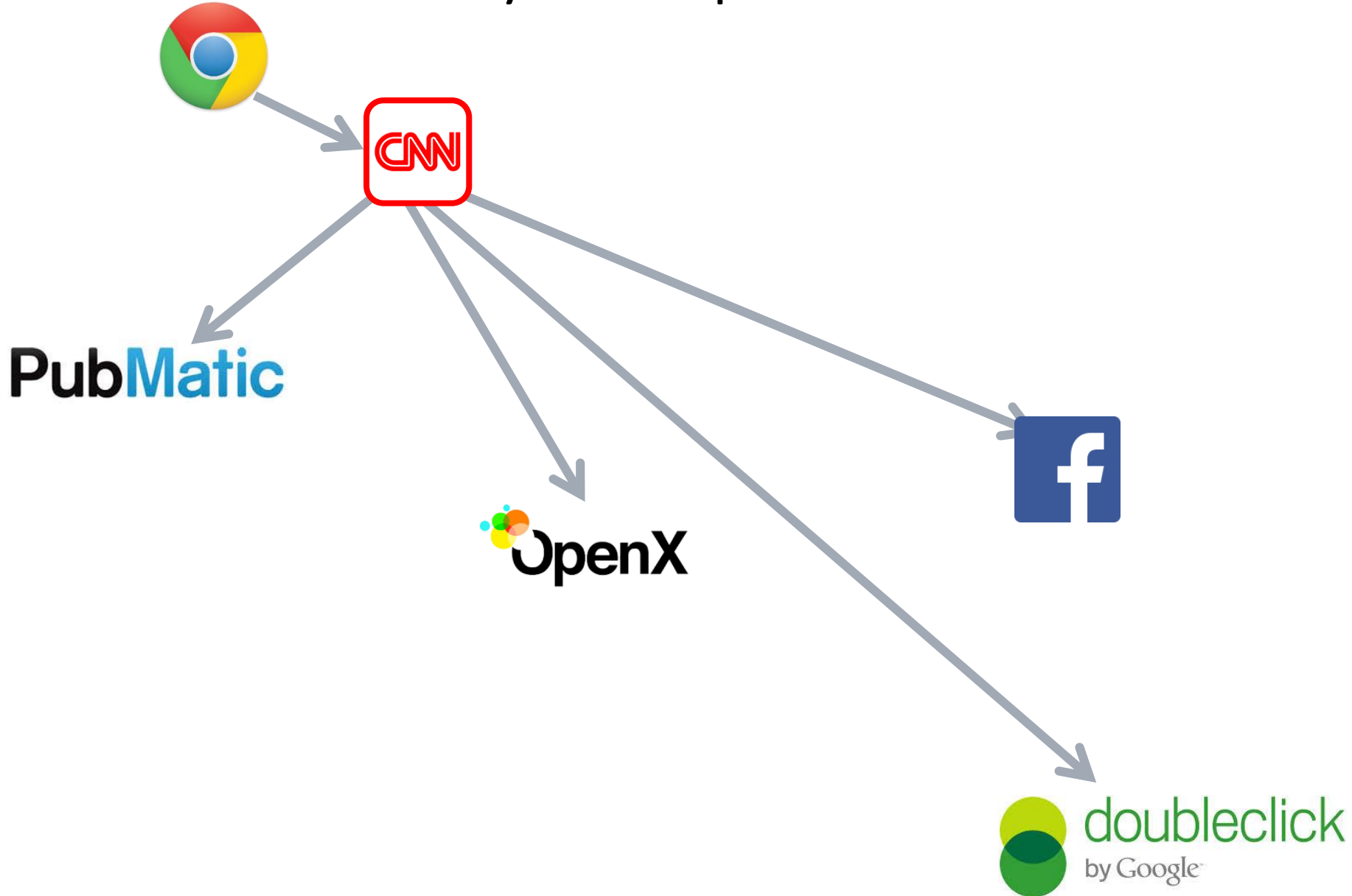
Northeastern University

Your Privacy Footprint

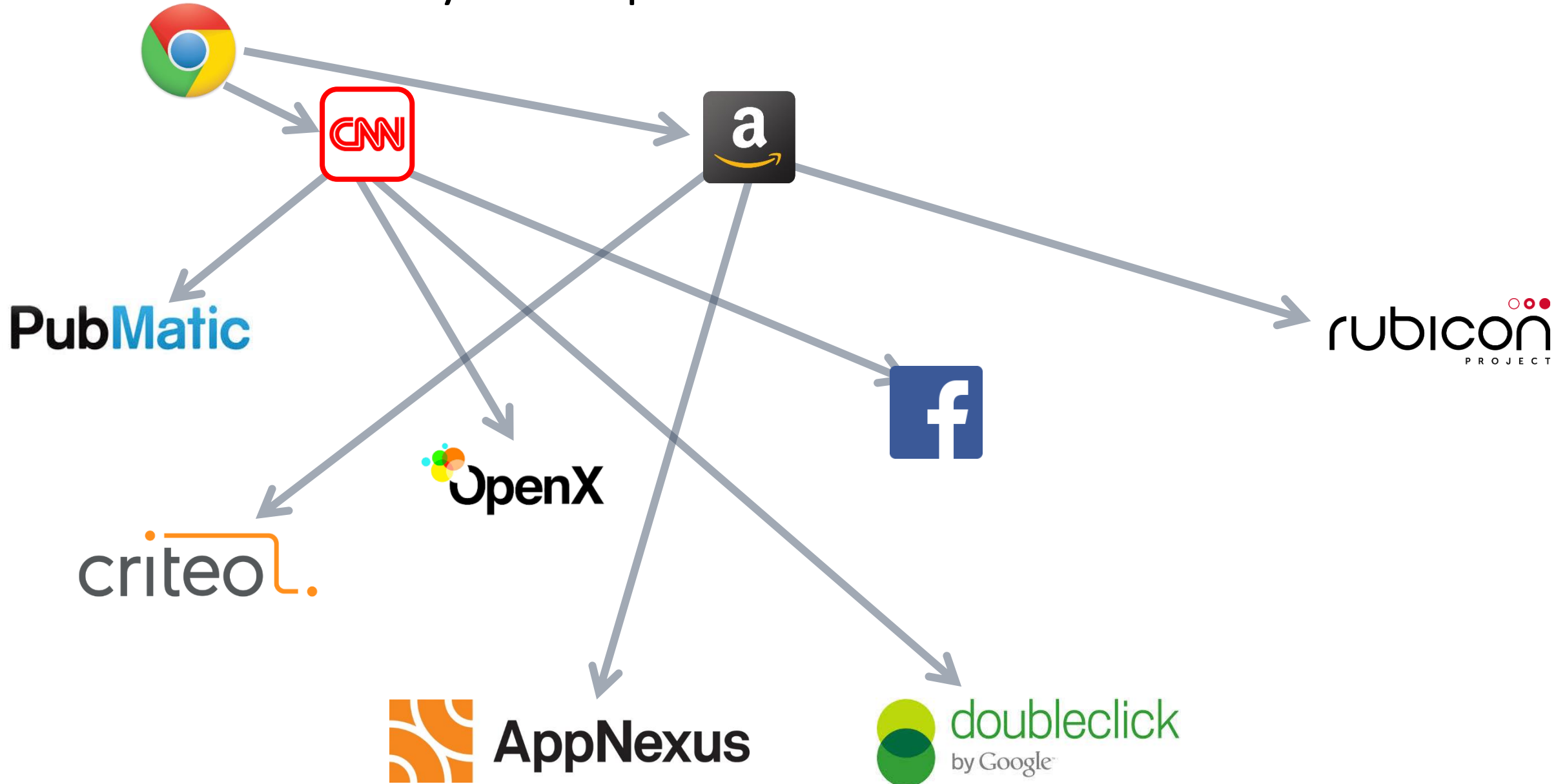
Your Privacy Footprint



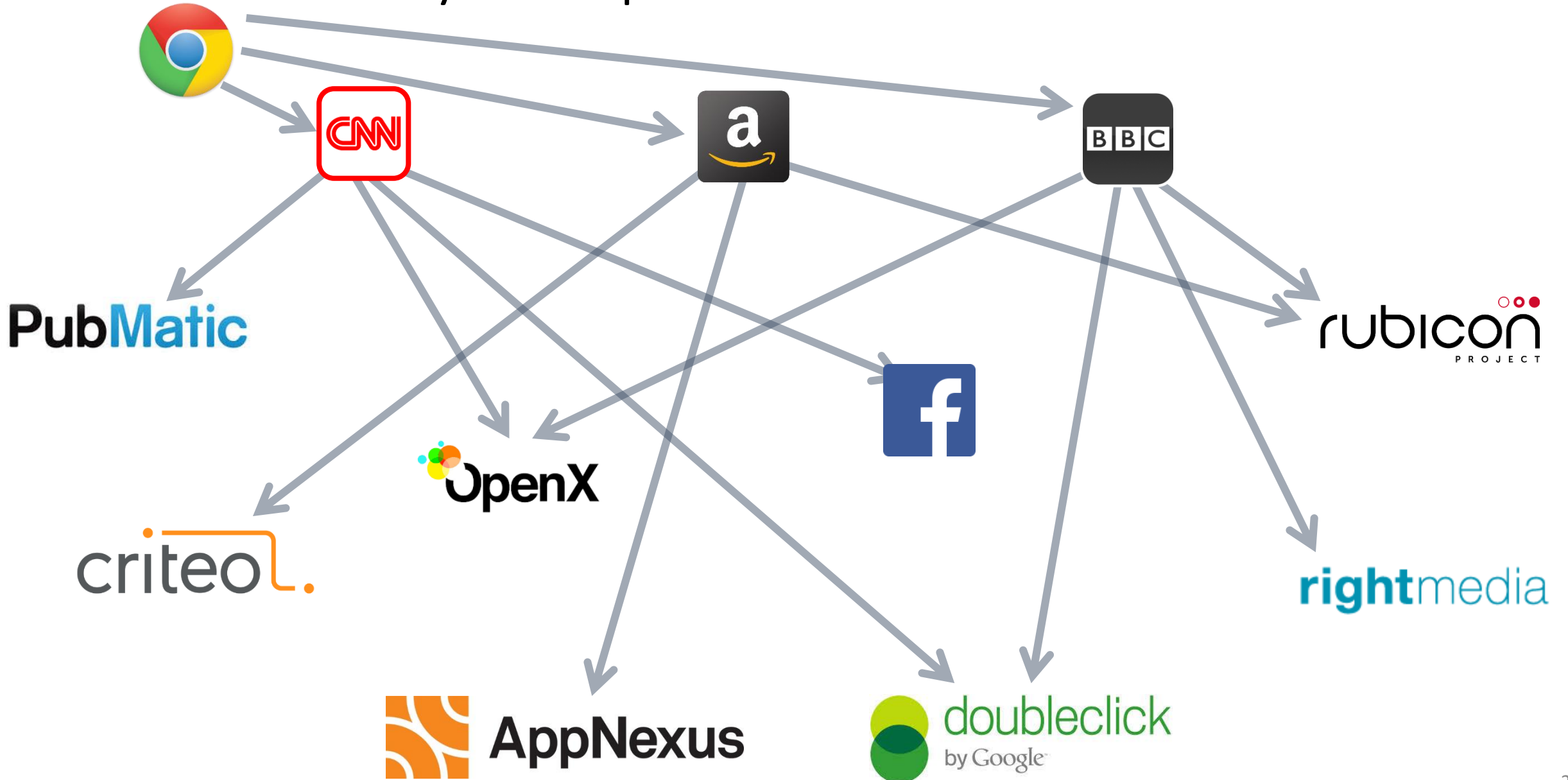
Your Privacy Footprint



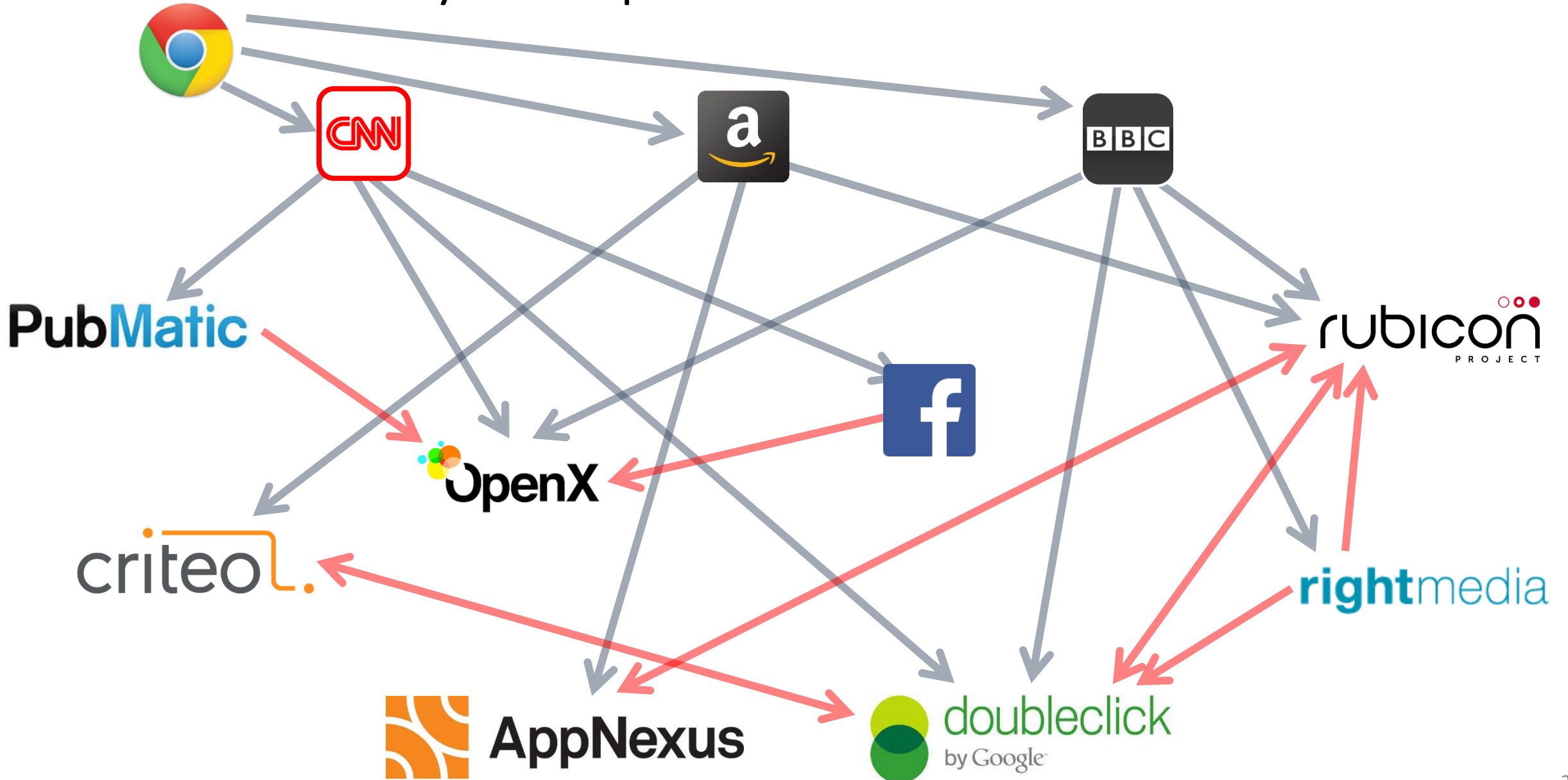
Your Privacy Footprint



Your Privacy Footprint

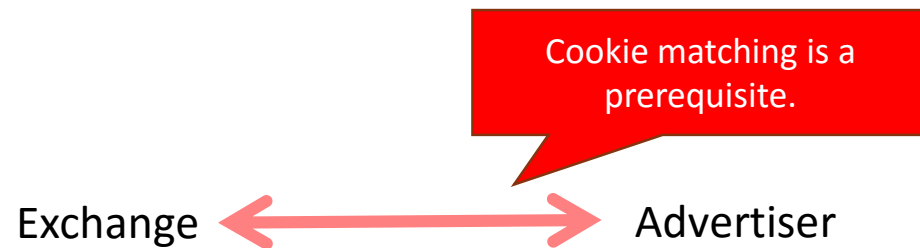


Your Privacy Footprint



Real Time Bidding

- RTB brings more flexibility in the ad ecosystem.
 - Ad request managed by an Ad Exchange which holds an auction.
 - Advertisers bid on each ad impression.



- RTB spending to cross \$20B by 2017^[1].
 - 49% annual growth.
 - Will account for 80% of US Display Ad spending by 2022.

