

# Diffusion of User Tracking Data in the Online Advertising Ecosystem

Muhammad Ahmad Bashir and Christo Wilson



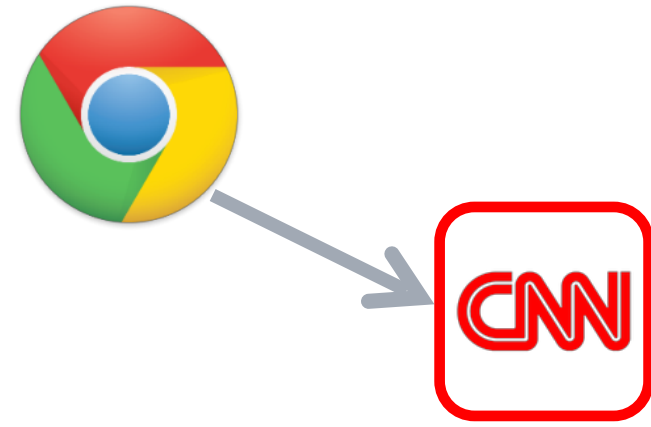
**Northeastern University**  
College of Computer and Information Science

# Your Digital Privacy Footprint

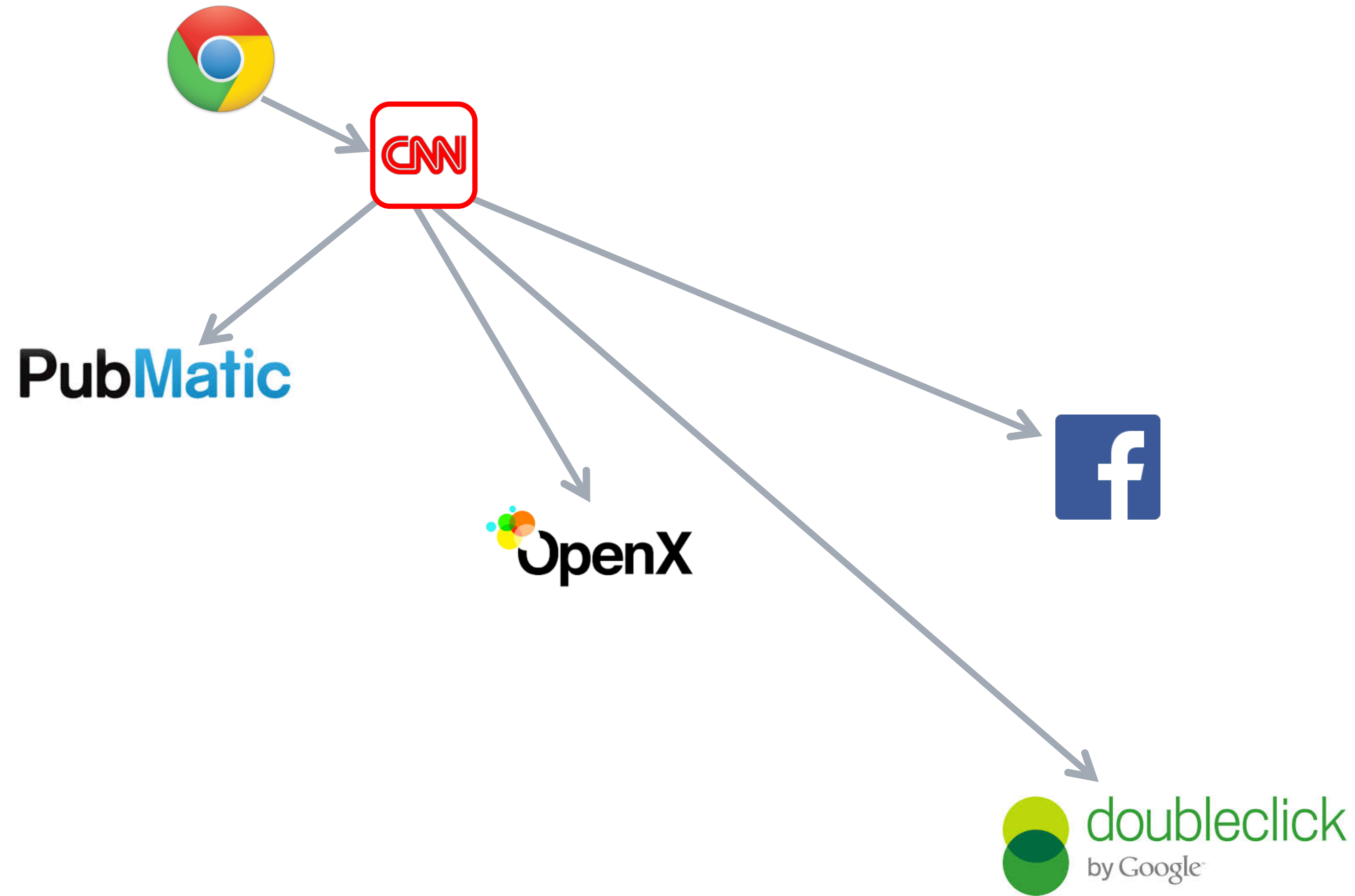
# Your Digital Privacy Footprint



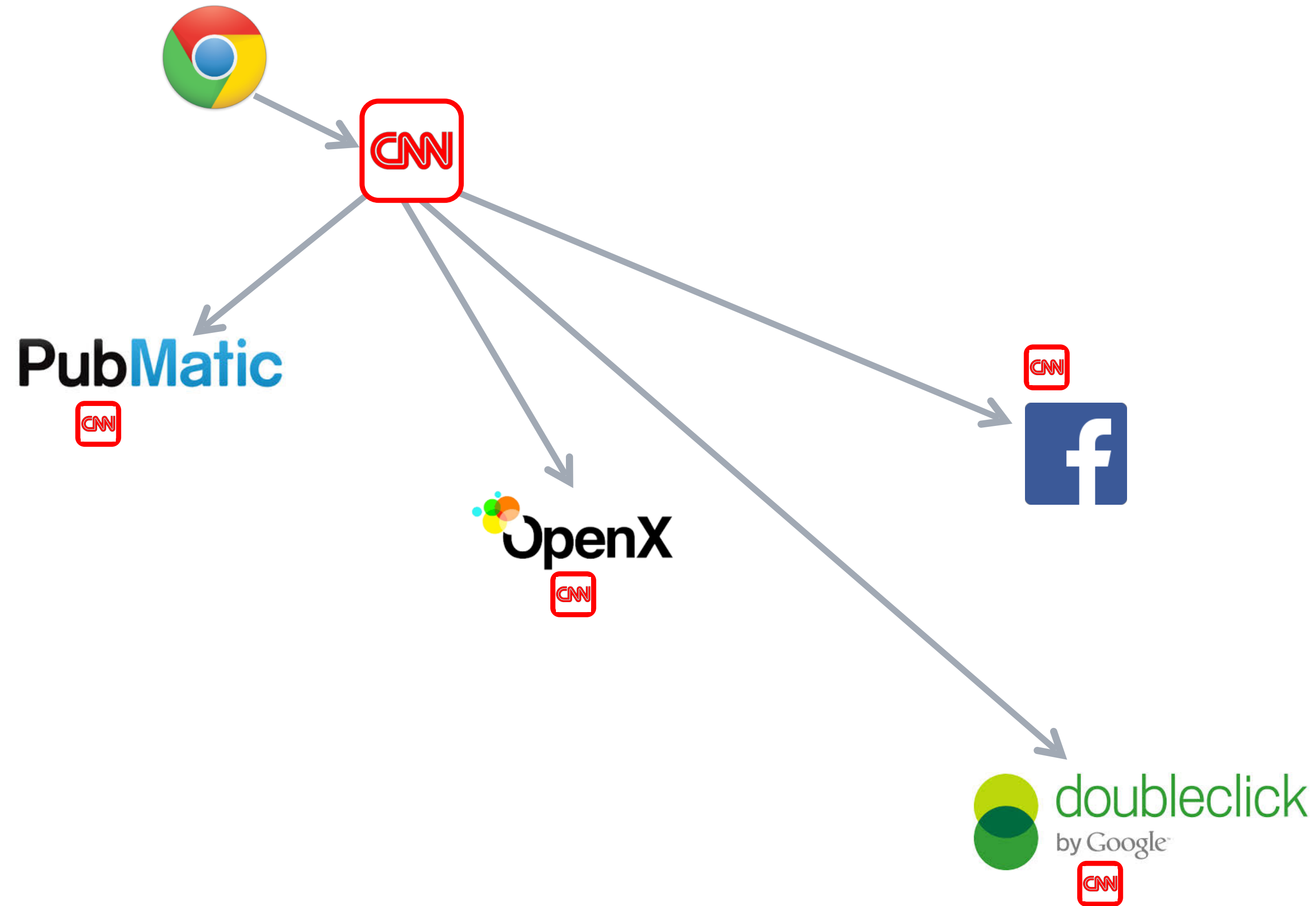
# Your Digital Privacy Footprint



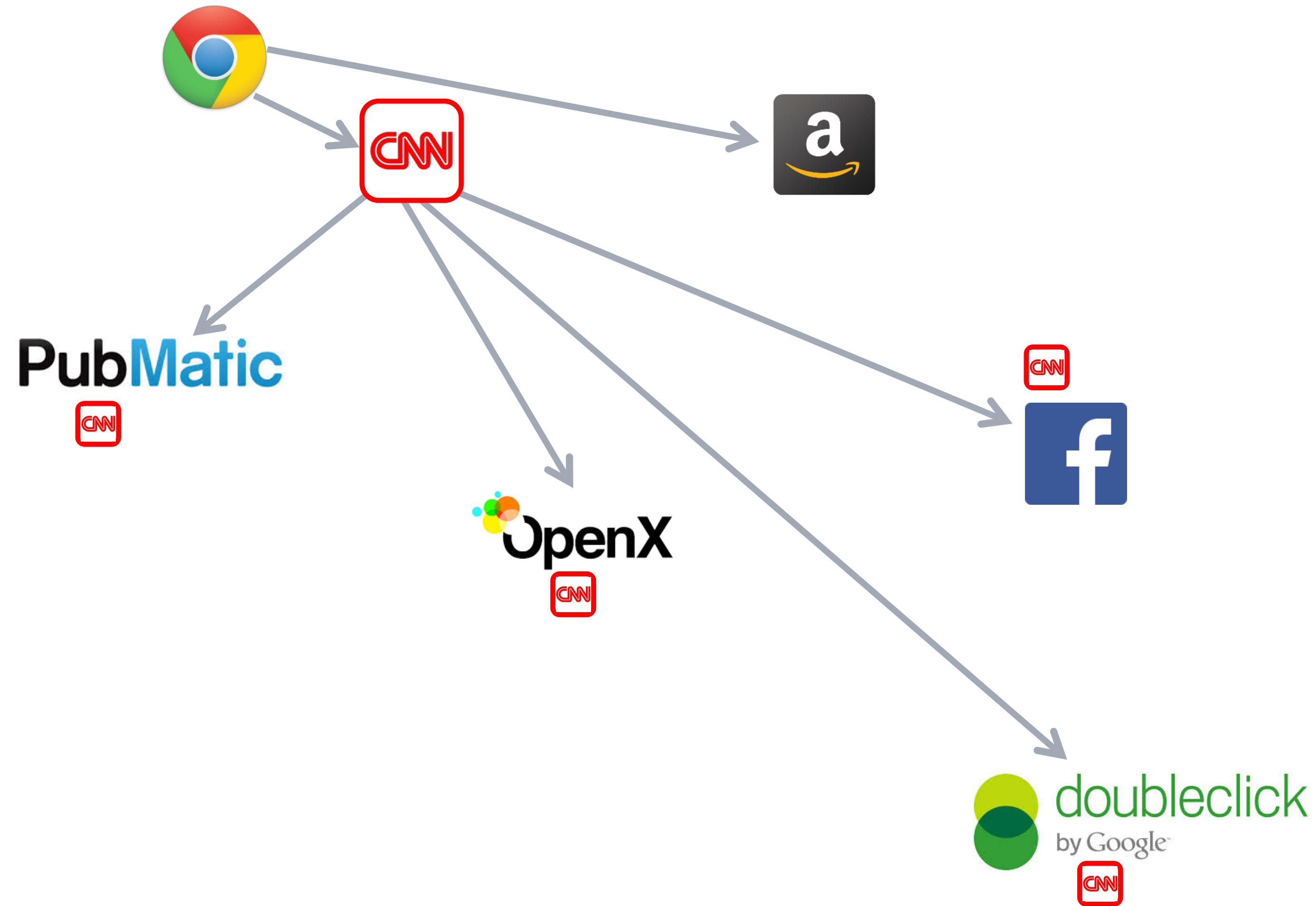
# Your Digital Privacy Footprint



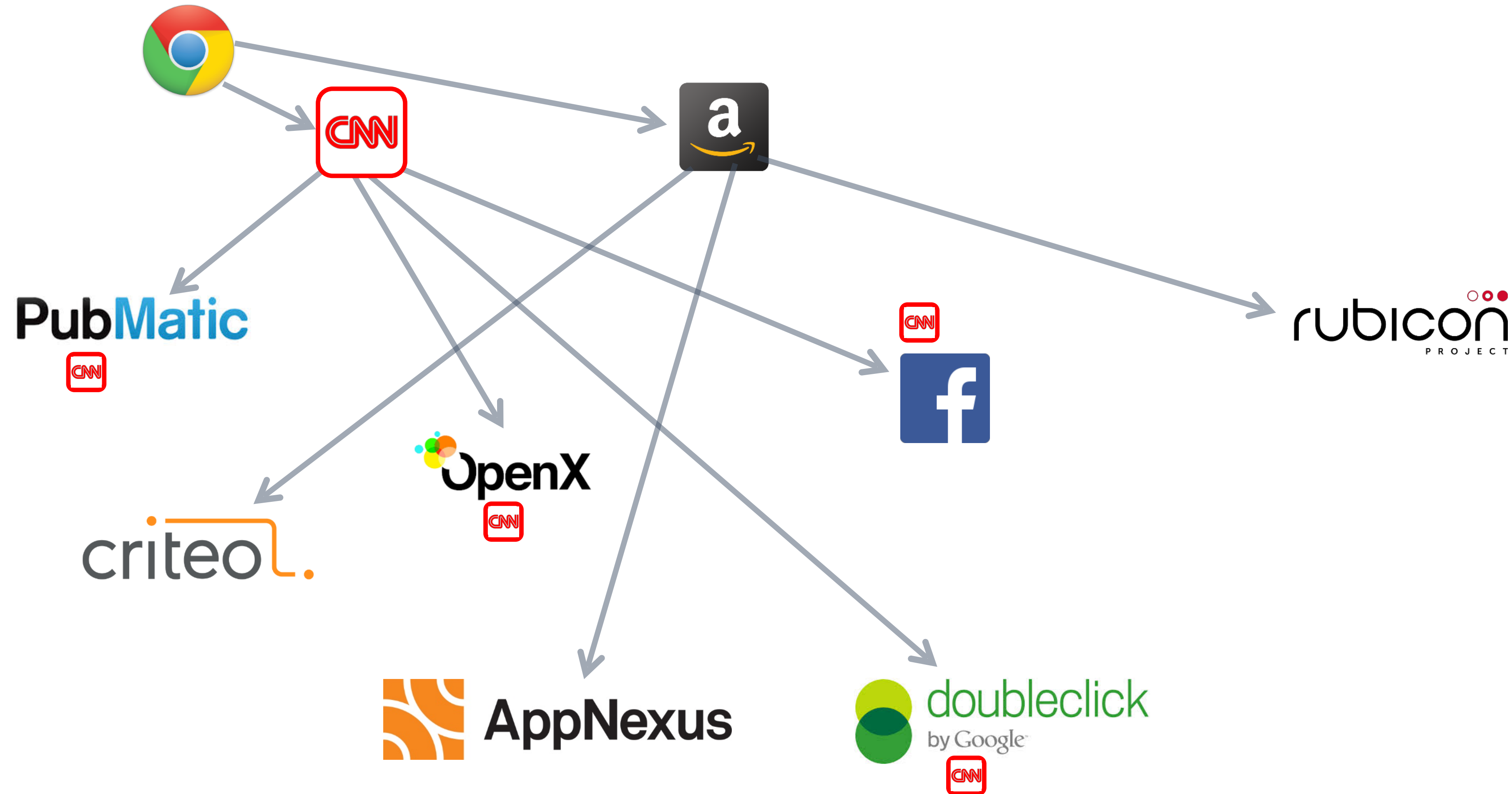