[1] <http://www.prnewswire.com/news-releases/new-idc-study-shows-real-time-bidding-rtb-display-ad-spend-to-grow-worldwide-to-208-billion-by-2017-228061051.html>

User



Publisher



GET, CNN's Cookie



GET, DoubleClick's Cookie



Real Time Bidding (RTB)

Ad Exchange



Advertisers



Solicit bids, DoubleClick's Cookie



Bid



User



Publisher



GET, CNN's Cookie



GET, DoubleClick's Cookie



GET, RightMedia's Cookie



Advertisement



Real Time Bidding (RTB)

Ad Exchange



Advertisers

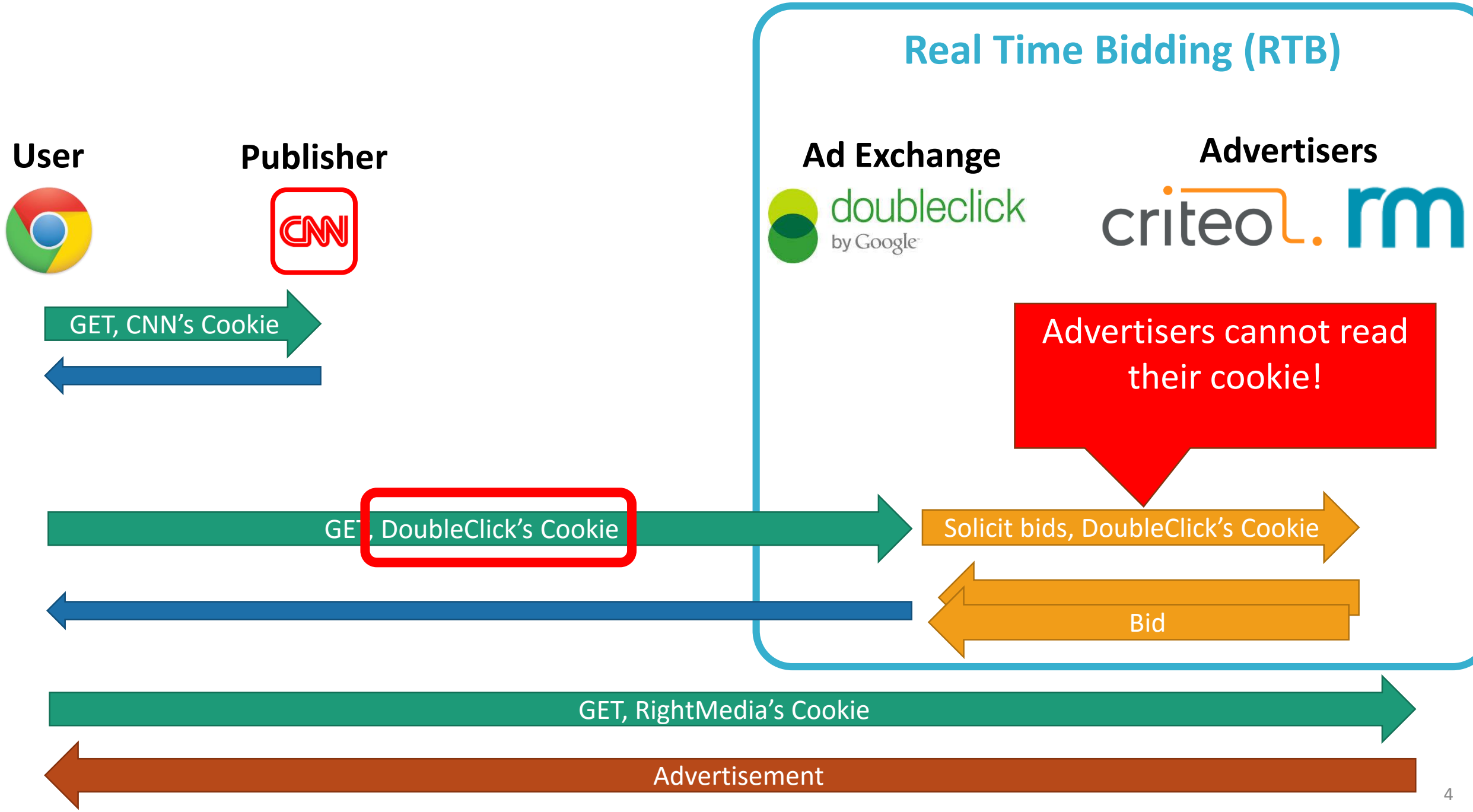


Solicit bids, DoubleClick's Cookie



Bid





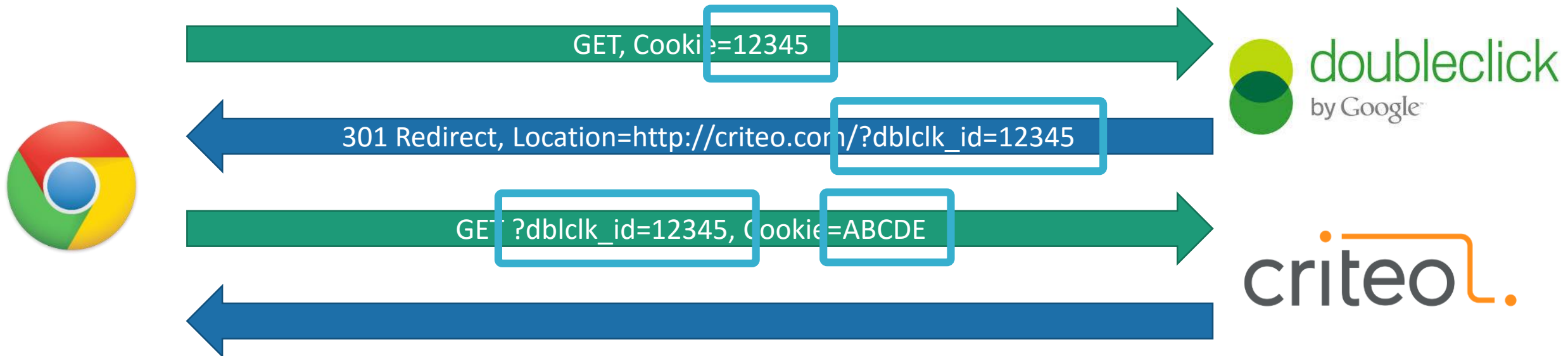
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: **cookie matching**

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



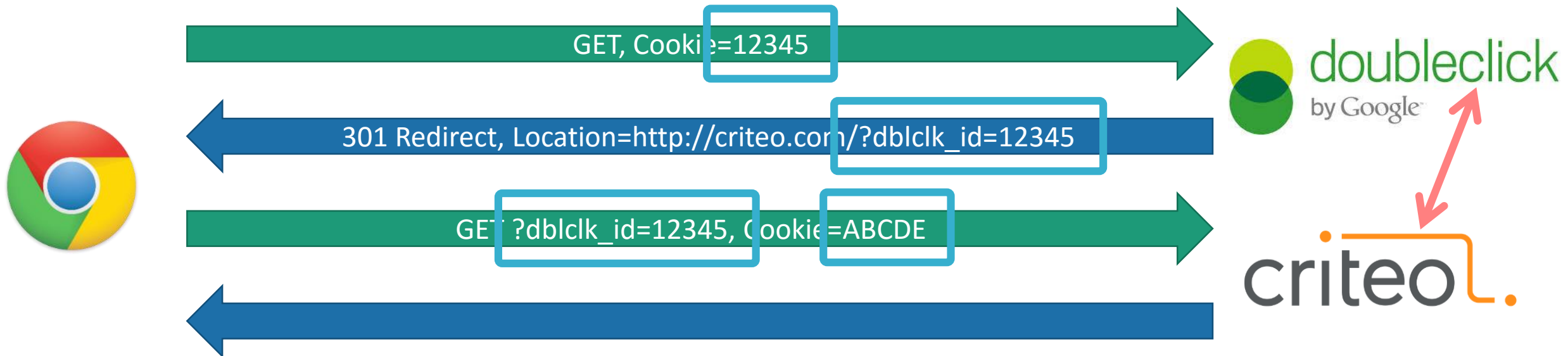
Cookie Matching

Key problem: Advertisers cannot read their cookies in the RTB auction

- How can they submit reasonable bids if they cannot identify the user?

Solution: **cookie matching**

- Also known as cookie synching
- Process of linking the identifiers used by two ad exchanges



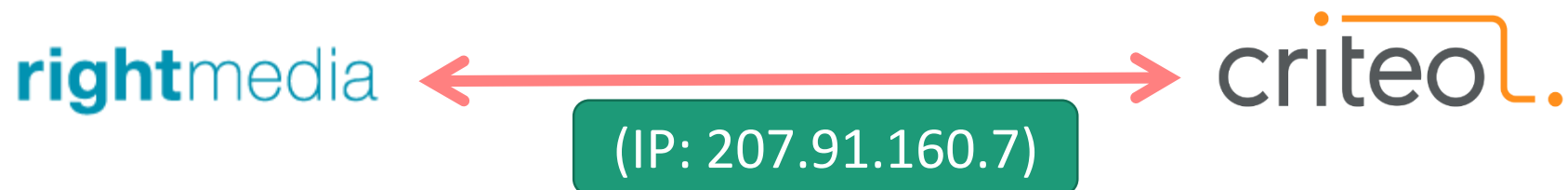
Prior Work

- Several studies have examined cookie matching
 - Acar *et al.* found hundreds of domains passing identifiers to each other
 - Olejnik *et al.* found 125 exchanges matching cookies
 - Falahrastegar *et al.* analyzed clusters of exchanges that share the exact same cookies
- These studies rely on studying HTTP requests/responses.

Challenge 1: Server Side Matching



Behind the scene, RightMedia and Criteo sync up.



Challenge 2: Obfuscation



Challenge 2: Obfuscation



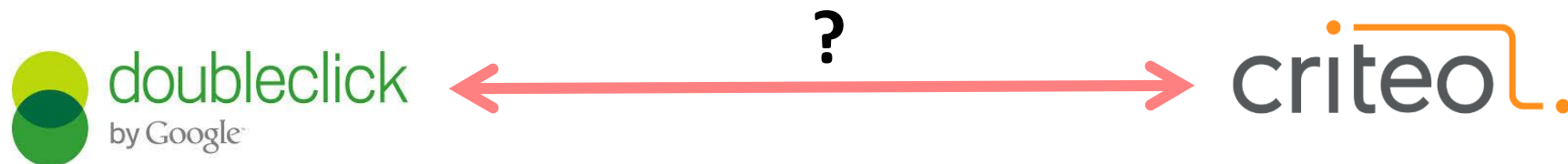
Challenge 2: Obfuscation



Goal

Develop a method to identify information flows (cookie matching) between ad exchanges

- Mechanism agnostic: resilient to obfuscation
- Platform agnostic: detect sharing on the client- and server-side



Key Insight: Use Retargeted Ads

Retargeted ads are the most highly targeted form of online ads



Cisco-Linksys AE1000 High-Performance Wireless-N Adapter

by Linksys

★★★★☆ 207 customer reviews | 10 answered questions

Price: \$15.99 

Only 1 left in stock.

Want it Tuesday, June 14? Order within 33 hrs 50 mins and choose One-Day Shipping at checkout. [Details](#)

Sold by Home Sweet Home Direct and Fulfilled by Amazon.

Eligible for  donation.



Want to hire a computer technician?

Buy professional computer technician services directly on Amazon. Backed by our Happiness Guarantee.

[Learn more](#)

• Networking Equipment Features: WEP Security, WPA Security, Easy Setup, WPA2



amazon.com [Shop now](#)

Linksys/Cisco AE1000
300Mbps 802.11n
Dual-Band Wireless...
~~\$39.99~~ \$15.99

Privacy

Key insight: because retargets are so specific, they can be used to conduct controlled experiments

- Information **must be** shared between ad exchanges to serve retargeted ads

Contributions

1. Novel methodology for identifying information flows between ad exchanges
2. Demonstrate the impact of ad network obfuscation in practice
 - 31% of cookie matching partners cannot be identified using heuristics
3. Develop a method to categorize information sharing relationships
4. Use graph analysis to infer the roles of actors in the ad ecosystem

Contributions

1. Novel methodology for identifying information flows between ad exchanges
2. Demonstrate the impact of ad network obfuscation in practice
 - 31% of cookie matching partners cannot be identified using heuristics
3. Develop a method to categorize information sharing relationships
4. ~~Use graph analysis to infer the roles of actors in the ad ecosystem~~

Data Collection

Classifying Ad Network Flows

Results

Using Retargets as an Experimental Tool

Key observation: retargets are only served under very specific circumstances

1)



Advertiser observes the user at a shop

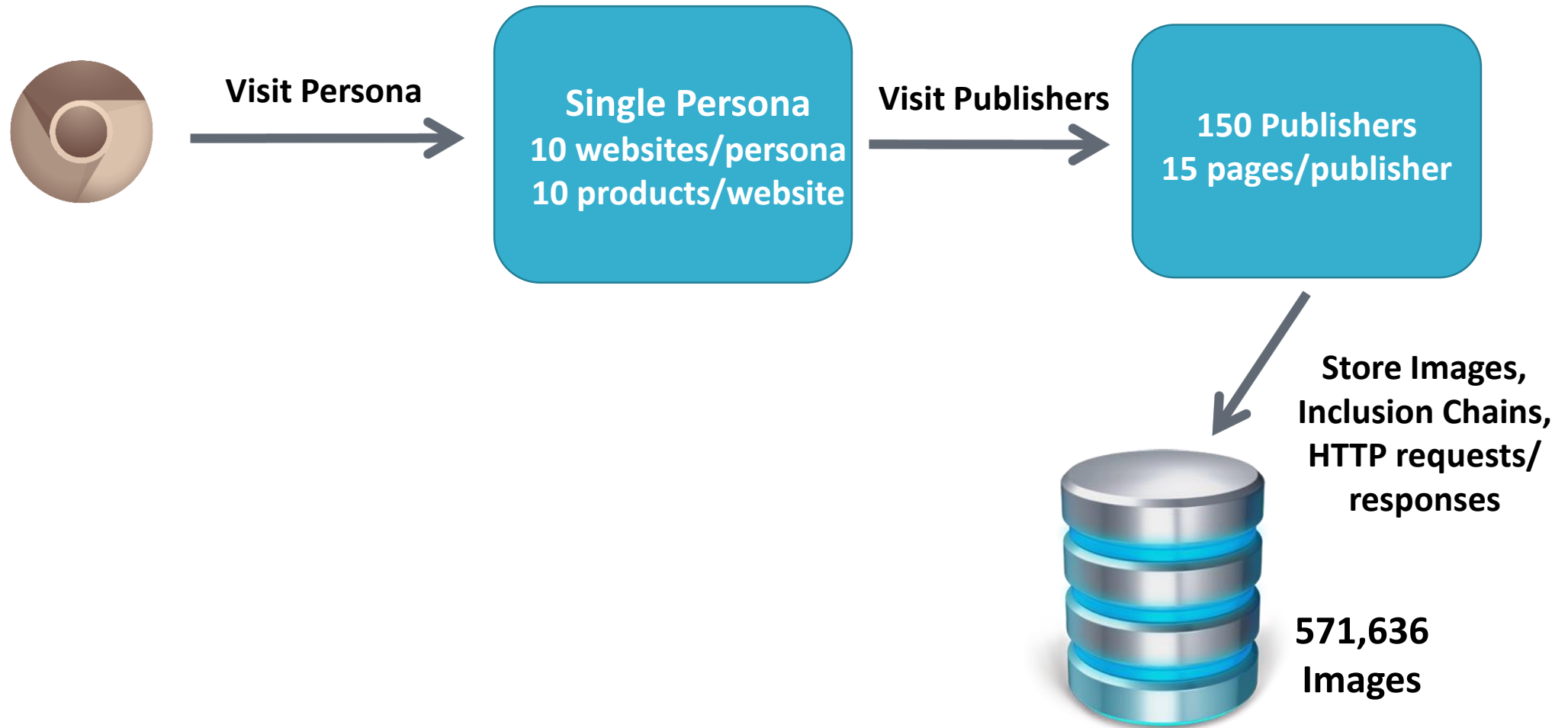
Advertiser and the exchange must have matched cookies

2)