# Your Digital Privacy Footprint



# Your Digital Privacy Footprint

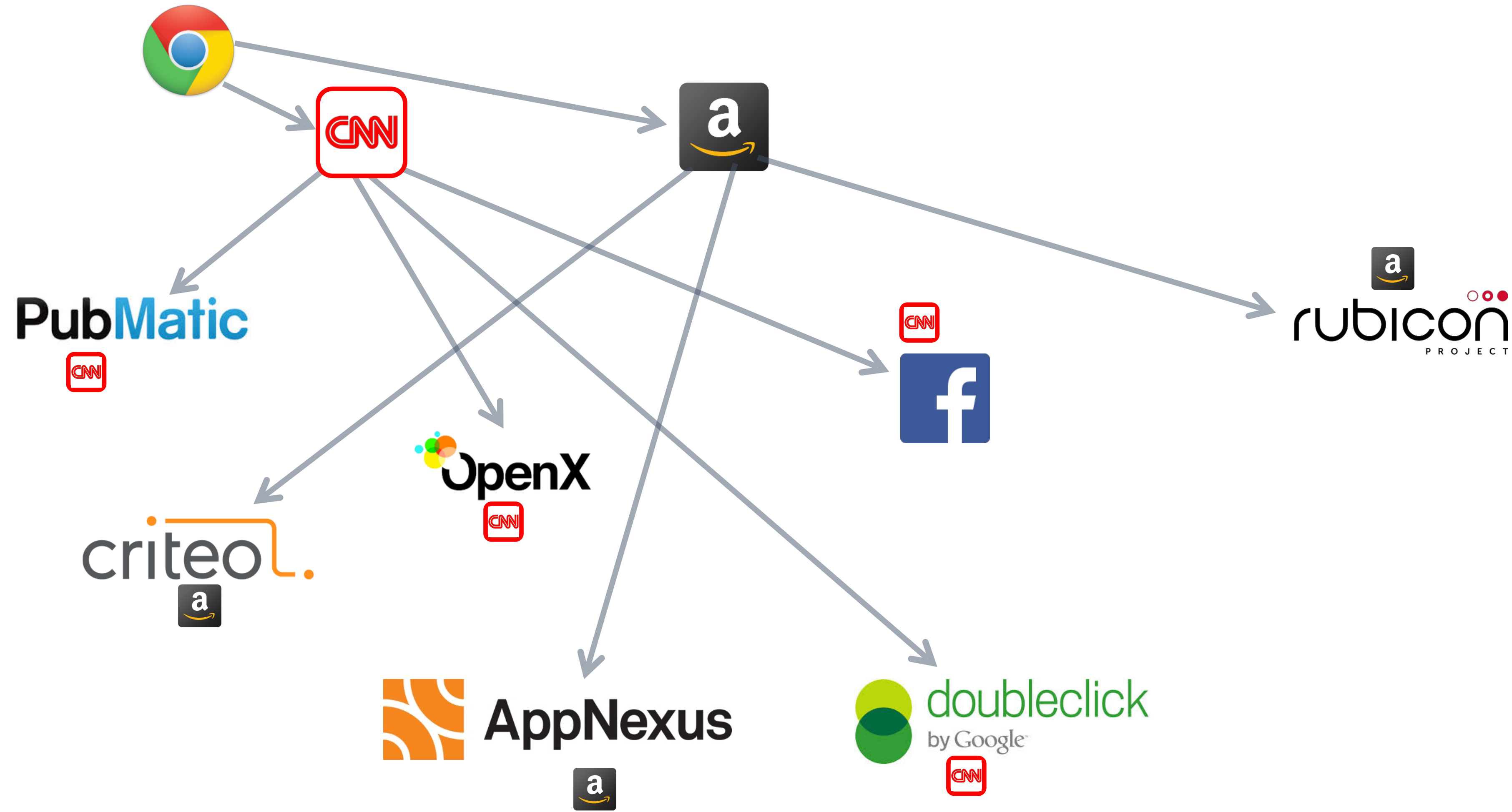


# Your Digital Privacy Footprint



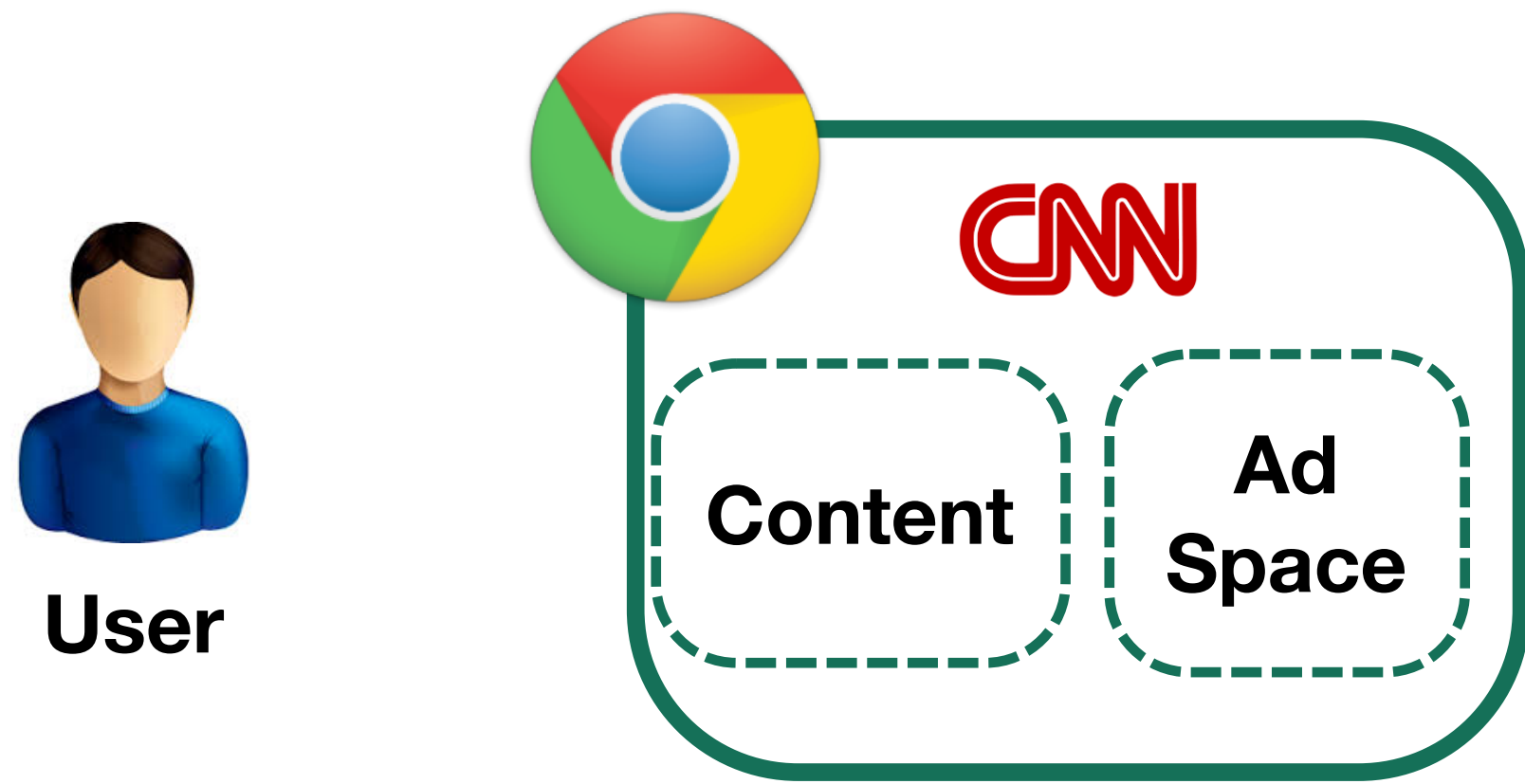


# Your Digital Privacy Footprint

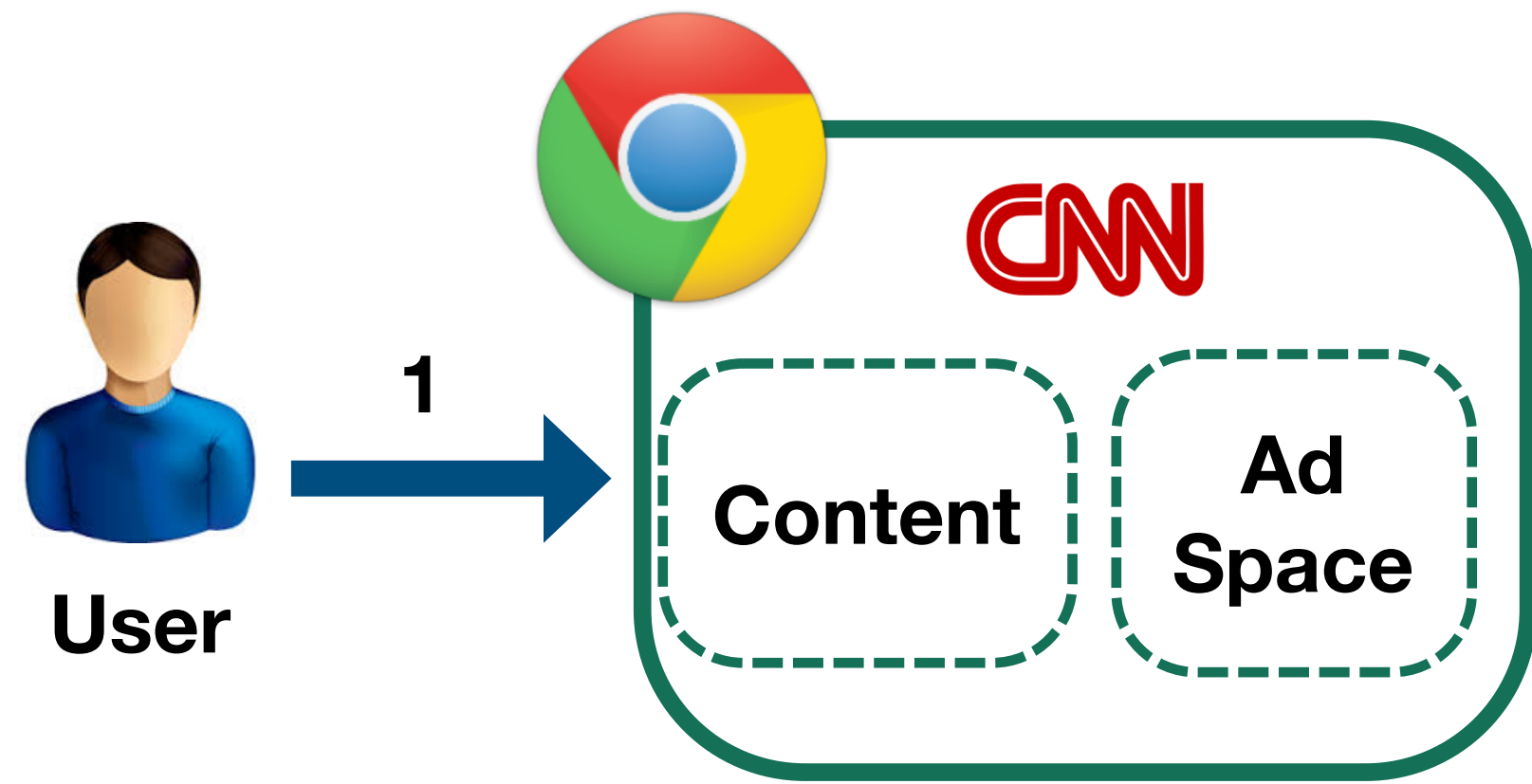


# Ads Under Real Time Bidding (RTB)

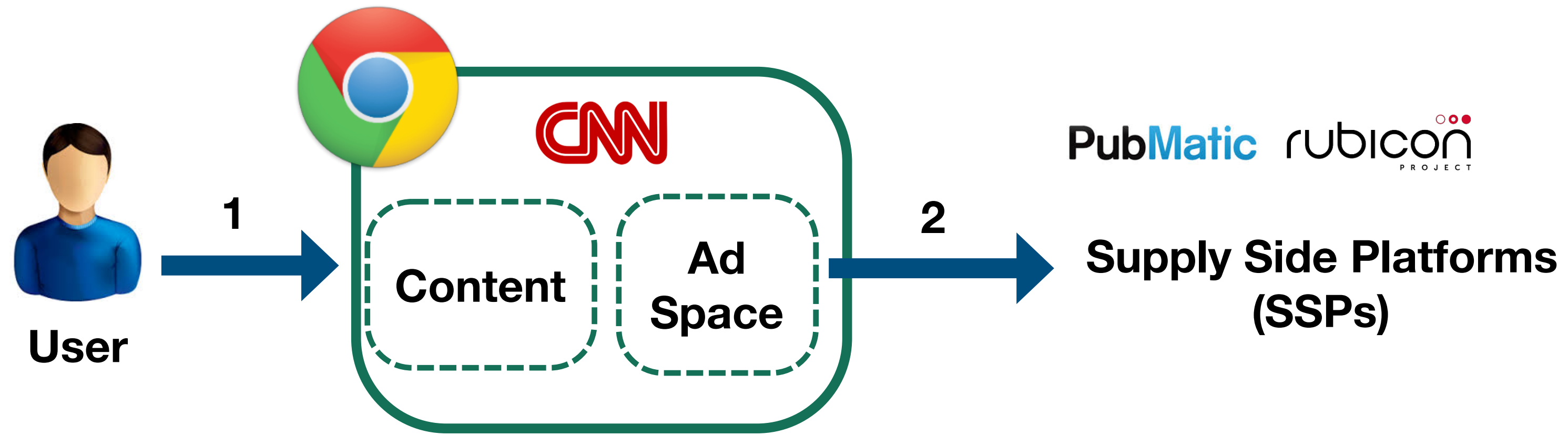
# Ads Under Real Time Bidding (RTB)



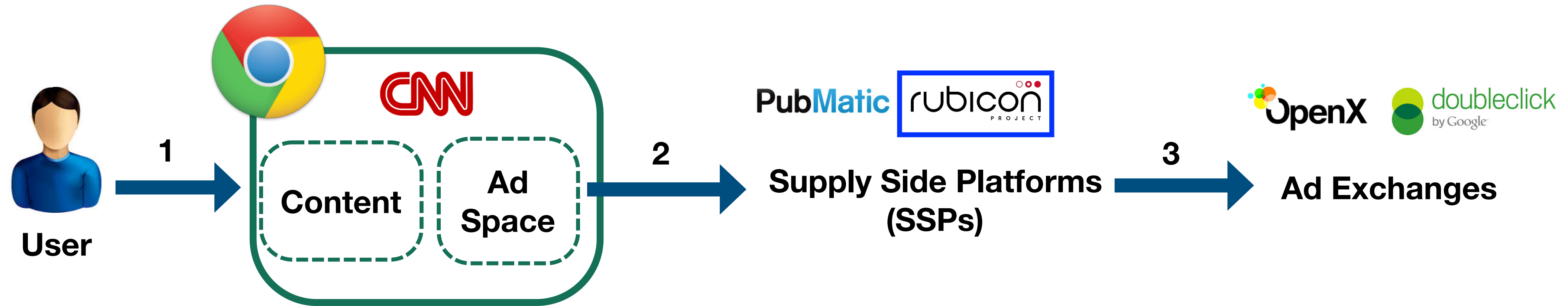
# Ads Under Real Time Bidding (RTB)



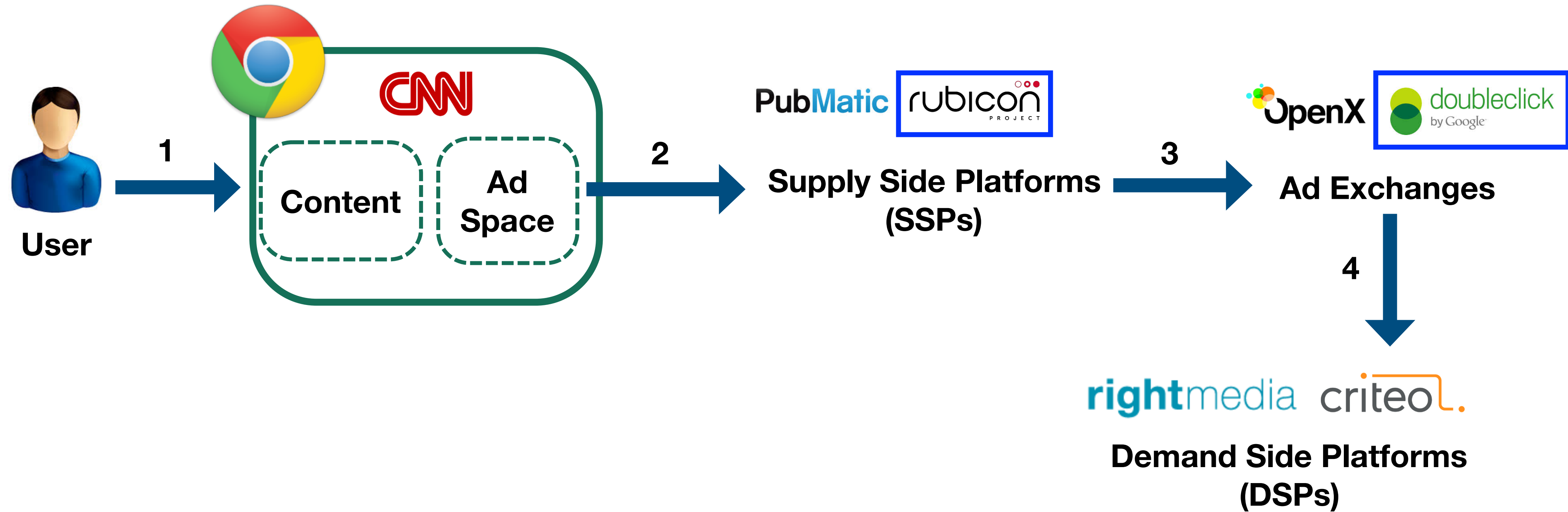
# Ads Under Real Time Bidding (RTB)