This implies a causal flow of information from Exchange → Advertiser

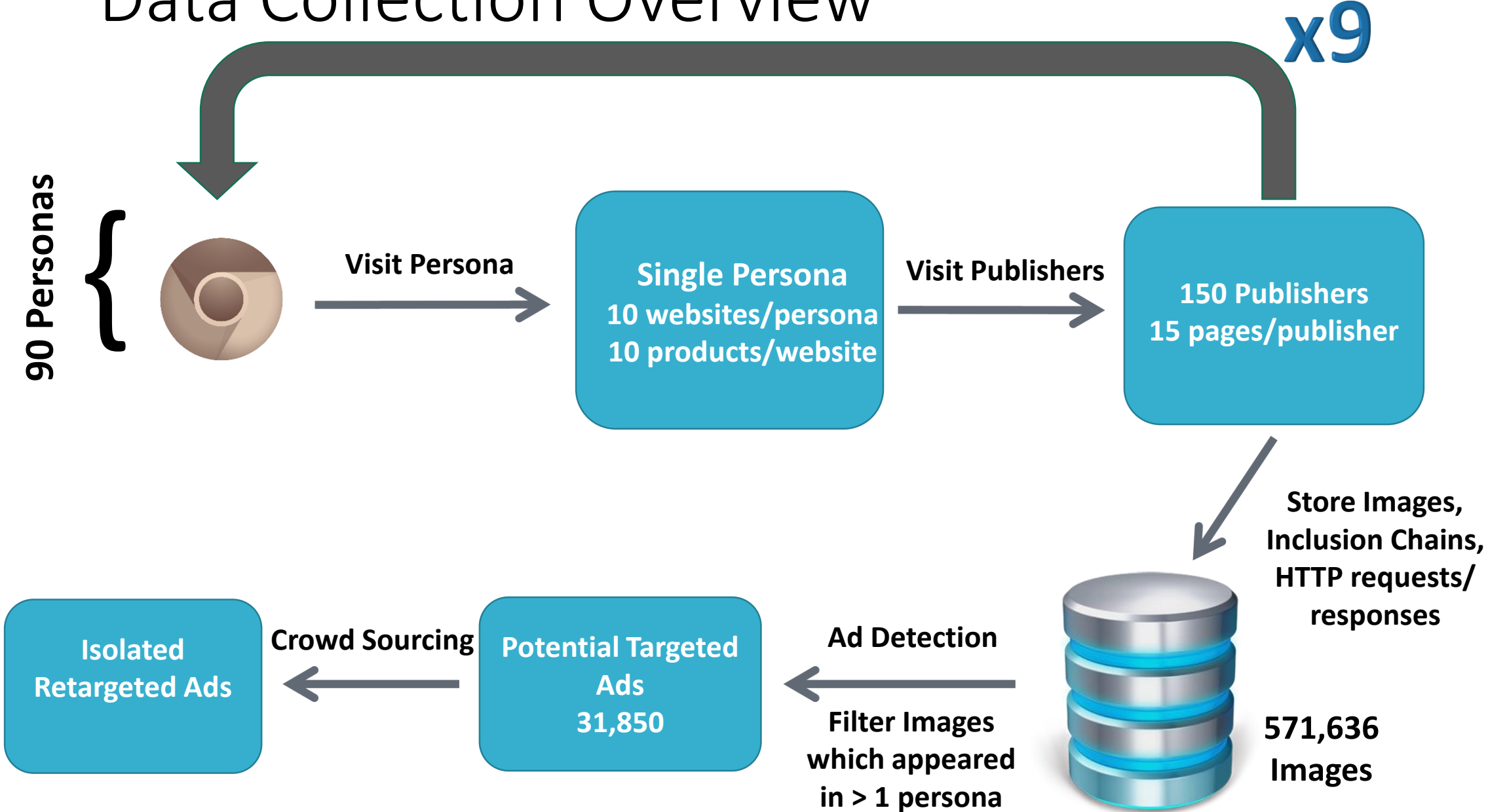
Data Collection Overview



Data Collection Overview



Data Collection Overview



Crowd Sourcing


We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- \$415 spent.

Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.


- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- \$415 spent.

Your response	Image
<p>Select the appropriate category for the image.</p> <p><input checked="" type="radio"/> shopping__jewelry__diamonds Rings, necklace, etc.</p> <p><input type="radio"/> None of the above</p> <p>Save and Continue</p>	

Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.


- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- \$415 spent.

Your response	Image
<p>Select the appropriate category for the image.</p> <p><input checked="" type="radio"/> shopping__jewelry__diamonds Rings, necklace, etc.</p> <p><input type="radio"/> None of the above</p> <p>Does this image say it came from one the following websites?</p> <p><input type="radio"/> adiamor.com</p> <p><input type="radio"/> bluenile.com</p> <p><input type="radio"/> gilletts.com.au</p> <p><input type="radio"/> lumeradiamonds.com</p> <p><input type="radio"/> shaneco.com</p> <p><input type="radio"/> szul.com</p> <p><input type="radio"/> totaram.com</p> <p><input type="radio"/> washingtondiamond.com</p> <p><input type="radio"/> whiteflash.com</p> <p><input type="radio"/> No</p> <p>Save and Continue</p>	

Crowd Sourcing

We used Amazon Mechanical Turk (AMT) to label 31,850 ads.

- Total 1,142 Tasks.
- 30 ads / Task.
- 27 unlabeled.
- 3 labeled by us.
- 2 workers per ad.
- \$415 spent.

Your response	Image
<p>Select the appropriate category for the image.</p> <p><input checked="" type="radio"/> shopping__jewelry__diamonds Rings, necklace, etc.</p> <p><input type="radio"/> None of the above</p> <p>Does this image say it came from one the following websites?</p> <p><input checked="" type="radio"/> adiamor.com</p> <p><input type="radio"/> bluenile.com</p> <p><input type="radio"/> gilletts.com.au</p> <p><input type="radio"/> lumeradiamonds.com</p> <p><input type="radio"/> shaneco.com</p> <p><input type="radio"/> szul.com</p> <p><input type="radio"/> totaram.com</p> <p><input type="radio"/> washingtondiamond.com</p> <p><input type="radio"/> whiteflash.com</p> <p><input type="radio"/> No</p> <p>Save and Continue</p>	

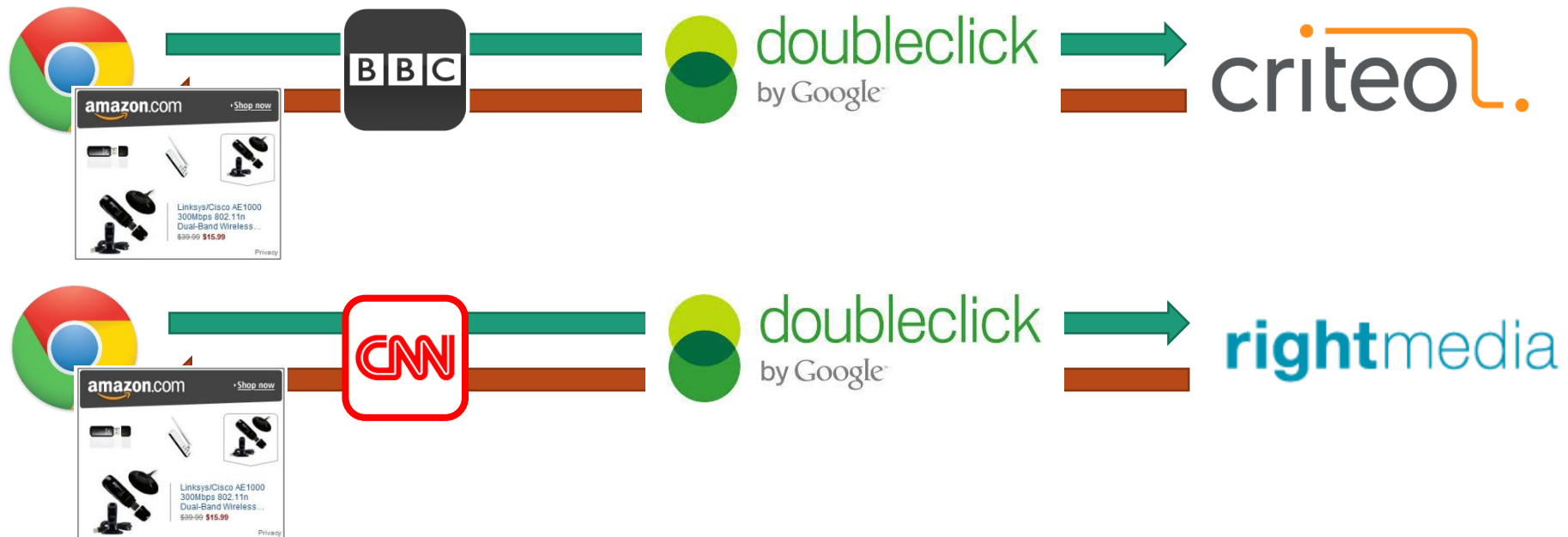
Final Dataset

5,102 unique retargeted ads

- From 281 distinct online retailers

35,448 **publisher-side** chains that served the retargets

- We observed some retargets multiple times



Data Collection

Classifying Ad Network Flows

Results

A look at Publisher Chains

Shopper-side chain



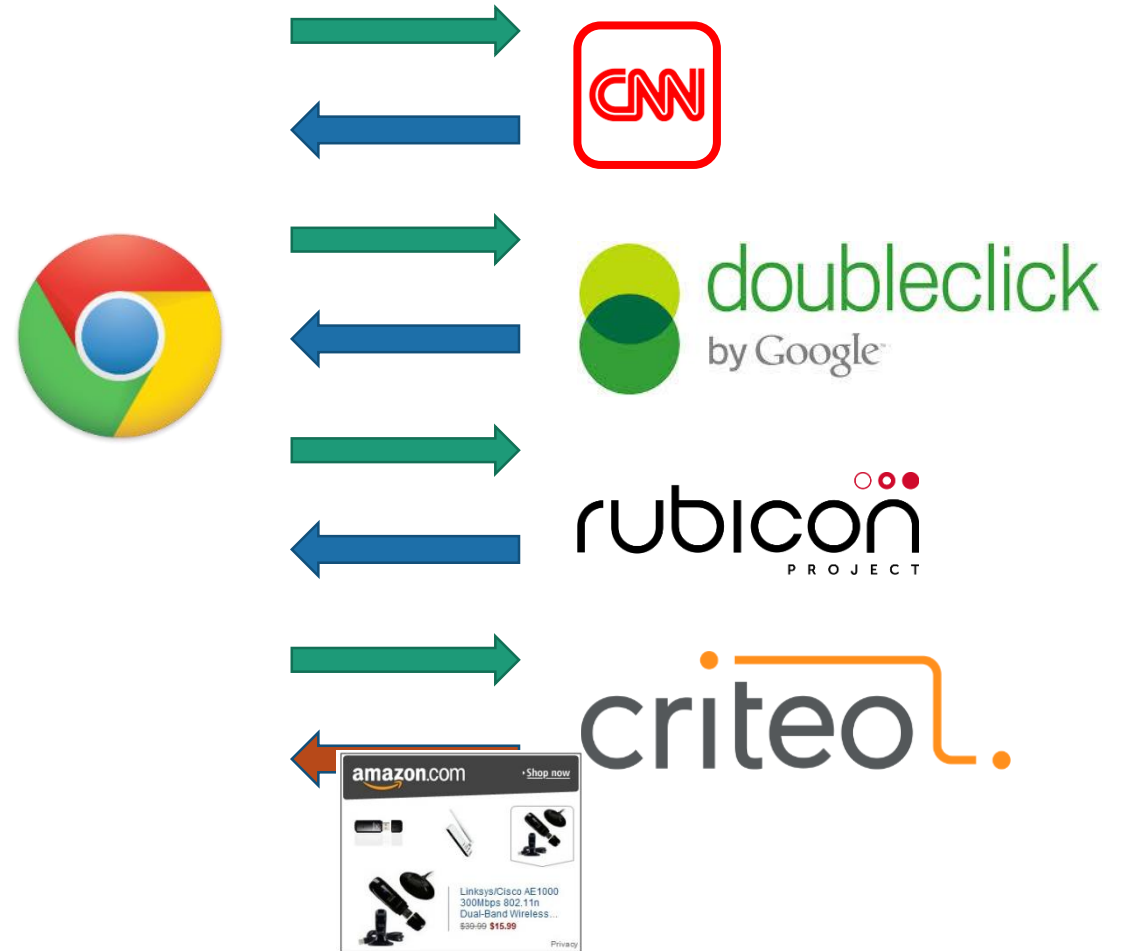
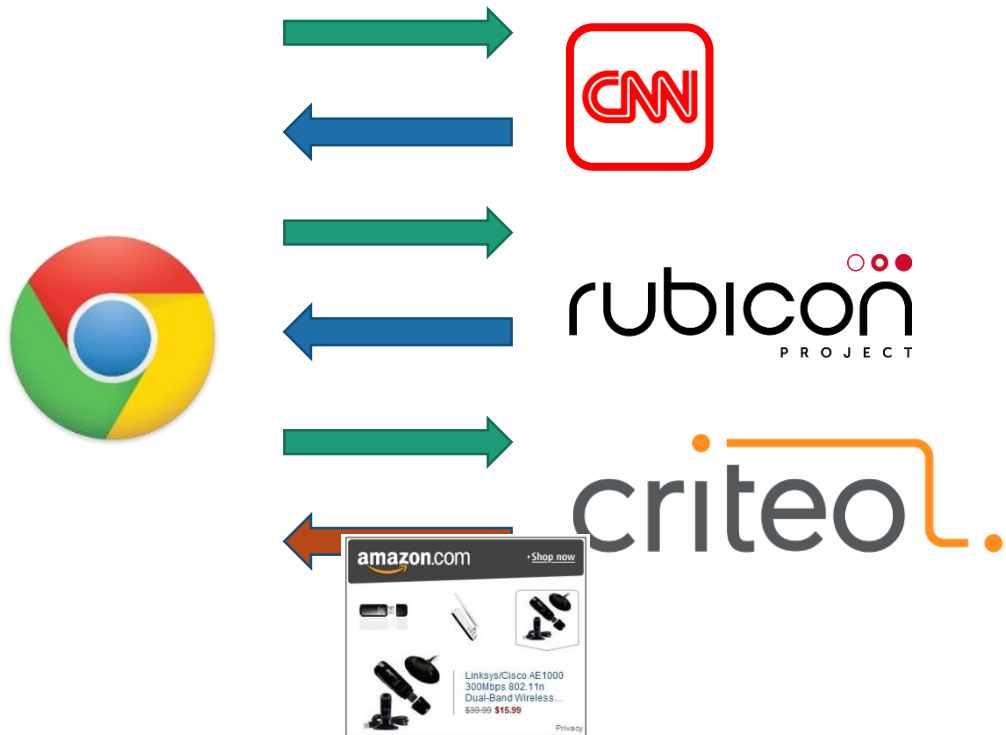
Publisher-side chain