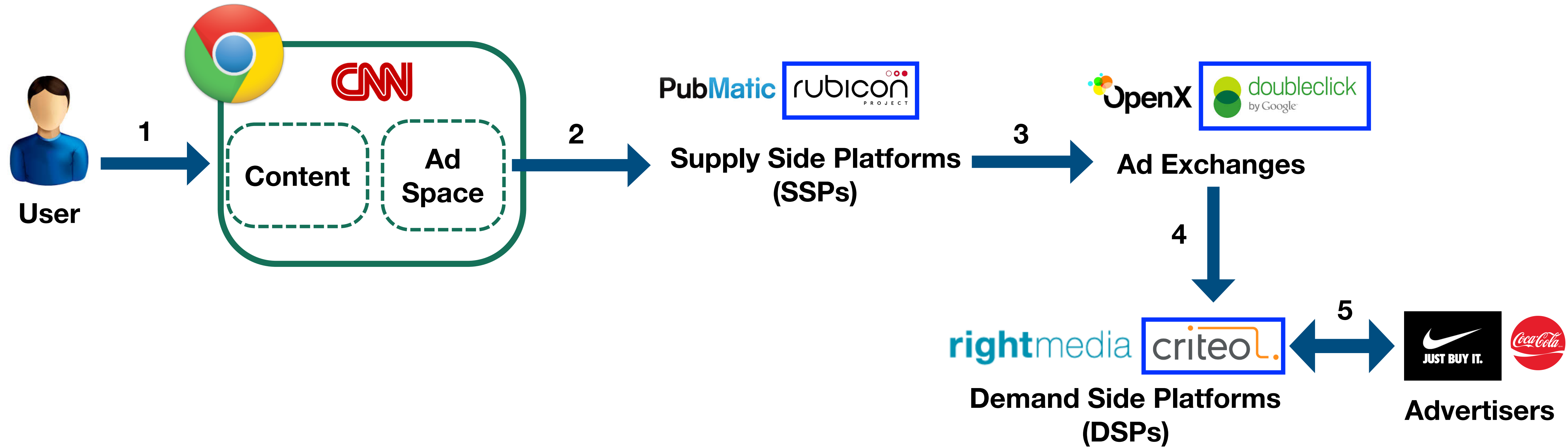
# Ads Under Real Time Bidding (RTB)



# Ads Under Real Time Bidding (RTB)

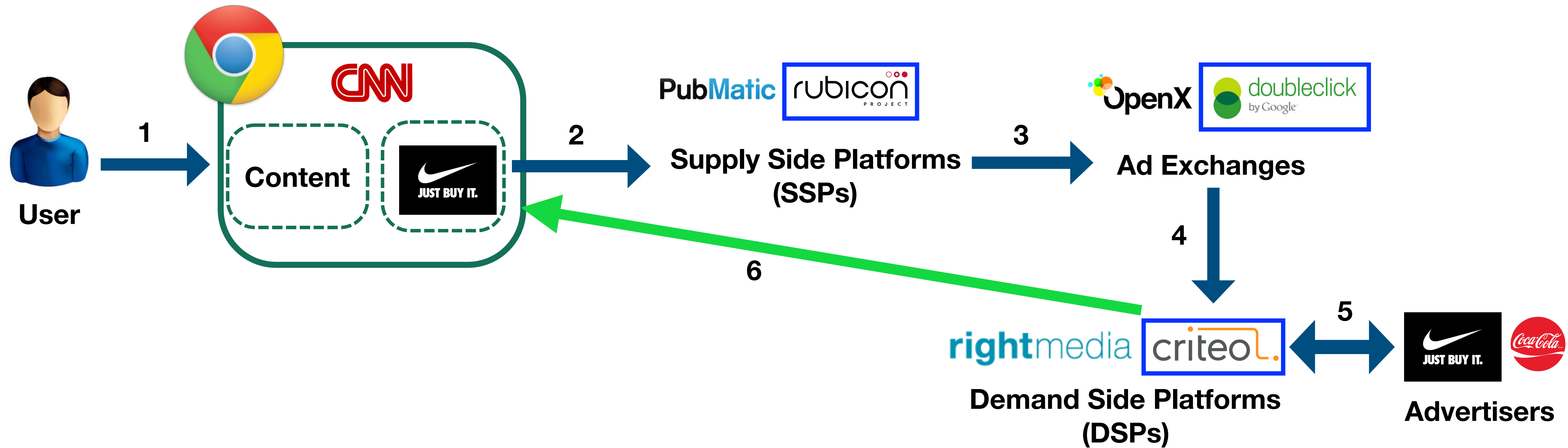


# Ads Under Real Time Bidding (RTB)

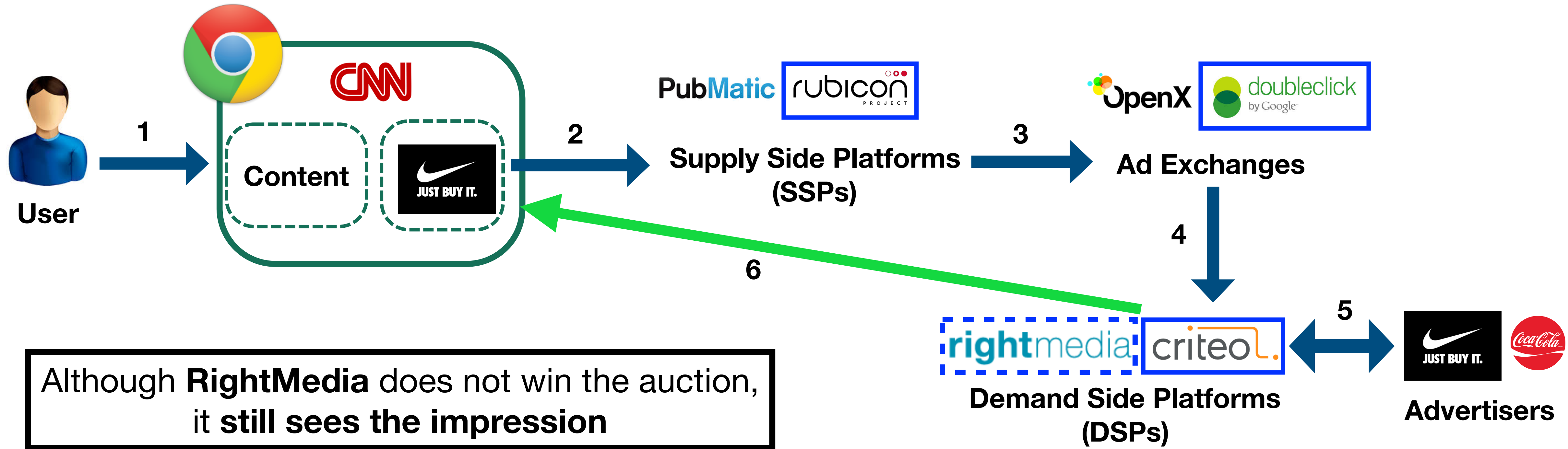




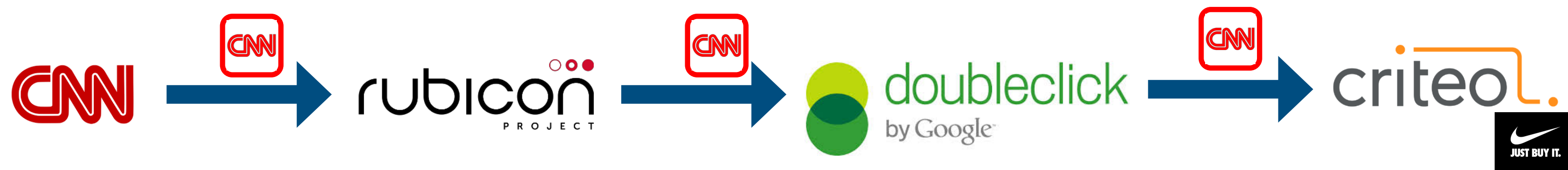
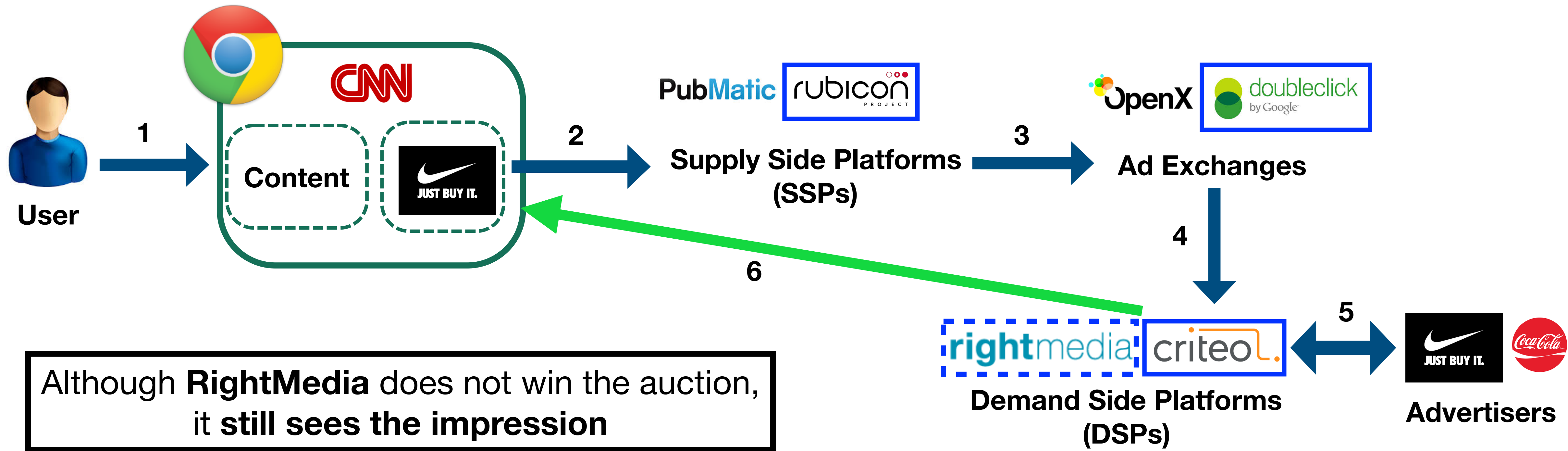
# Ads Under Real Time Bidding (RTB)



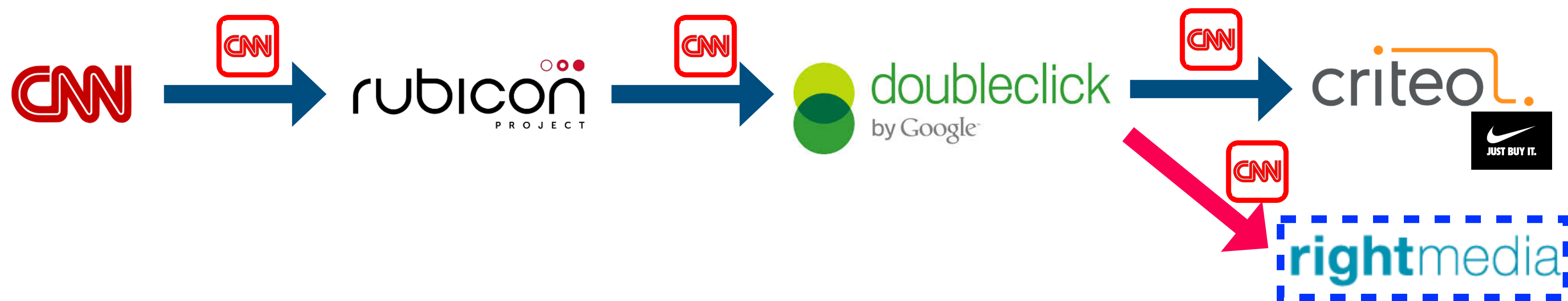
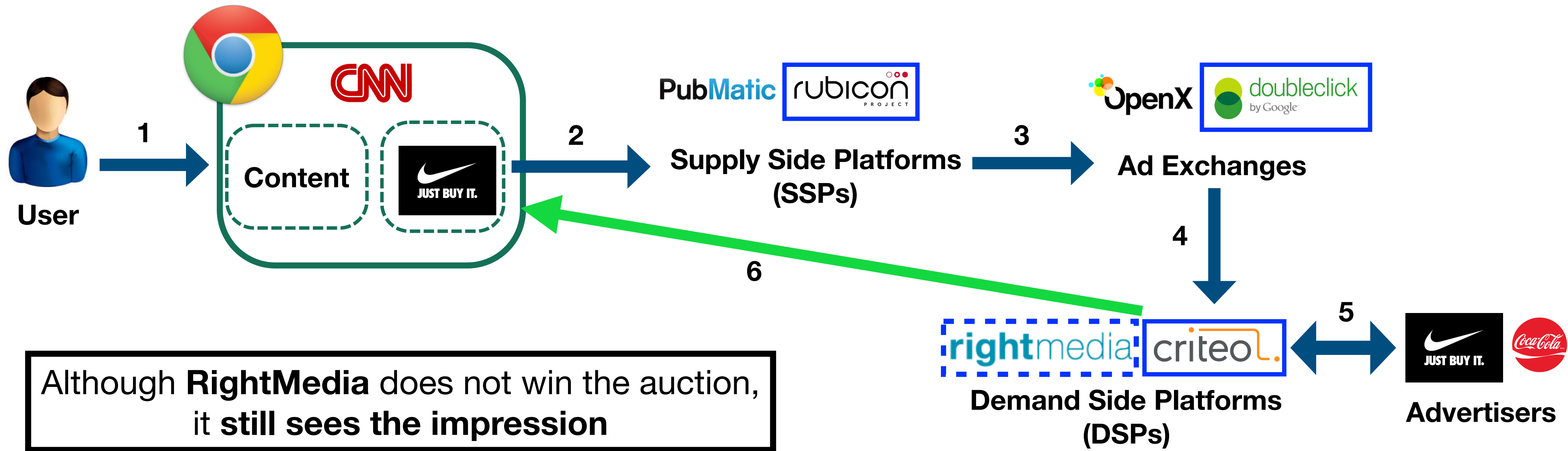
# Ads Under Real Time Bidding (RTB)



# Ads Under Real Time Bidding (RTB)

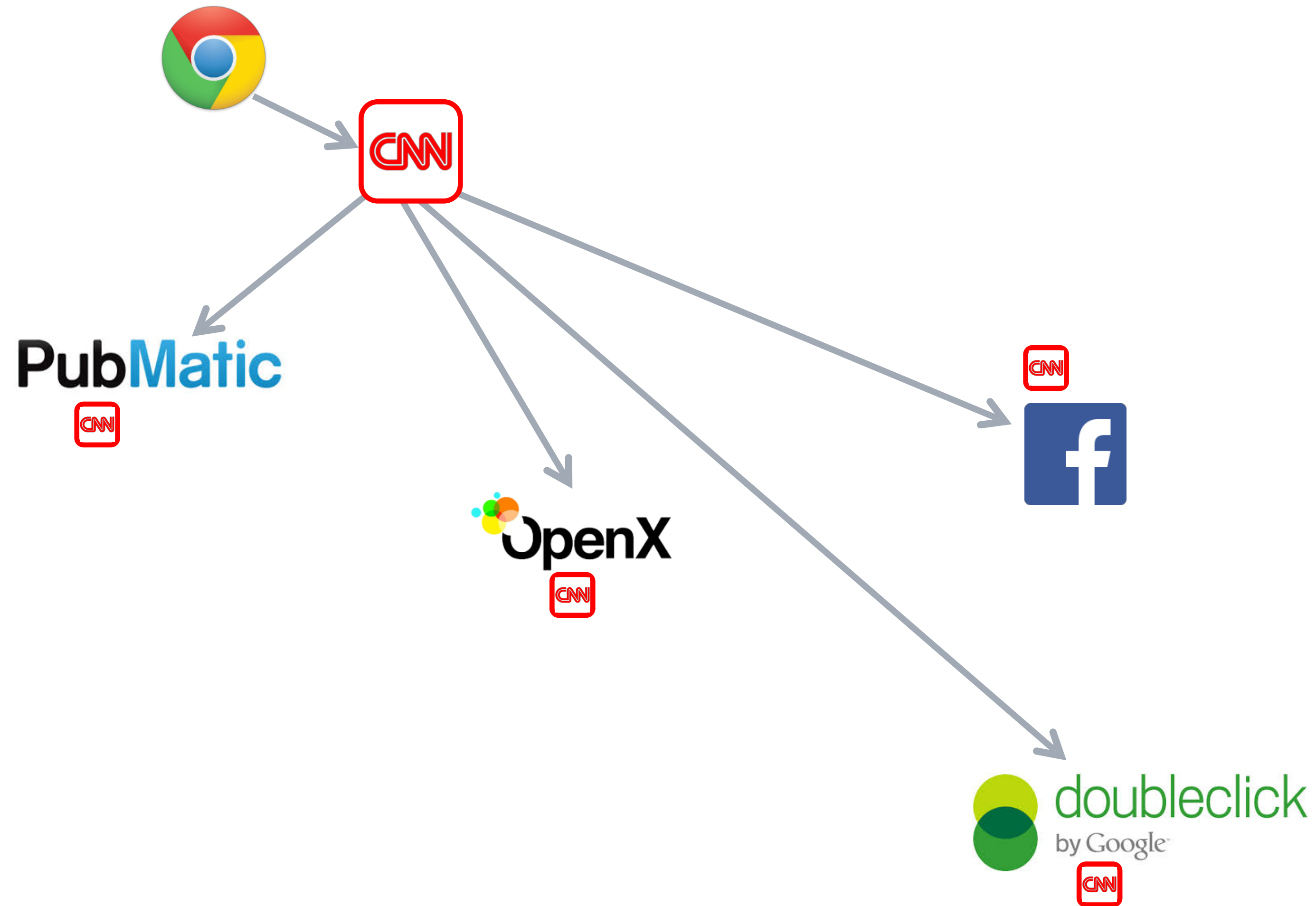


# Ads Under Real Time Bidding (RTB)



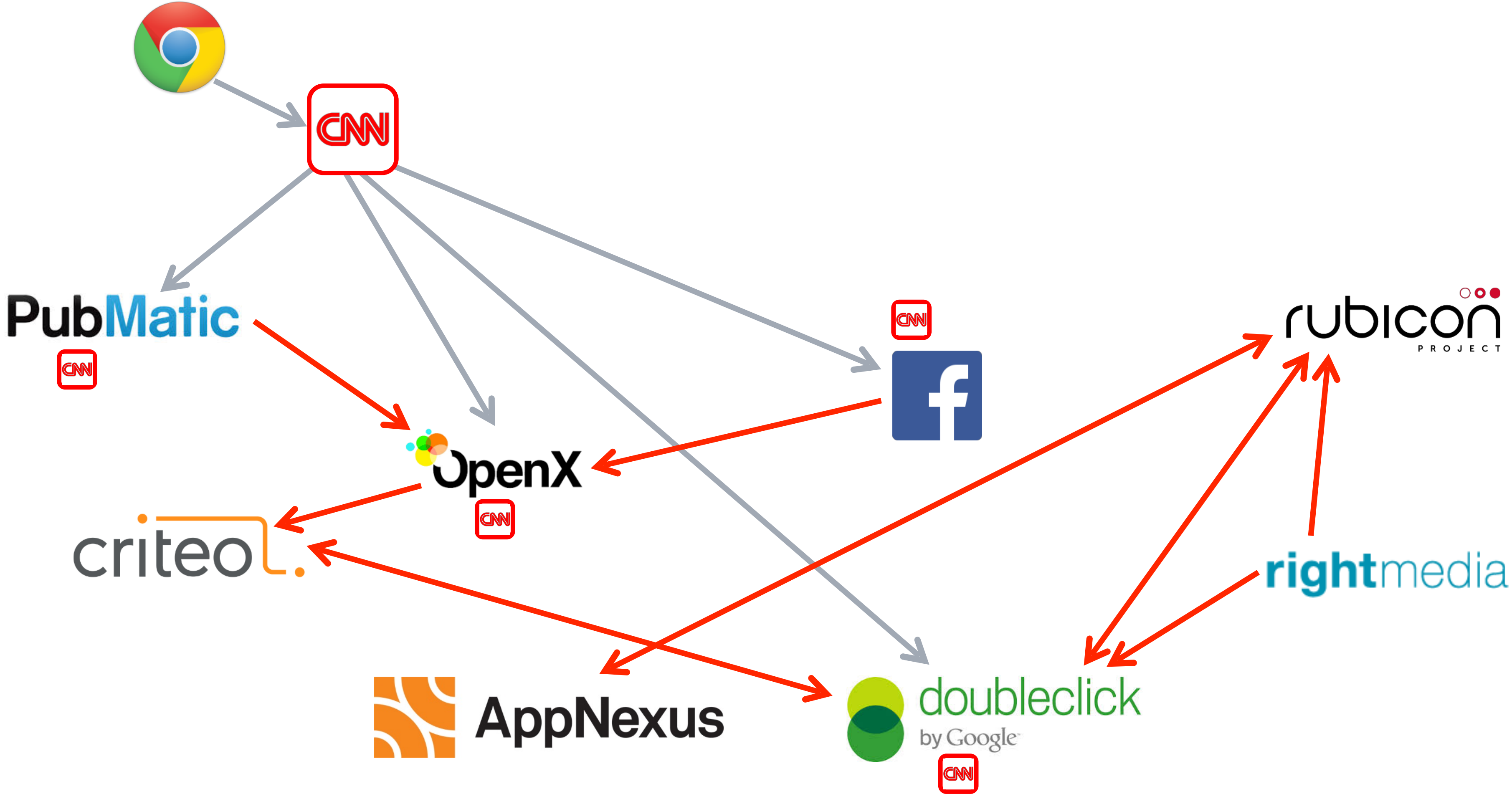
# Adjusted with Real Time Bidding

# Adjusted with Real Time Bidding

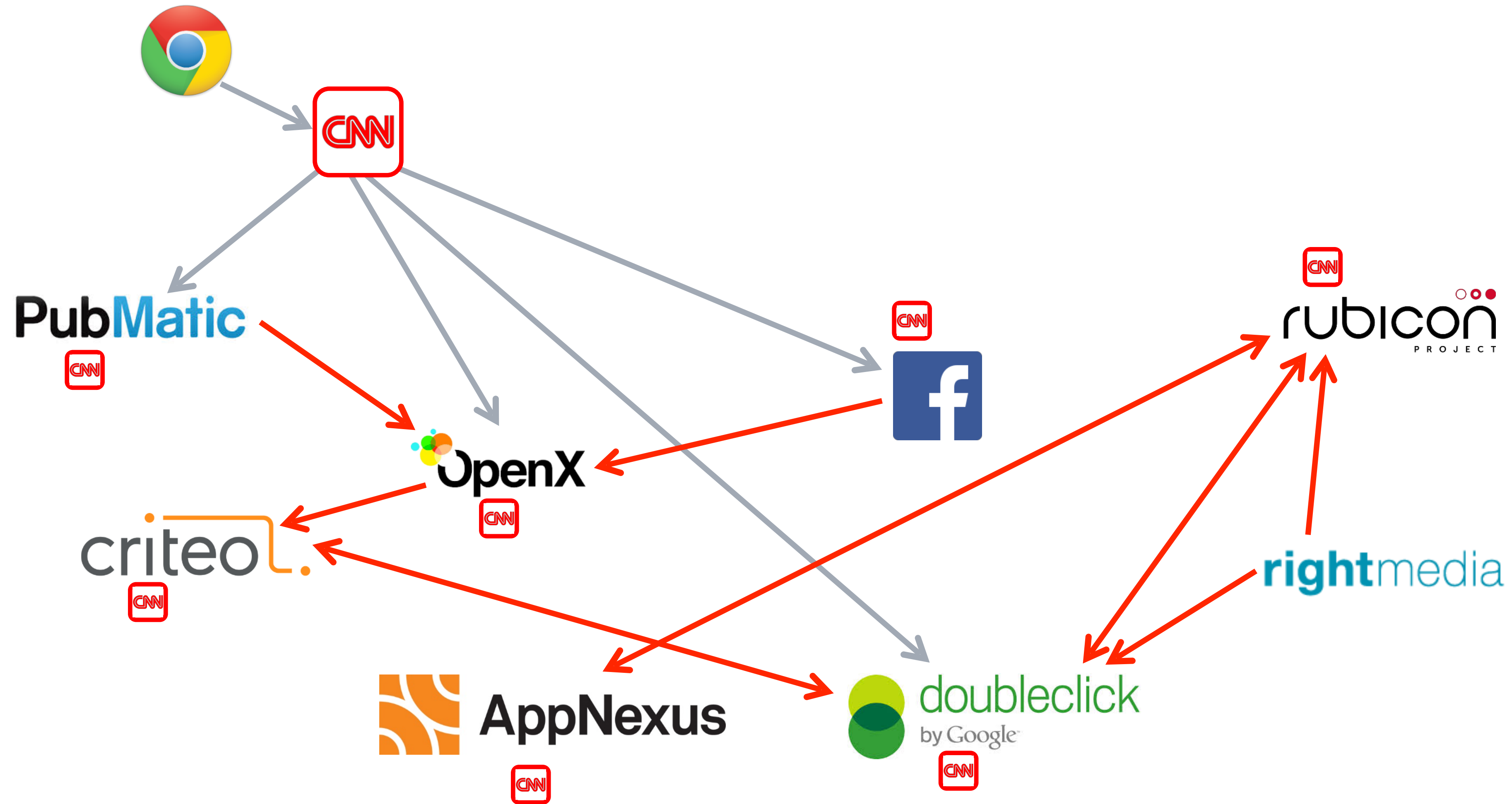




# Adjusted with Real Time Bidding

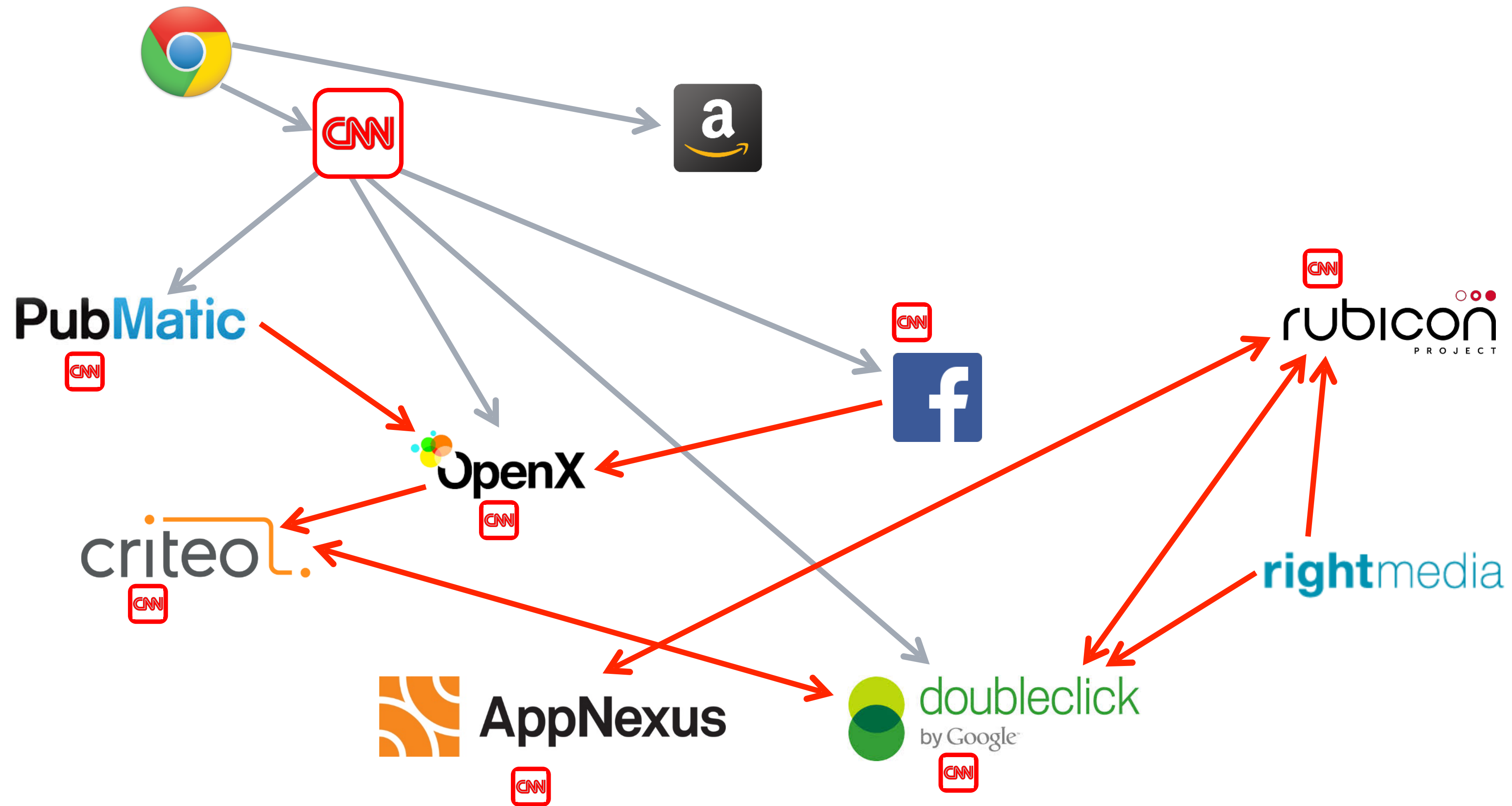


# Adjusted with Real Time Bidding

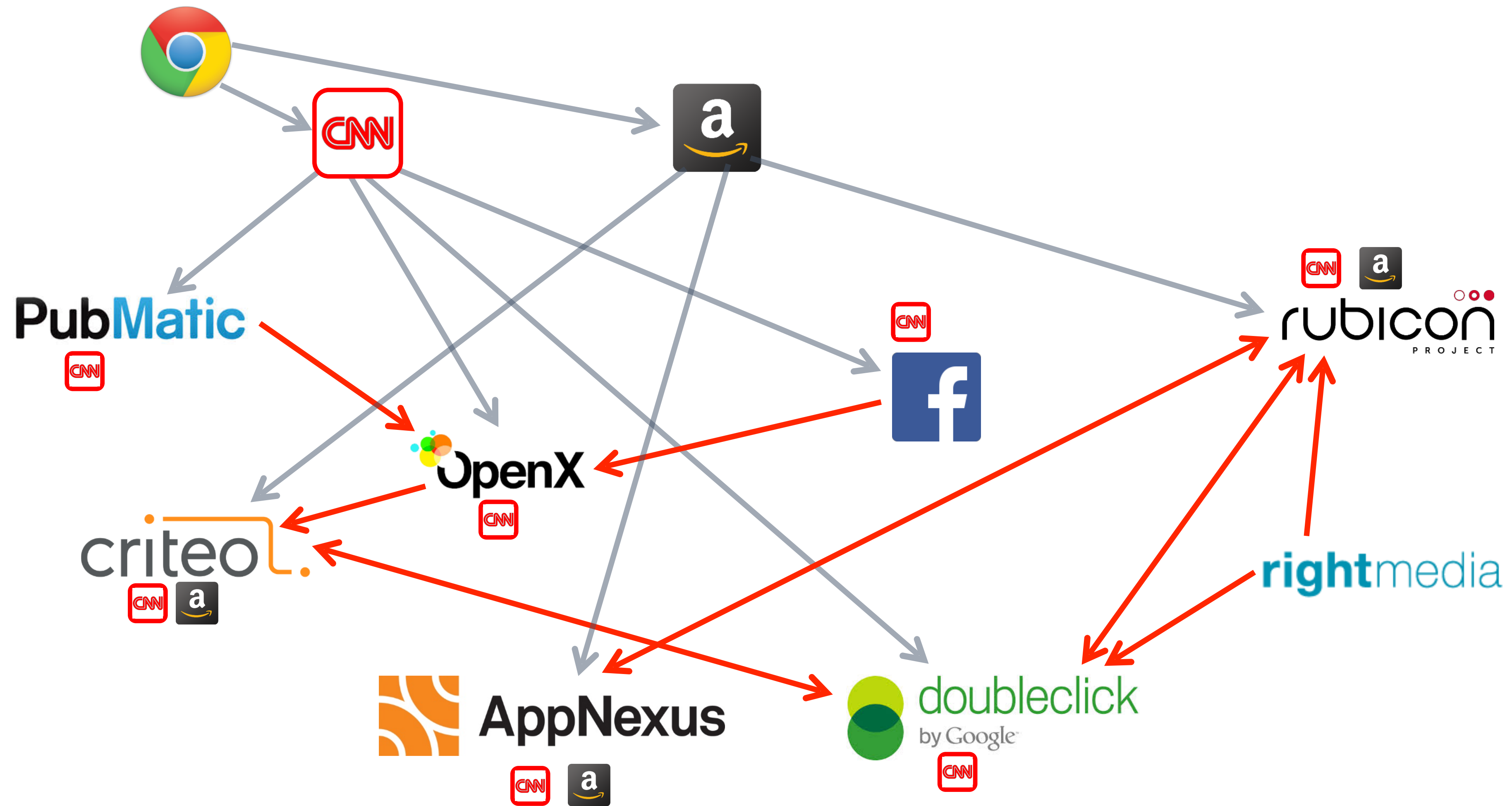




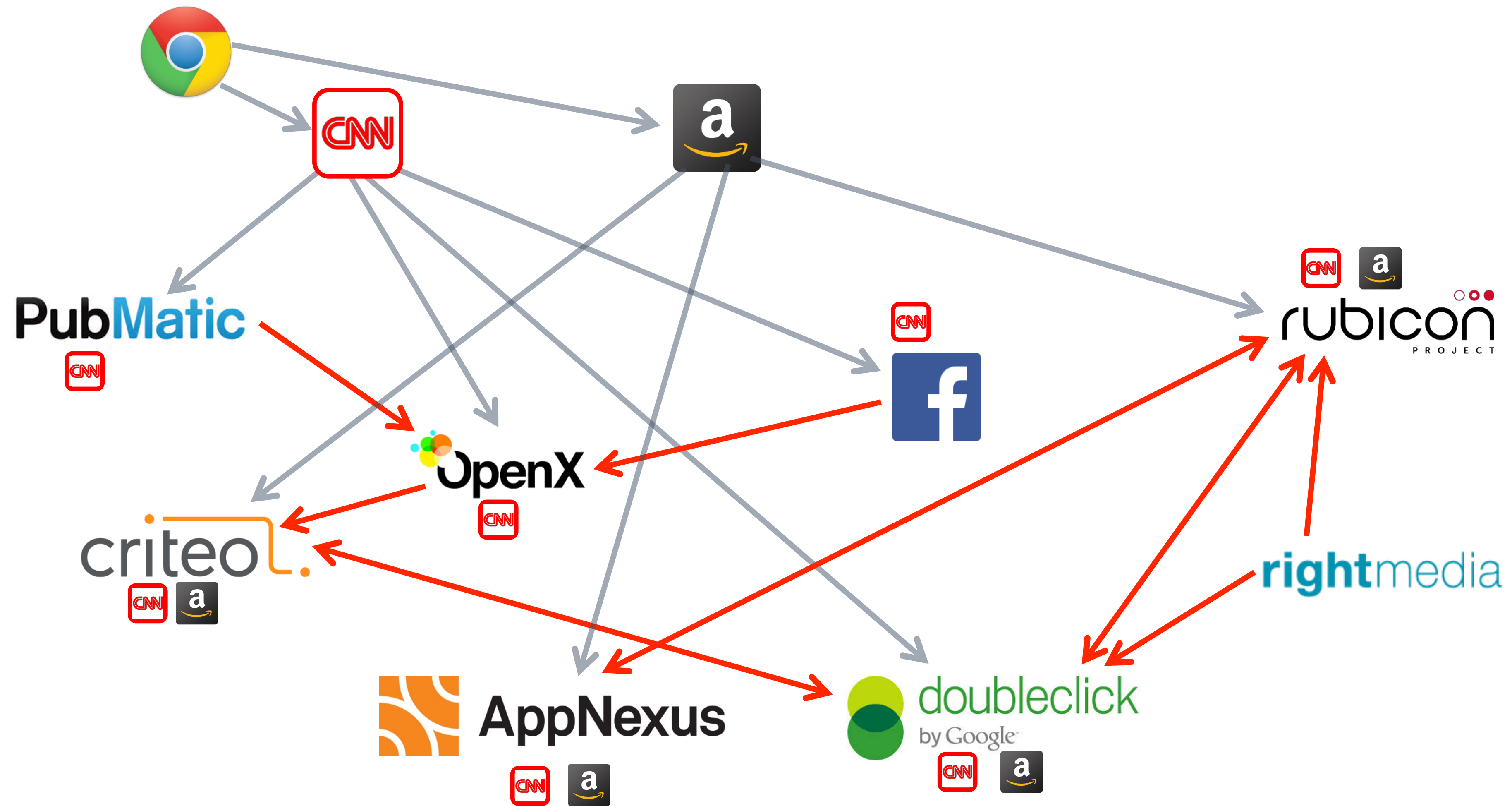
# Adjusted with Real Time Bidding



# Adjusted with Real Time Bidding



# Adjusted with Real Time Bidding



# Goal of the Study

**Model the Diffusion of Impressions in the Advertising Ecosystem**

# Goal of the Study

## **Model the Diffusion of Impressions in the Advertising Ecosystem**

- We want to take Real Time Bidding (RTB) into account
  - This goes beyond the current techniques being used

# Goal of the Study

## Model the Diffusion of Impressions in the Advertising Ecosystem

- We want to take Real Time Bidding (RTB) into account
  - This goes beyond the current techniques being used
- We want to answer the following:
  1. What fraction of user impressions are viewed by ad companies?
  2. How much ad and tracker blocking extensions help?

# Goal of the Study

## Model the Diffusion of Impressions in the Advertising Ecosystem

- We want to take Real Time Bidding (RTB) into account
  - This goes beyond the current techniques being used
- We want to answer the following:
  1. What fraction of user impressions are viewed by ad companies?
  2. How much ad and tracker blocking extensions help?

### Key Terms:

1. **Impressions:** Page Visits
2. **Publishers:** First party websites visited by users (e.g. cnn, bbc, espn)
3. **A&A:** Advertising and Analytics related companies / domains

# Overview

1. Dataset used in our study
2. Our Simulations
3. Results
4. Ad & Tracker Blocking



# Dataset Used

# Dataset Used

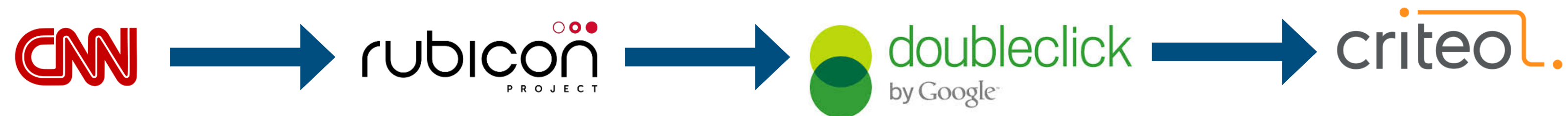
- We use data from our prior work.<sup>1</sup>
- We trained personas using 738 popular e-commerce websites and solicit retargeted ads through 150 top publishers.

# Dataset Used

- We use data from our prior work.<sup>1</sup>
- We trained personas using 738 popular e-commerce websites and solicit retargeted ads through 150 top publishers.
- Using this data, we have inclusion chains of all resources

# Dataset Used

- We use data from our prior work.<sup>1</sup>
- We trained personas using 738 popular e-commerce websites and solicit retargeted ads through 150 top publishers.
- Using this data, we have inclusion chains of all resources