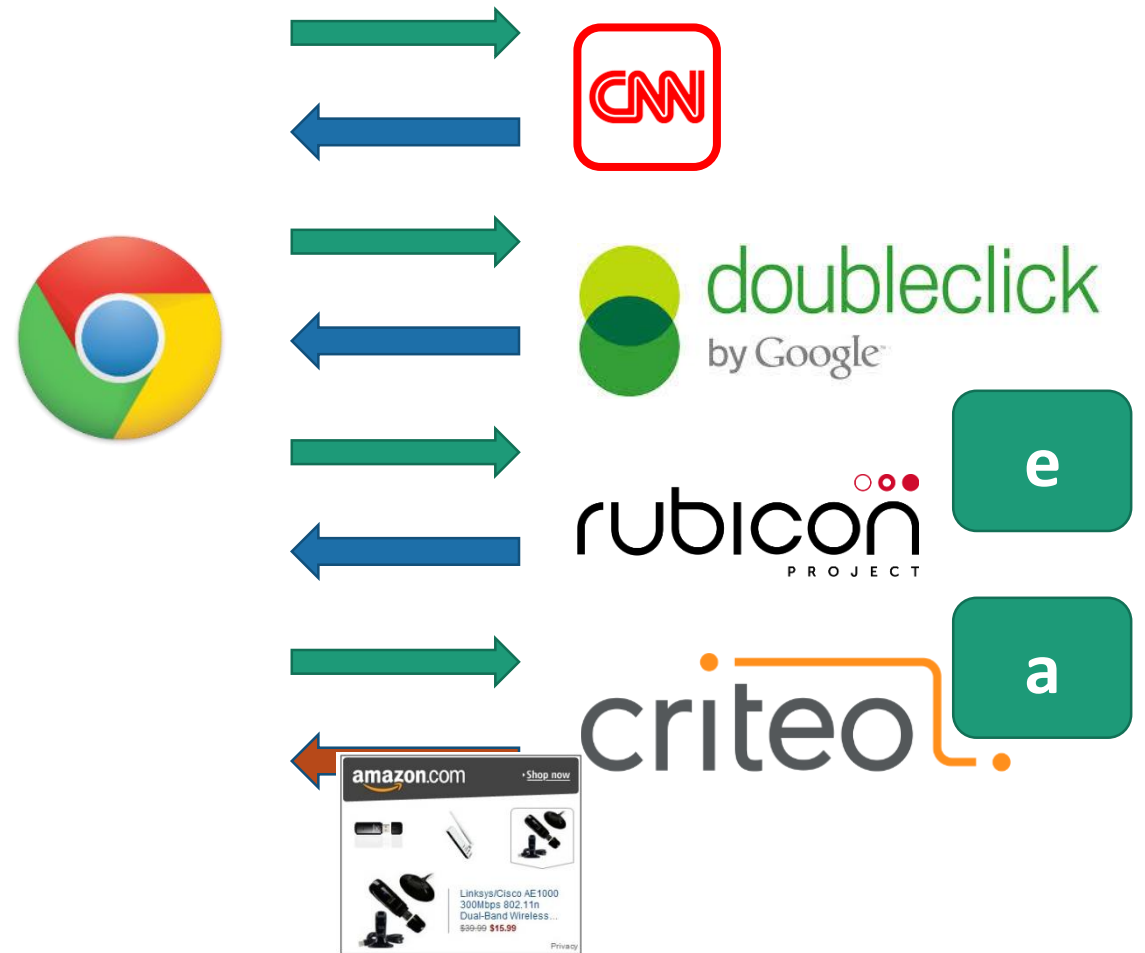
Example

- How does Criteo know to serve ad on BBC?
 - In this case it is pretty trivial.
 - Criteo observed us on the shopper.
- Can we classify all such publisher-side chains?

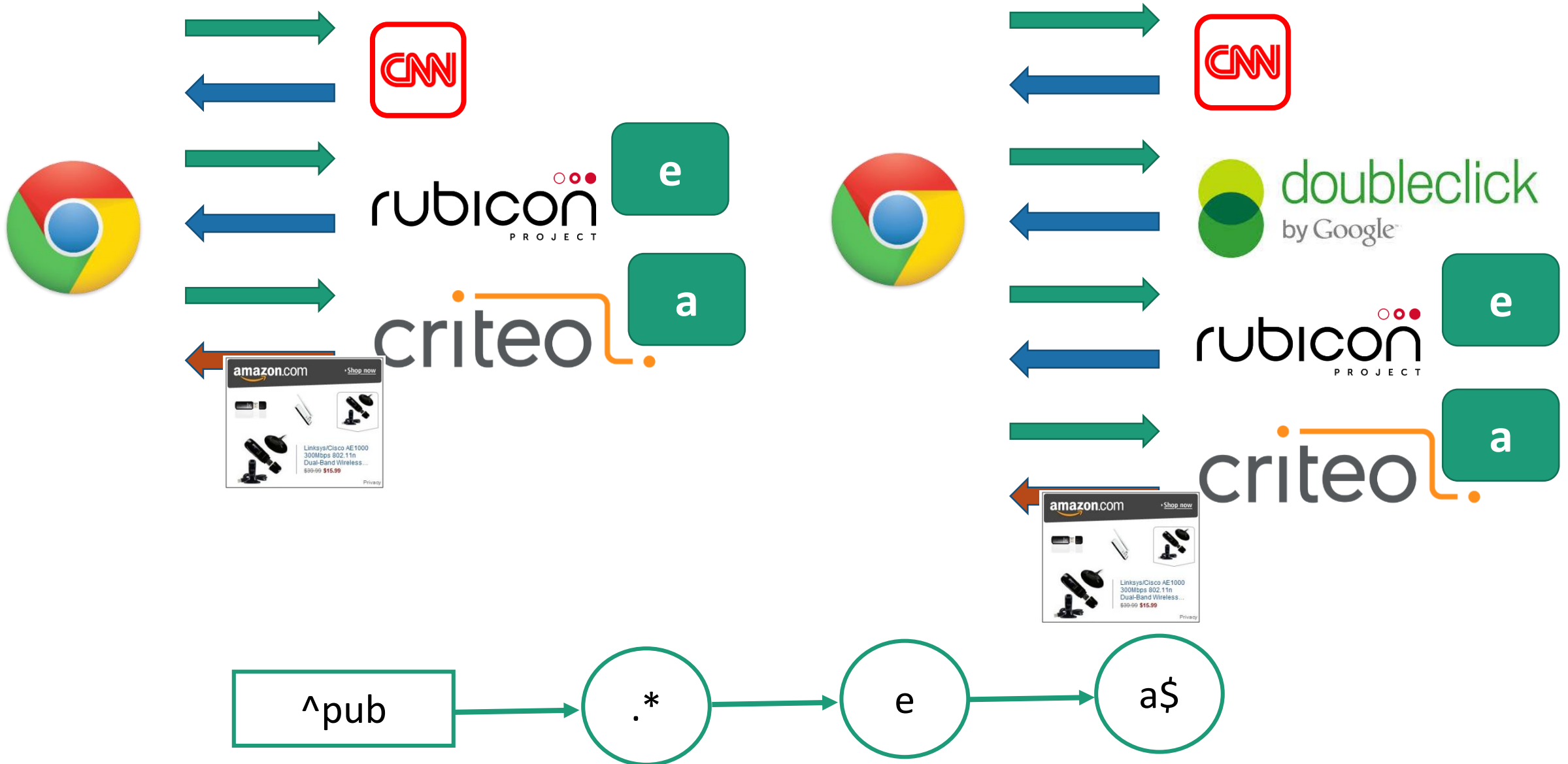
What is a Chain?



What is a Chain?



What is a Chain?



Four Classifications

Four possible ways for a retargeted ad to be served

1. Direct (Trivial) Matching
2. Cookie Matching
3. Indirect Matching
4. Latent (Server-side) Matching

Four Classifications

Four possible ways for a retargeted ad to be served

1. Direct (Trivial) Matching
2. Cookie Matching
- ~~3. Indirect Matching~~
4. Latent (Server-side) Matching

1) Direct (Trivial) Matching

Shopper-side

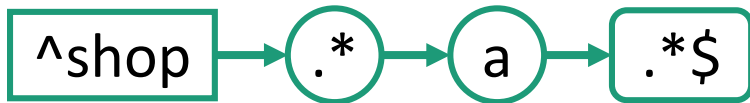


Publisher-side



Example

Rule



a is the advertiser that serves the retarget

1) Direct (Trivial) Matching

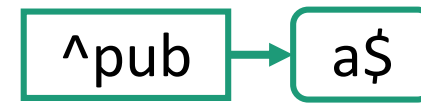
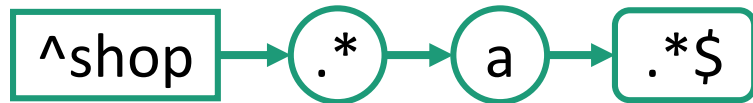
Shopper-side

Publisher-side

Example



Rule



a must appear on the shopper-side...

... but other trackers may also appear

a is the advertiser that serves the retarget

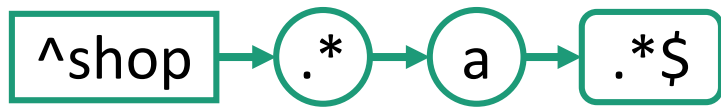
2) Cookie Matching

Shopper-side

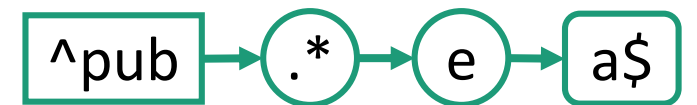
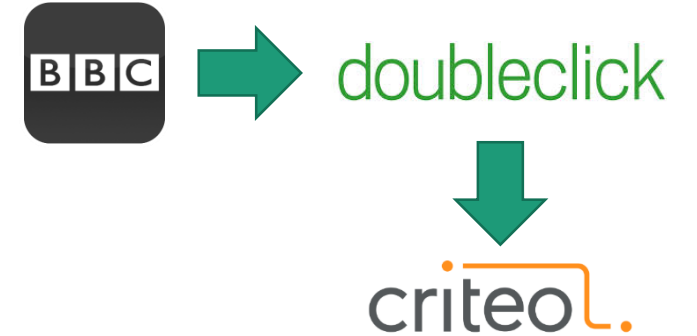
Example