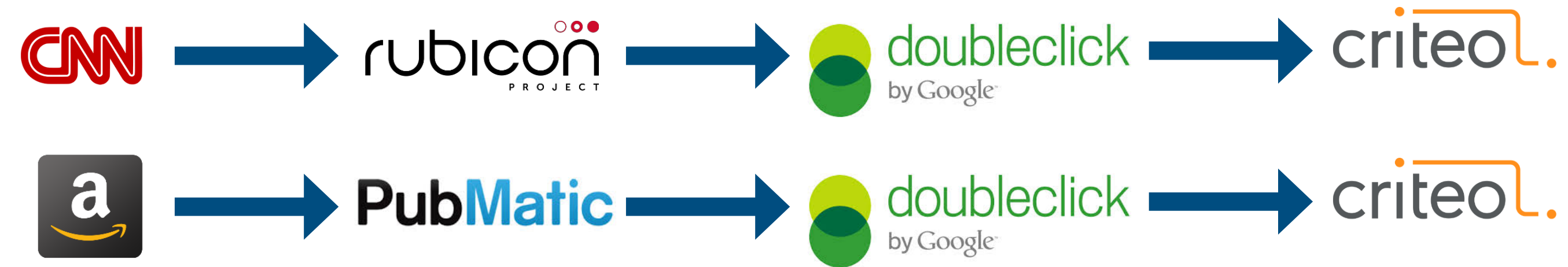


## Example Inclusion Chain

# From Chains to Graph

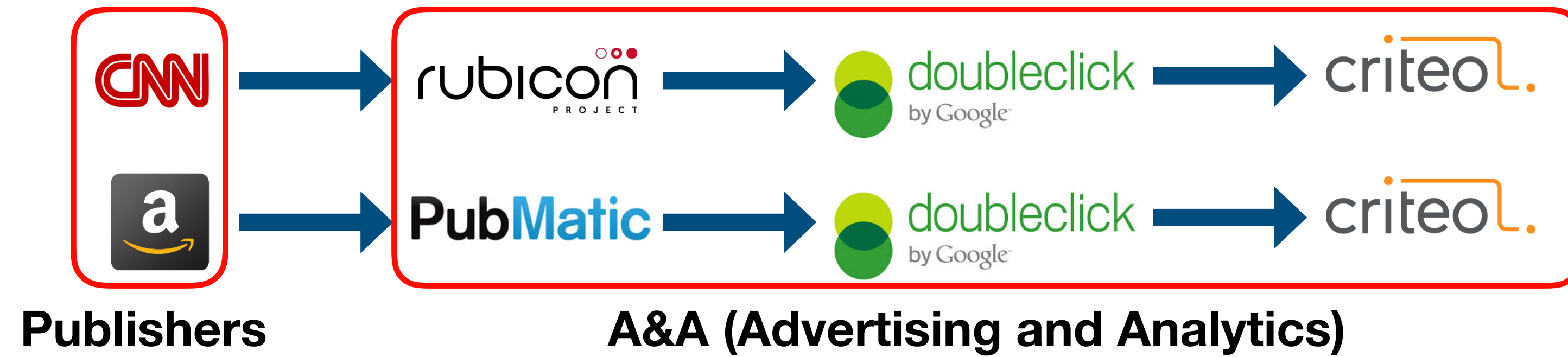
# From Chains to Graph

## Inclusion Chains



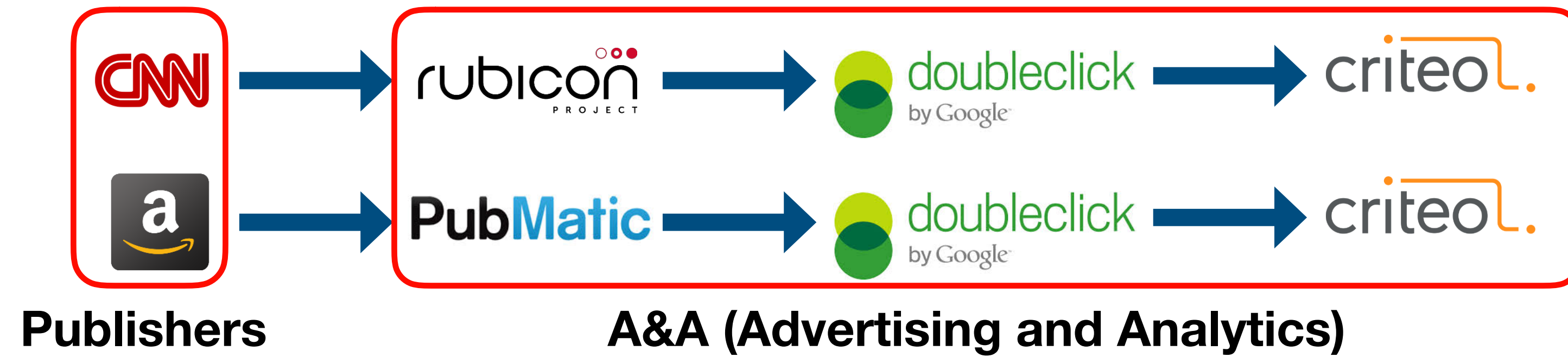
# From Chains to Graph

## Inclusion Chains



# From Chains to Graph

## Inclusion Chains

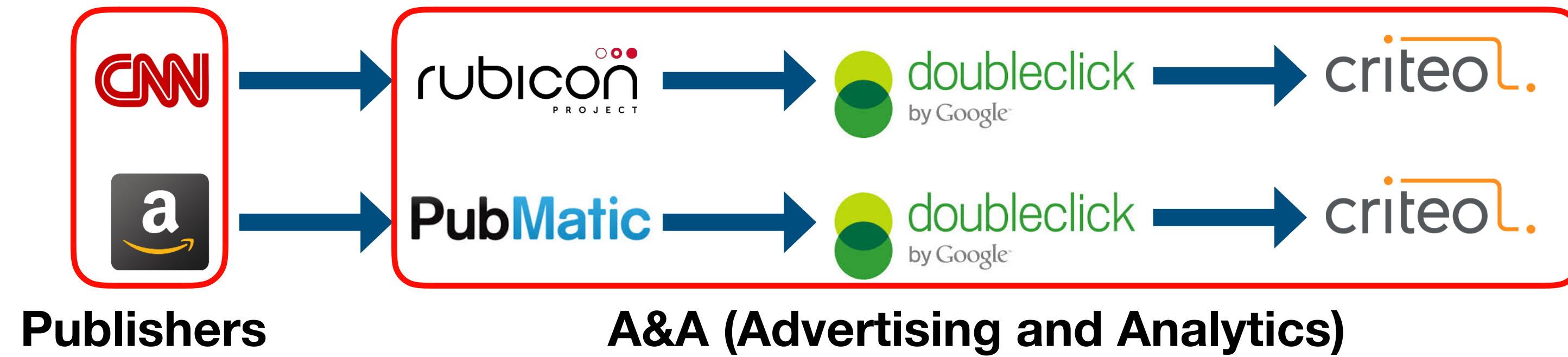


## Graph Representation

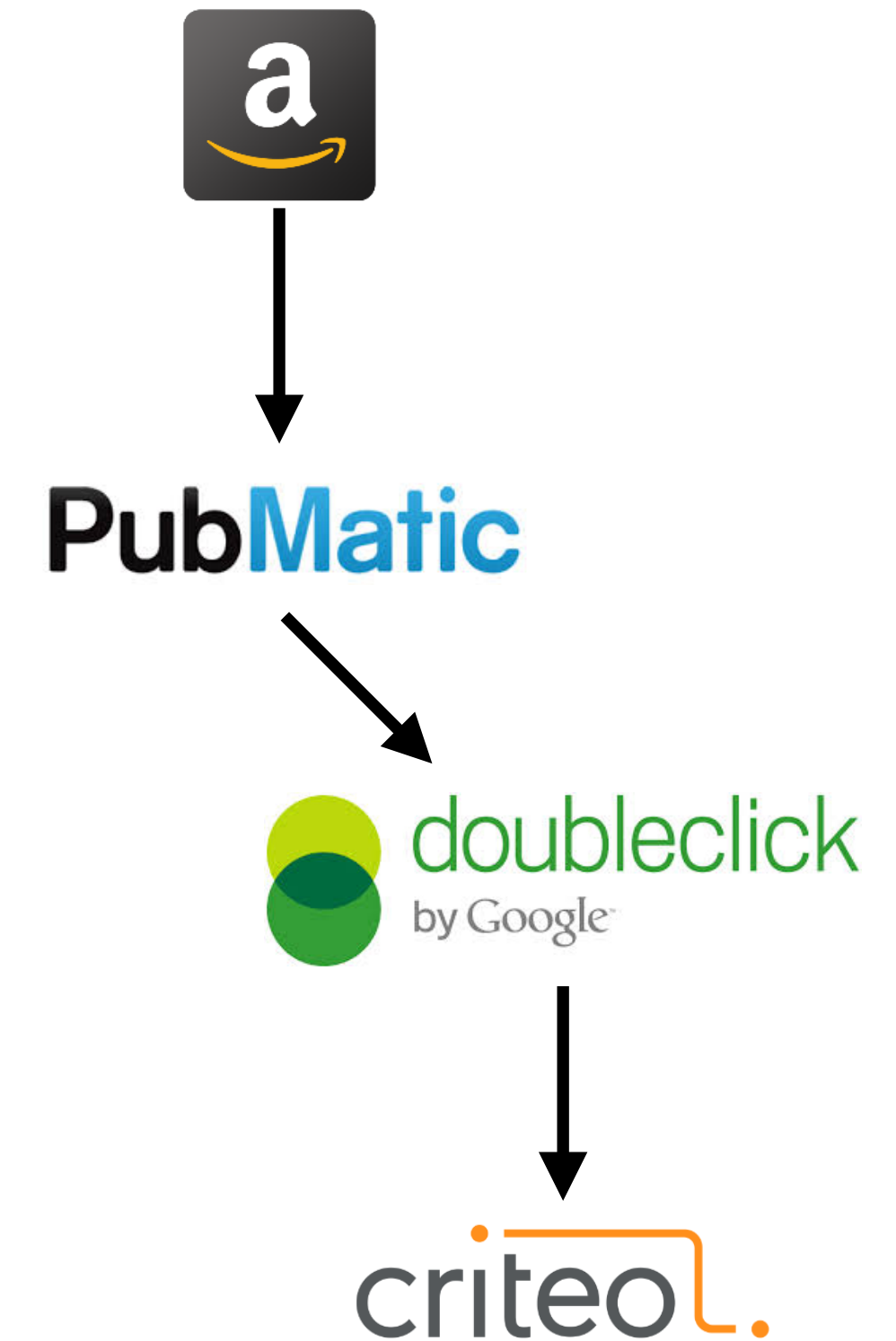


# From Chains to Graph

## Inclusion Chains

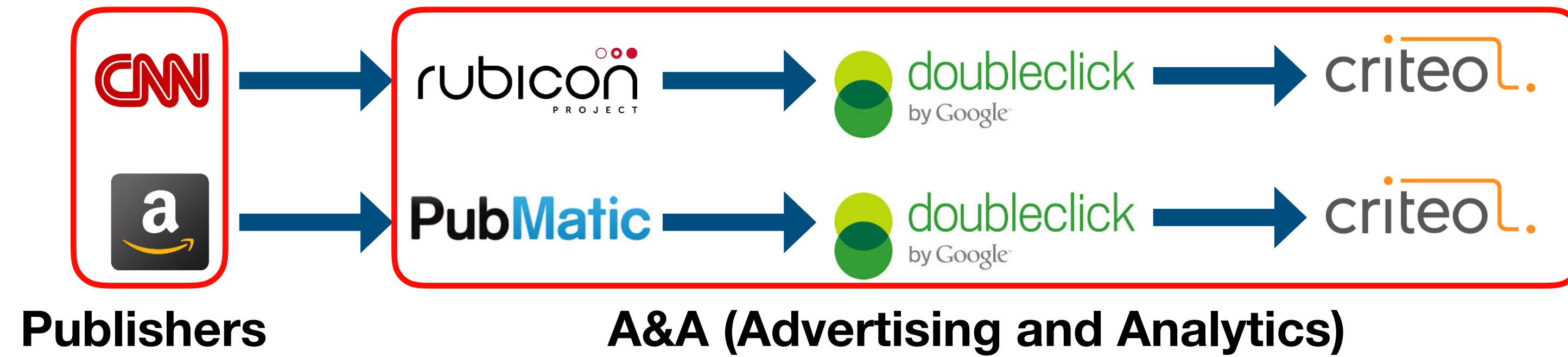


## Graph Representation



# From Chains to Graph

## Inclusion Chains

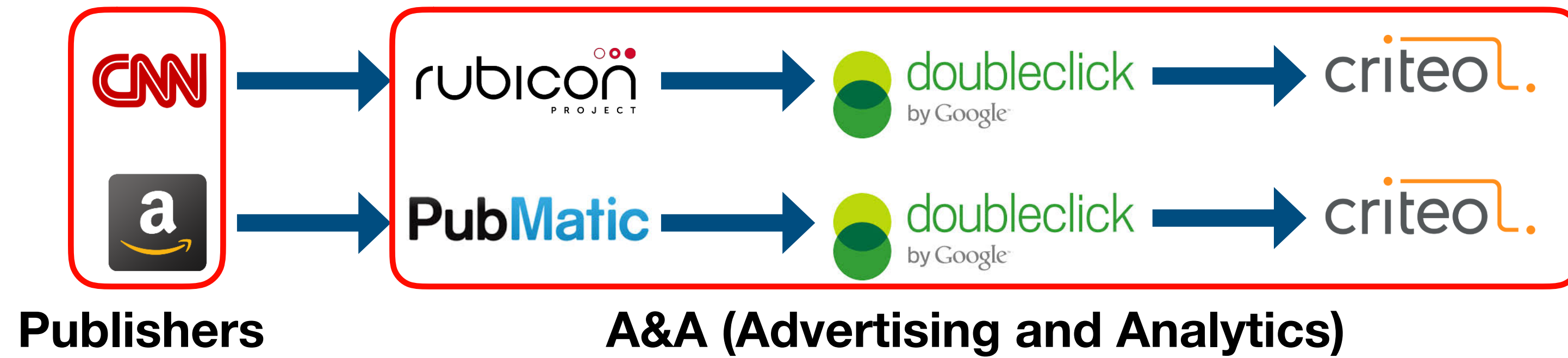


## Graph Representation

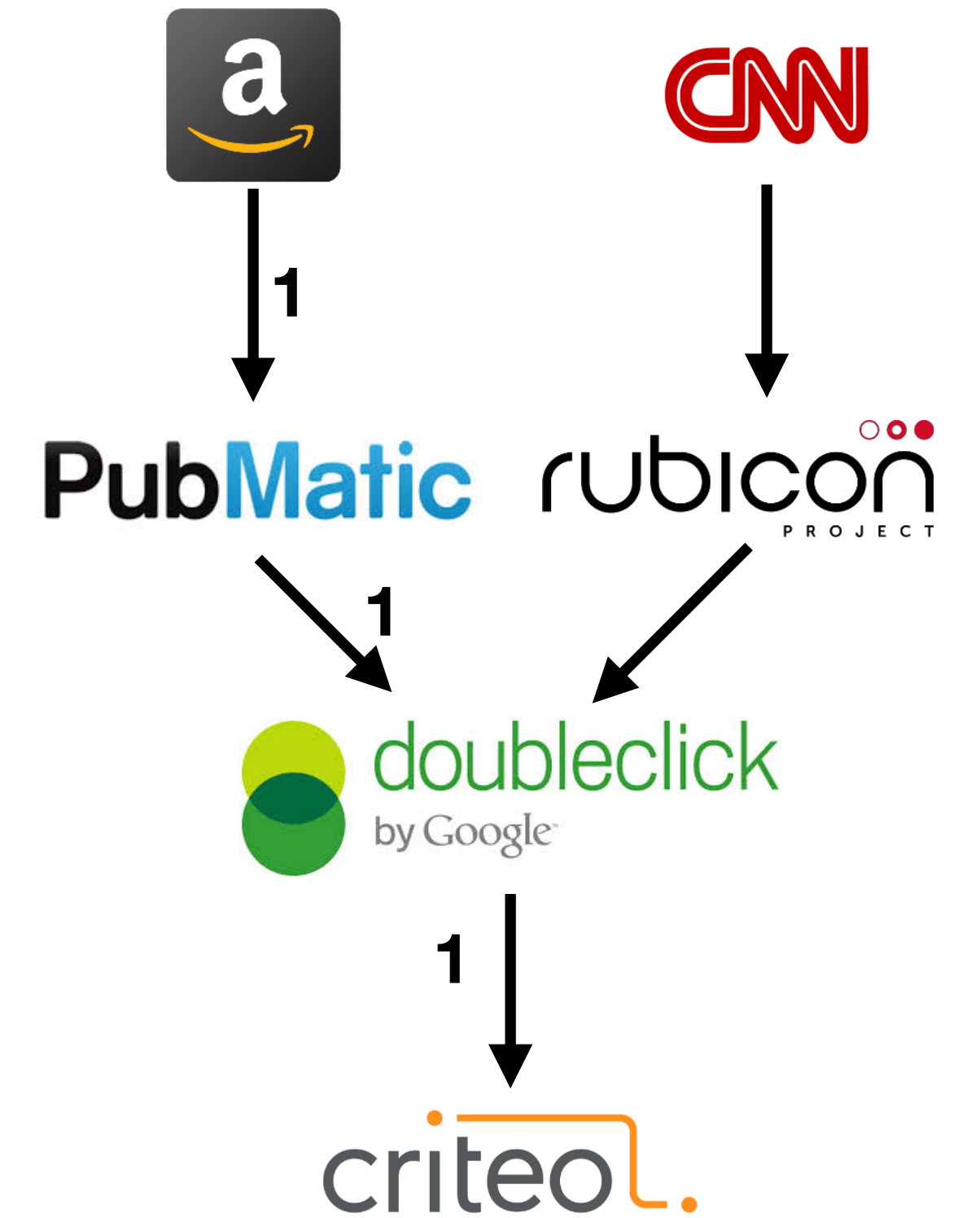


# From Chains to Graph

## Inclusion Chains

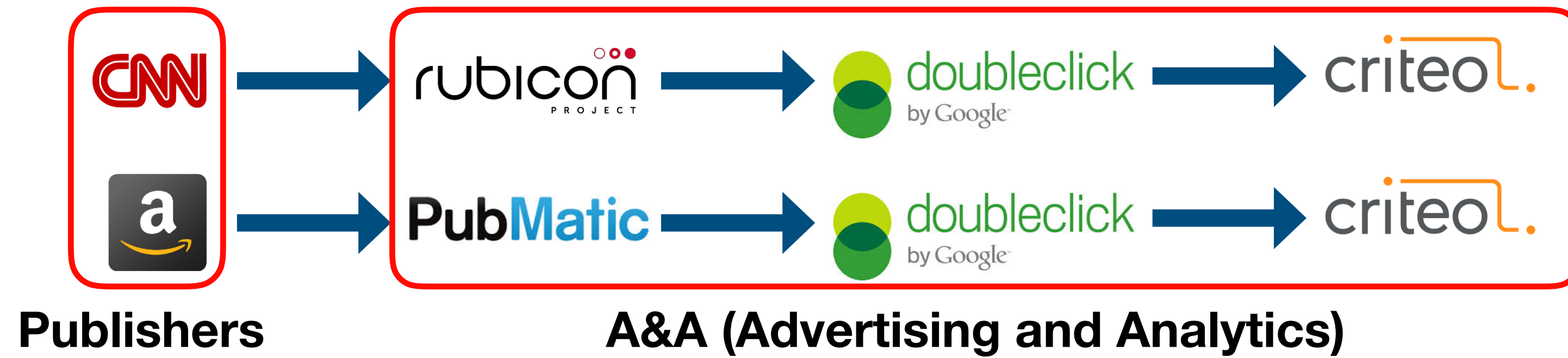


## Graph Representation

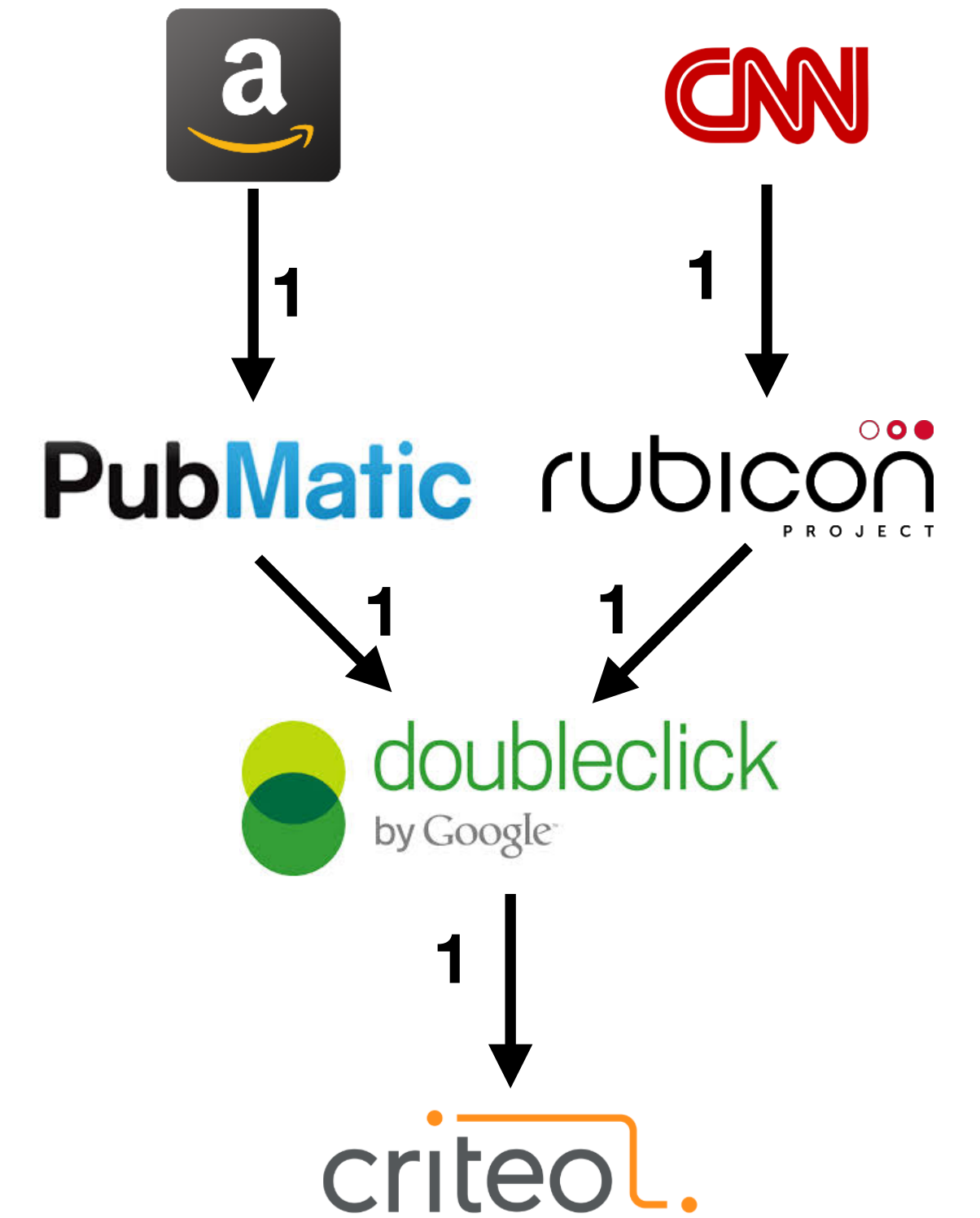


# From Chains to Graph

## Inclusion Chains

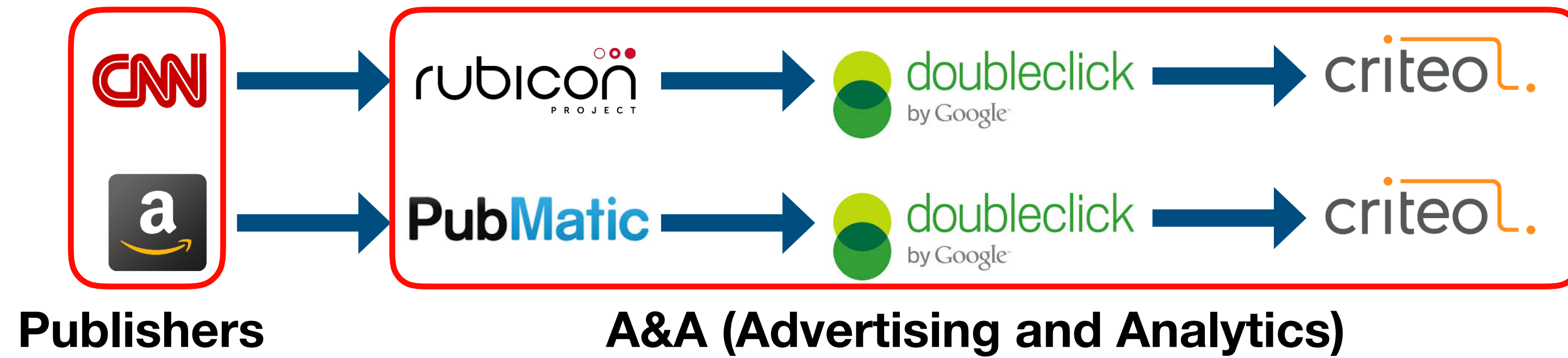


## Graph Representation

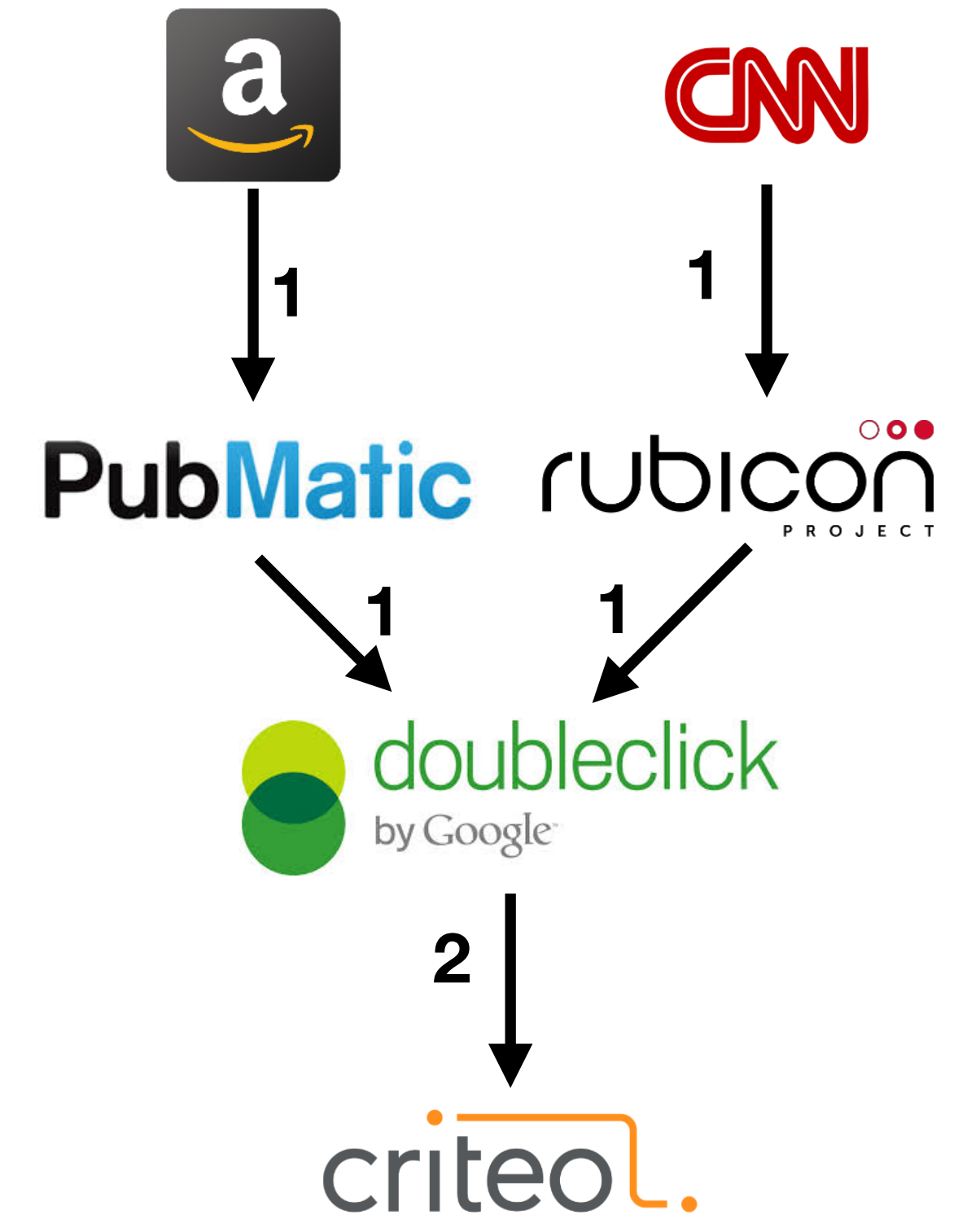


# From Chains to Graph

## Inclusion Chains



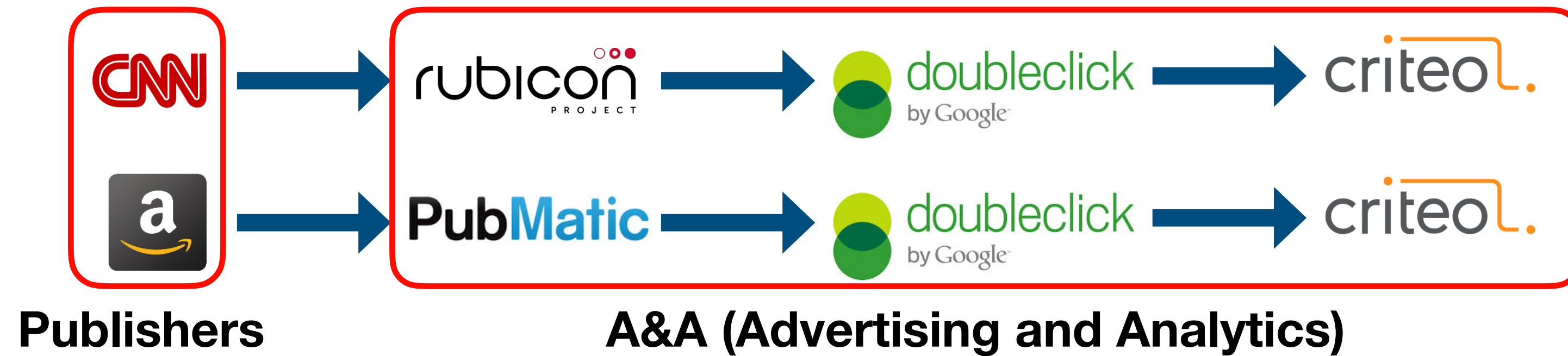
## Graph Representation



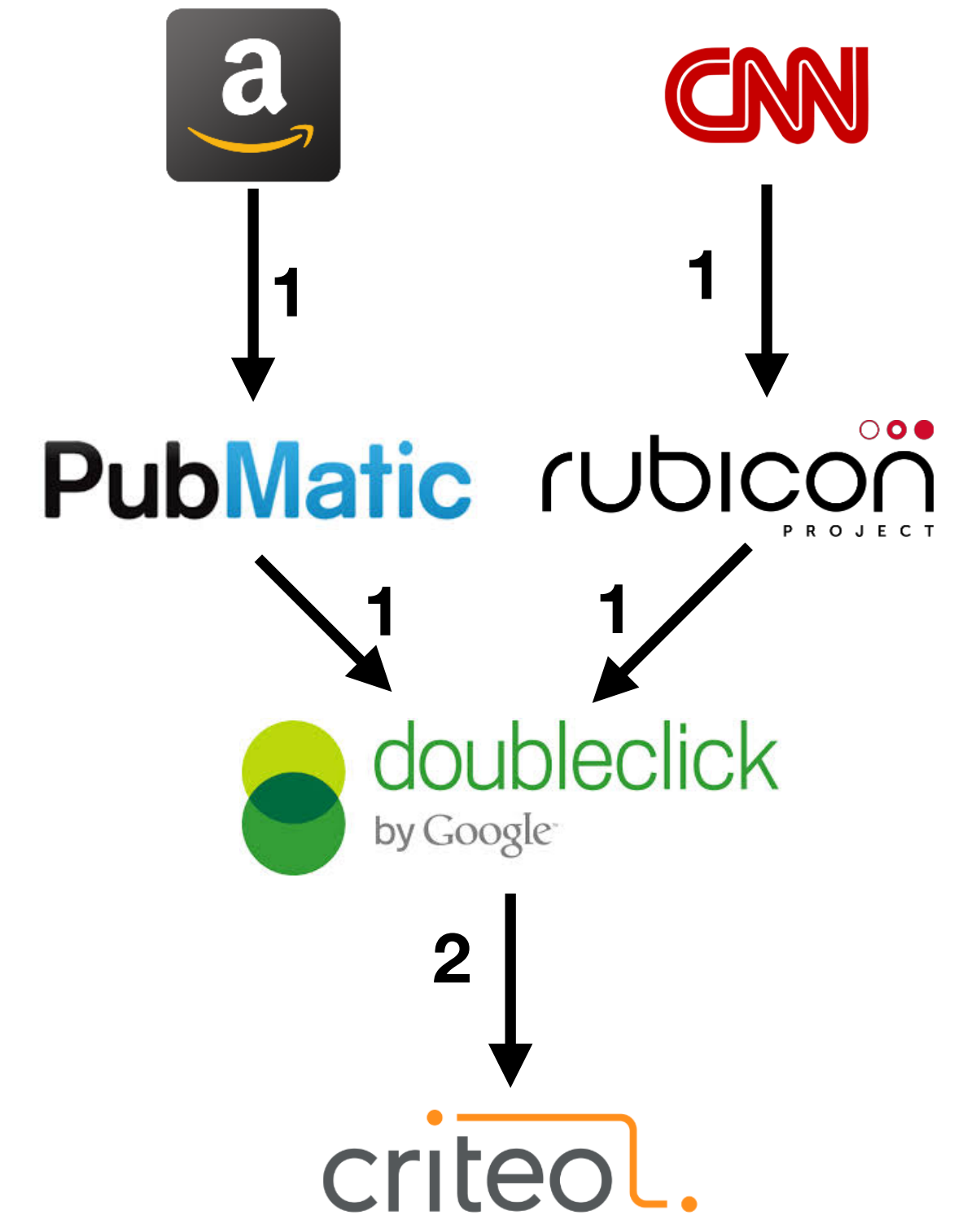


# From Chains to Graph

## Inclusion Chains



## Graph Representation

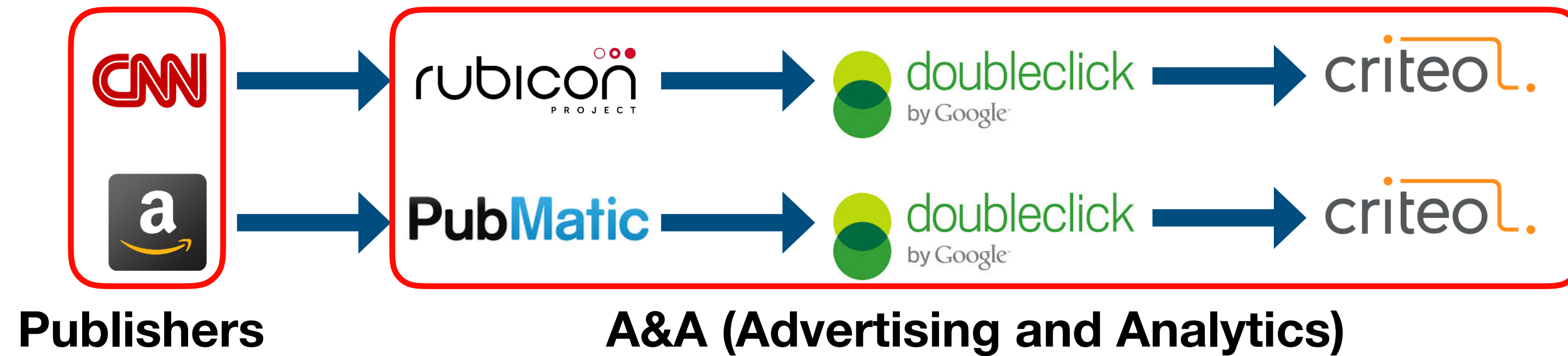


Nodes: Publishers or A&A domains

Edges: Publisher  $\rightarrow$  A&A  
A&A  $\rightarrow$  A&A

# From Chains to Graph

## Inclusion Chains



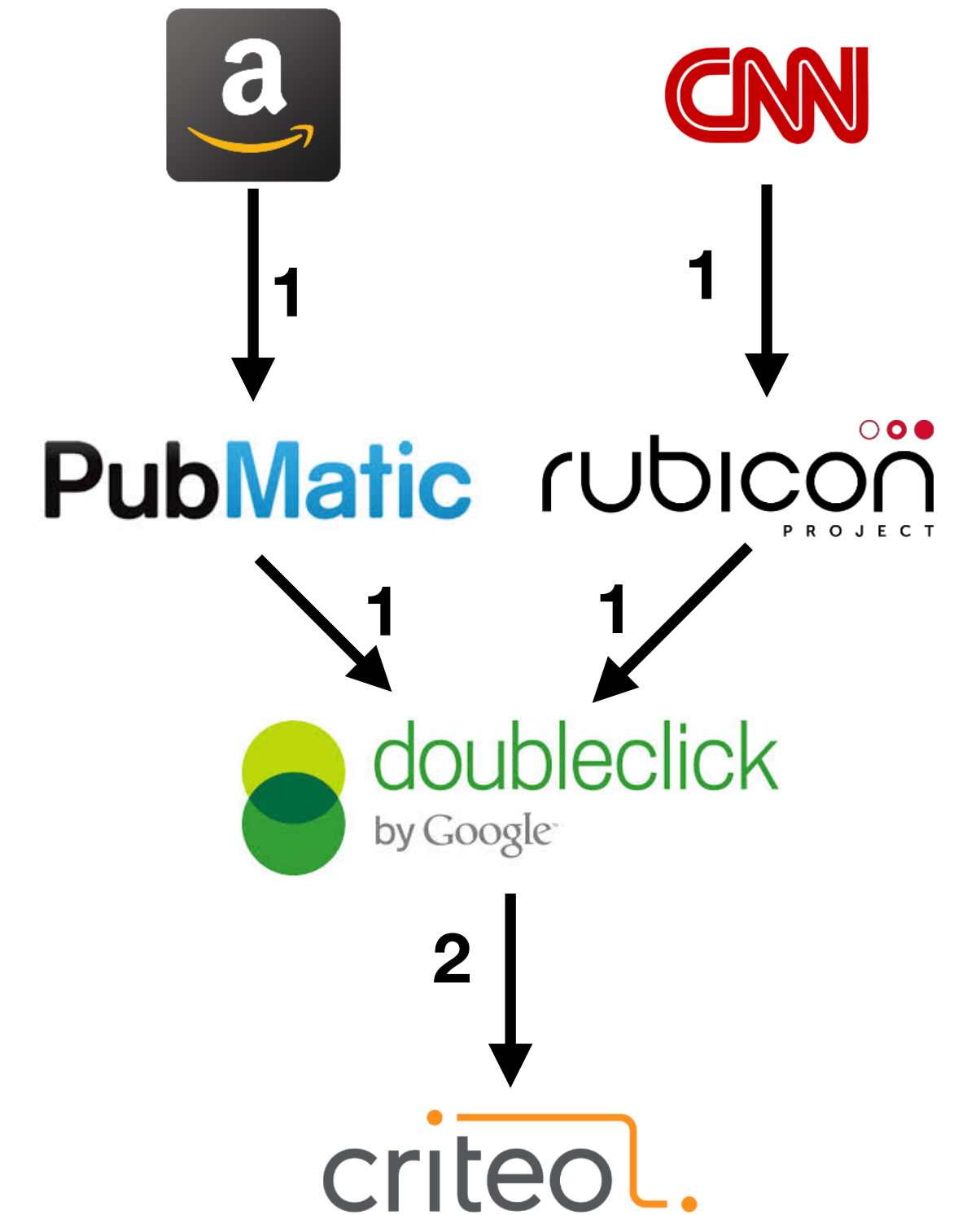
### Nodes:

- **Total:** ~1.9K
- **A&A:** ~1K

### Edges:

- **Total:** ~26K
- **Pub → A&A:** ~10.5K
- **A&A → A&A:** ~15.5K

## Graph Representation



**Nodes:** Publishers or A&A domains

**Edges:** Publisher → A&A  
A&A → A&A

# Some Properties of the Graph



# Some Properties of the Graph

- The graph is very dense.

# Some Properties of the Graph

- The graph is very dense.
- No distinct communities
  - Web is not necessary balkanized into distinct groups

# Some Properties of the Graph

- The graph is very dense.
- No distinct communities
  - Web is not necessary balkanized into distinct groups
- Expected top nodes with PageRank and Betweenness Centrality