Rule



Publisher-side



e precedes *a*,
which implies an
RTB auction

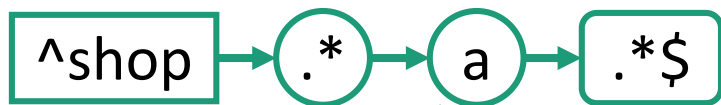
2) Cookie Matching

Shopper-side

Example

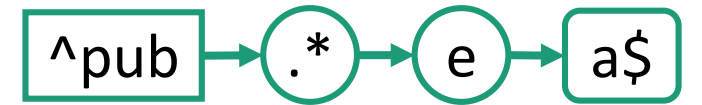
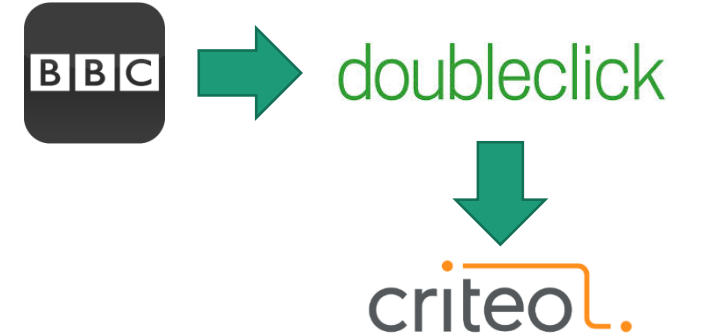


Rule



a must appear on the shopper-side

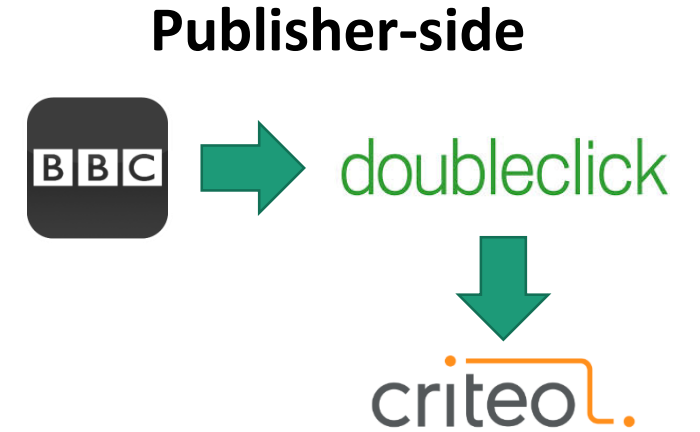
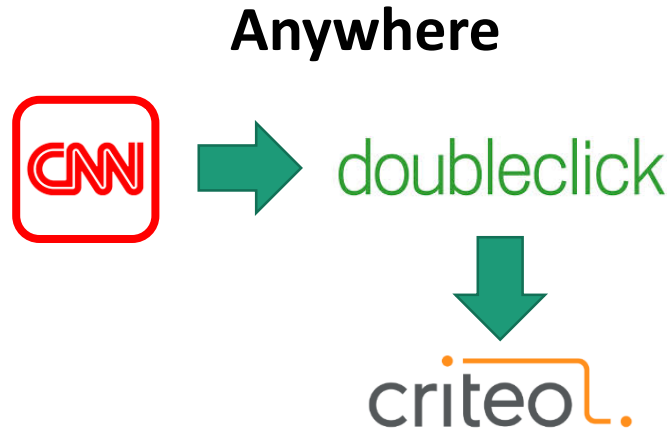
Publisher-side



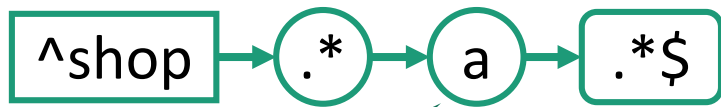
e precedes *a*, which implies an RTB auction

2) Cookie Matching

Example



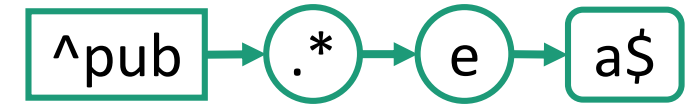
Rule



a must appear on the shopper-side



Transition $e \rightarrow a$ is where cookie match occurs

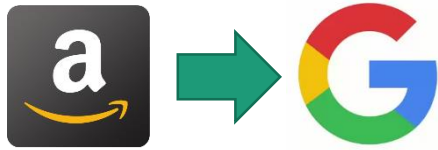


e precedes *a*, which implies an RTB auction

3) Latent (Server-side) Matching

Example

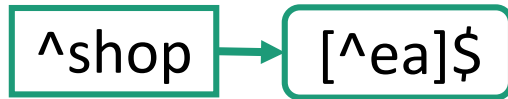
Shopper-side



Publisher-side



Rule



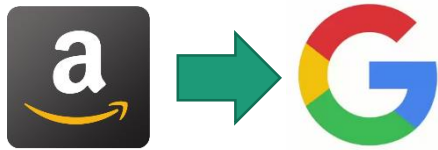
Neither *e* nor *a* appears on the shopper-side



3) Latent (Server-side) Matching

Example

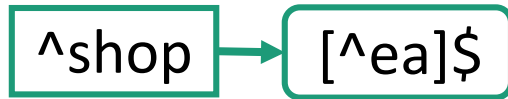
Shopper-side



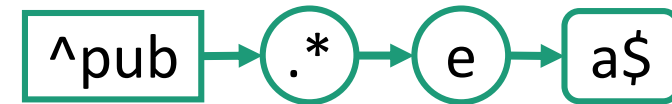
Publisher-side



Rule



Neither *e* nor *a* appears on the shopper-side

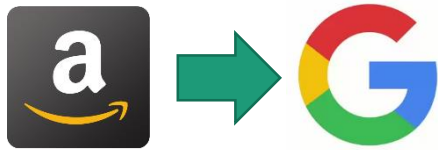


a must receive information from some shopper-side tracker

3) Latent (Server-side) Matching

Example

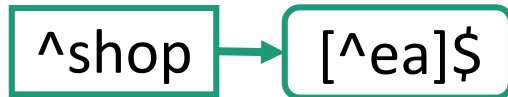
Shopper-side



Publisher-side



Rule



Neither *e* nor *a* appears on the shopper-side



a must receive information from some shopper-side tracker

We find latent matches in practice!

Data Collection
Classifying Ad Network Flows
Results

Categorizing Chains

Raw Chains

Type	Chains	%
Direct (Trivial) Match	1770	5
Cookie Match	25049	71
Latent (Server-side) Match	5362	15
<i>No Match</i>	775	2

Take away:

- 1- As expected, most retargets are due to cookie matching
- 2- Very small number of chains that cannot be categorized
 - Suggests low false positive rate of AMT image labeling task
- 3- Surprisingly large amount latent matches...

Categorizing Chains

Type	Raw Chains		Clustered Chains	
	Chains	%	Chains	%
Direct (Trivial) Match	1770	5	8449	24
Cookie Match	25049	71	25873	73
Latent (Server-side) Match	5362	15	343	1
<i>No Match</i>	775	2	183	1

Cluster together domains by “owner”

- E.g. google.com, doubleclick.com, googlesyndication.com

Categorizing Chains

Type	Raw Chains		Clustered Chains	
	Chains	%	Chains	%
Direct (Trivial) Match	1770	5	8449	24
Cookie Match	25049	71	25873	73
Latent (Server-side) Match	5362	15	343	1
<i>No Match</i>	775	2	183	1

Cluster together domains by “owner”

- E.g. google.com, doubleclick.com, googlesyndication.com

Latent matches essentially disappear

- The vast majority of these chains involve Google
- Suggests that Google shares tracking data across their services

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsrvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsrvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsrvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

Who is Cookie Matching?

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔ P
criteo	↔	doubleclick	3610	1144	→ E, P ← DC, P
criteo	↔	adnxs	3263	1066	↔ E, P
criteo	↔	rubiconproject	1586	749	↔ E, P
criteo	↔	servedbyopenx	707	460	↔ P
doubleclick	↔	steelhousemedia	362	27	→ P ← E, P
mathtag	↔	mediaforge	360	124	↔ E, P
netmng	↔	scene7	267	119	→ E ← ?
googlesyndication	↔	adsvr	107	29	↔ P
rubiconproject	↔	steelhousemedia	86	30	↔ E
googlesyndication	↔	steelhousemedia	47	22	?
adtechus	→	adacado	36	18	?
atwola	→	adacado	32	6	?
adroll	↔	adnxs	31	8	?

Heuristics Key (used by prior work)

E – share exact cookies

P – special URL parameters

DC – DoubleClick URL parameters

? – Unknown sharing method

31% of cookie matching partners would be missed.

Summary

We develop a novel methodology to detect information flows between ad exchanges

- Controlled methodology enables causal inference
- Defeats obfuscation attempts
- Detects client- and server-side flows

Dataset gives a better picture of ad ecosystem

- Reveals which ad exchanges are linking information about users
- Allows us to reason about how information is being transferred