# Goal of the Study

**Model the Diffusion of Impressions in the Advertising Ecosystem**

1. What fraction of user impressions are viewed by ad companies?
2. How much ad and tracker blocking extensions help?

# Our Simulation Setup

# Our Simulation Setup

We simulate browsing traces for 200 users using method from [1].


# Our Simulation Setup

We simulate browsing traces for 200 users using method from [1].

1. User generates an impression on  $N$  selected publishers.

# Our Simulation Setup


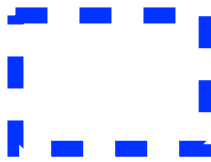
We simulate browsing traces for 200 users using method from [1].

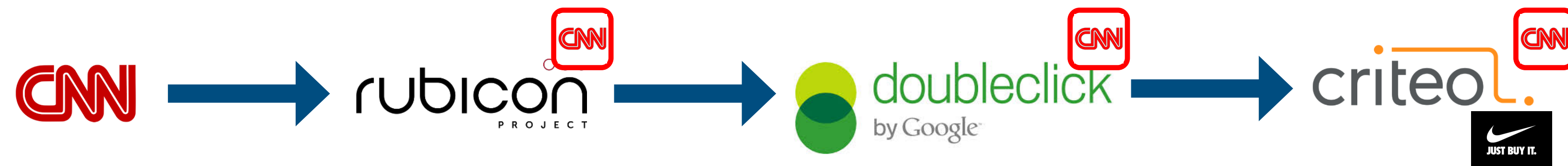
1. User generates an impression on N selected publishers.
2. Impressions are forwarded to A&A domains via:
  - A. **Direct Propagation:** 
    - Present on publisher or won RTB auction. **Observable** (goes through the browser)
  - B. **Indirect Propagation:** 
    - A&A domains learn impressions through RTB participation. **Non-observable**



# Our Simulation Setup

We simulate browsing traces for 200 users using method from [1].

1. User generates an impression on N selected publishers.
2. Impressions are forwarded to A&A domains via:
  - A. **Direct Propagation:** 
    - Present on publisher or won RTB auction. **Observable** (goes through the browser)
  - B. **Indirect Propagation:** 
    - A&A domains learn impressions through RTB participation. **Non-observable**



# Our Simulation Setup

We simulate browsing traces for 200 users using method from [1].

1. User generates an impression on N selected publishers.

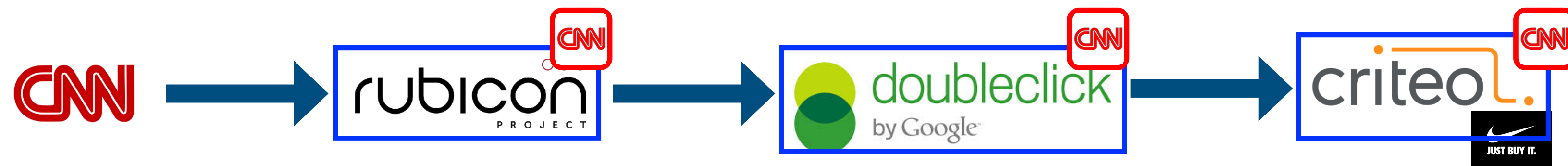
2. Impressions are forwarded to A&A domains via:

A. **Direct Propagation:** 

- Present on publisher or won RTB auction. **Observable** (goes through the browser)

B. **Indirect Propagation:** 

- A&A domains learn impressions through RTB participation. **Non-observable**



# Our Simulation Setup

We simulate browsing traces for 200 users using method from [1].

1. User generates an impression on N selected publishers.

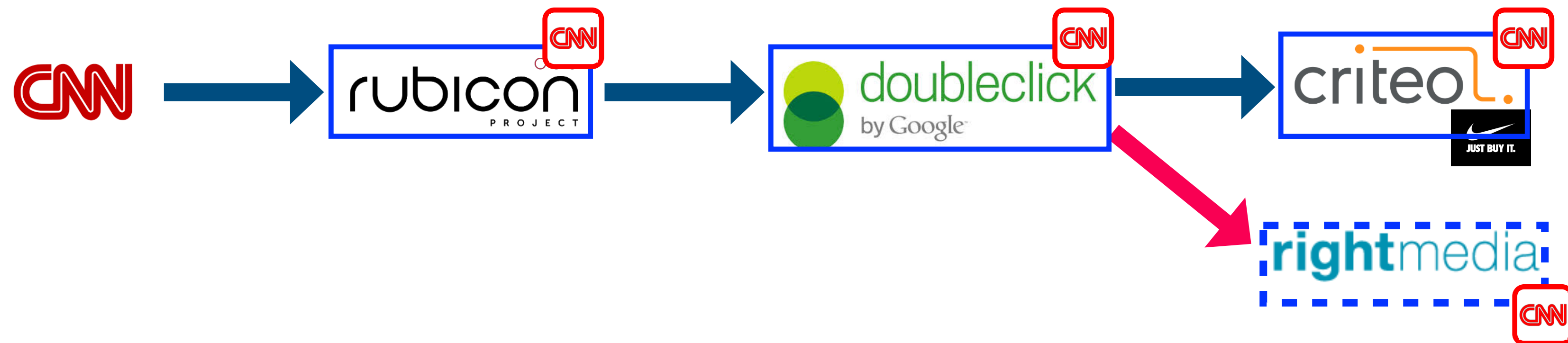
2. Impressions are forwarded to A&A domains via:

A. **Direct Propagation:** 

- Present on publisher or won RTB auction. **Observable** (goes through the browser)


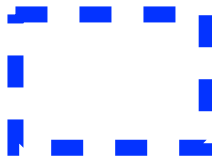
B. **Indirect Propagation:** 

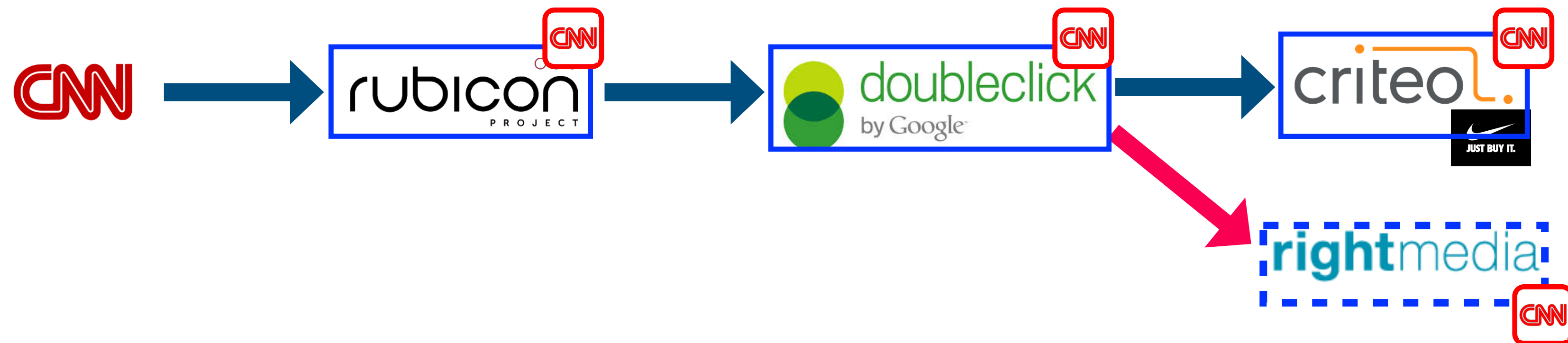
- A&A domains learn impressions through RTB participation. **Non-observable**



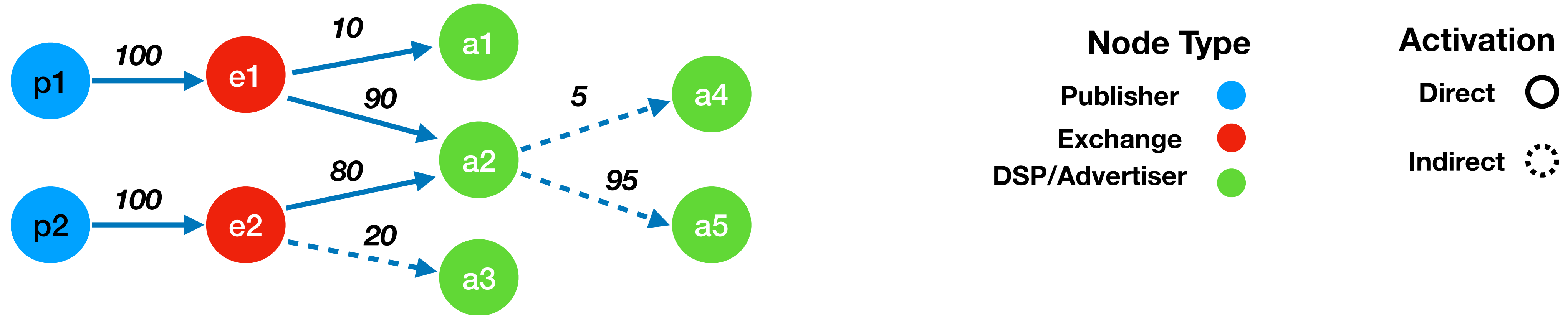
# Our Simulation Setup

We simulate browsing traces for 200 users using method from [1].

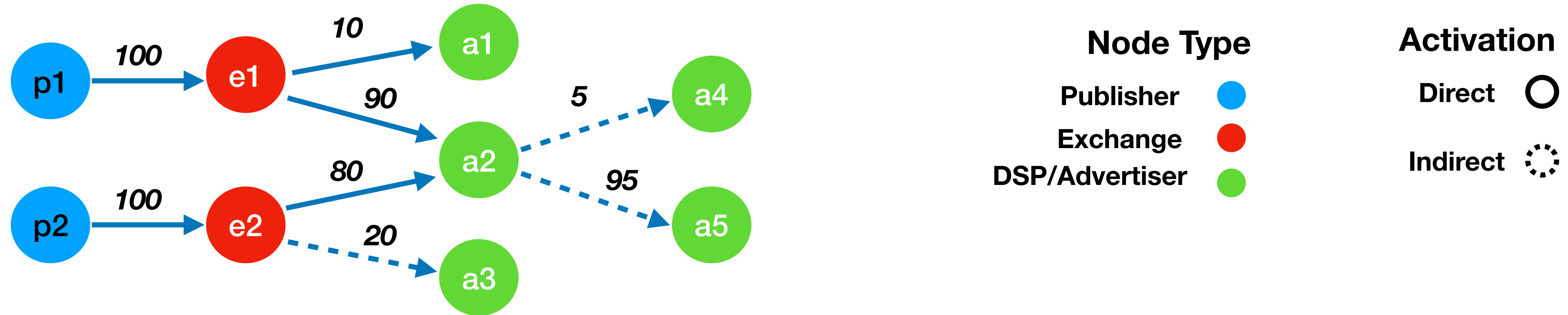
1. User generates an impression on N selected publishers.
2. Impressions are forwarded to A&A domains via:
  - A. **Direct Propagation:** 
    - Present on publisher or won RTB auction. **Observable** (goes through the browser)
  - B. **Indirect Propagation:** 
    - A&A domains learn impressions through RTB participation. **Non-observable**
3. RTB winner is decided based on probability (function of edge weights).



# Simulation Example (RTB Constrained)

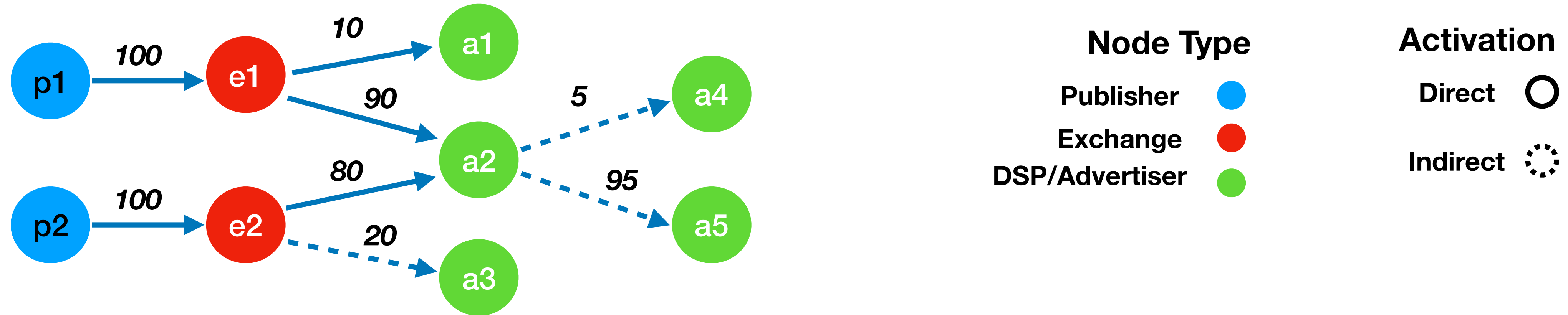


# Simulation Example (RTB Constrained)



We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)

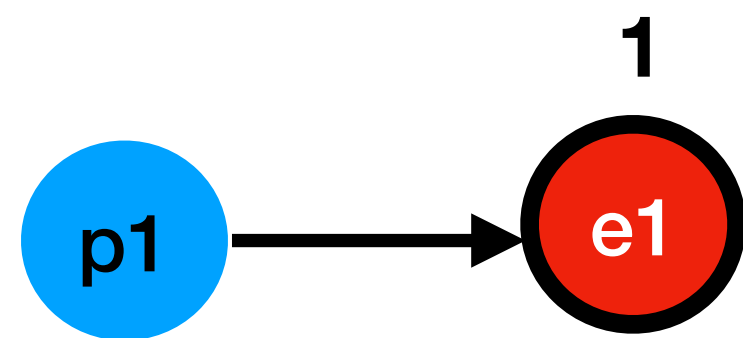
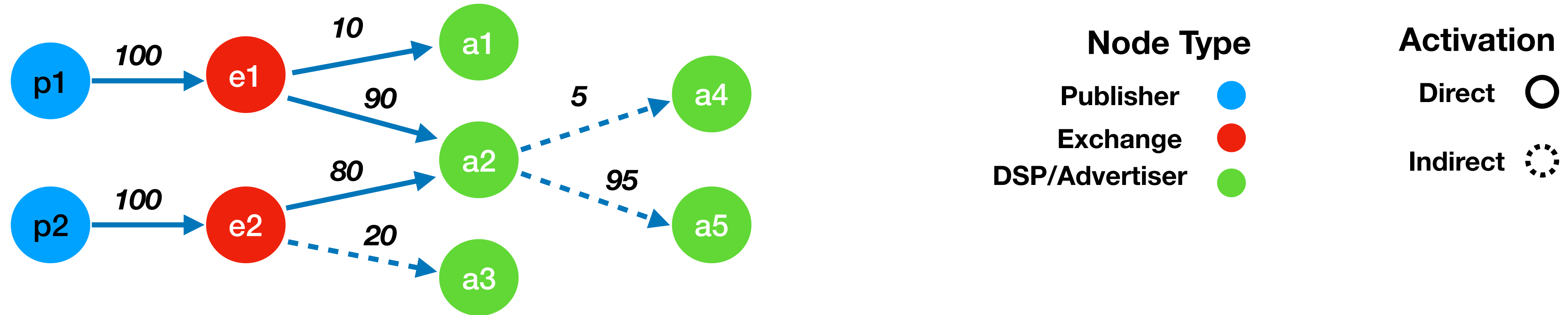


p1

We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB



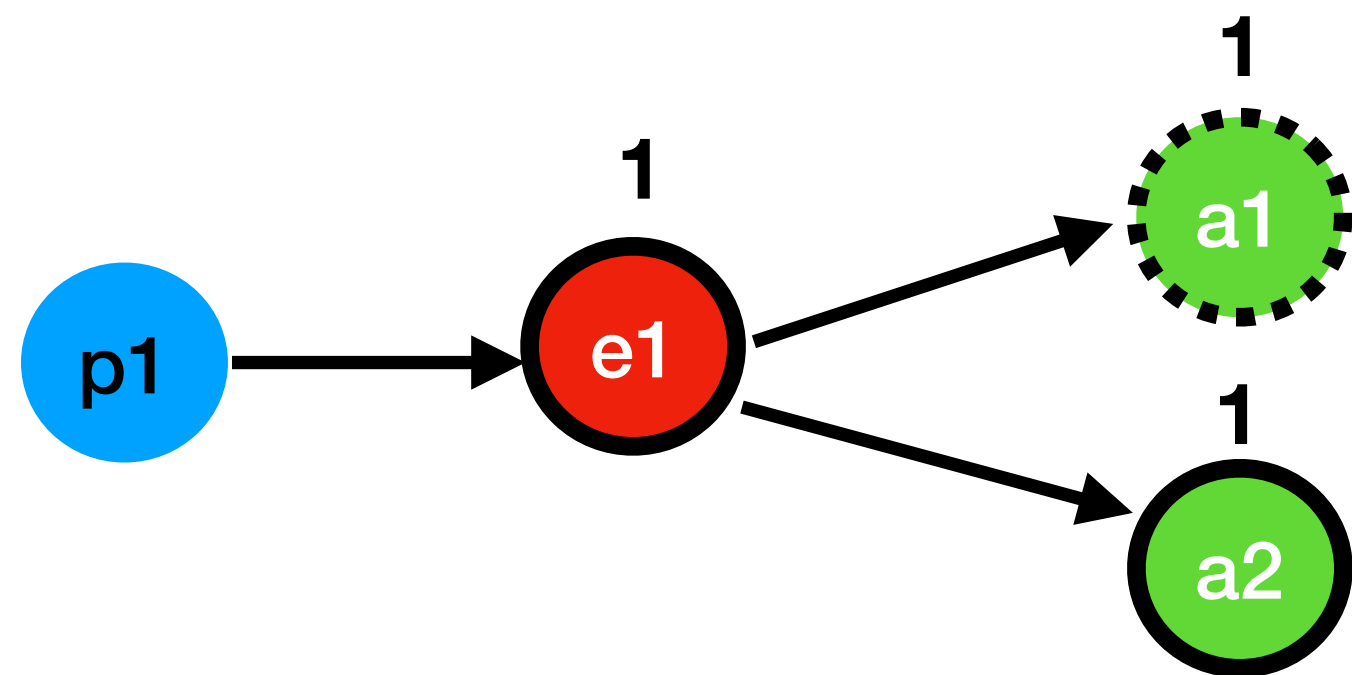
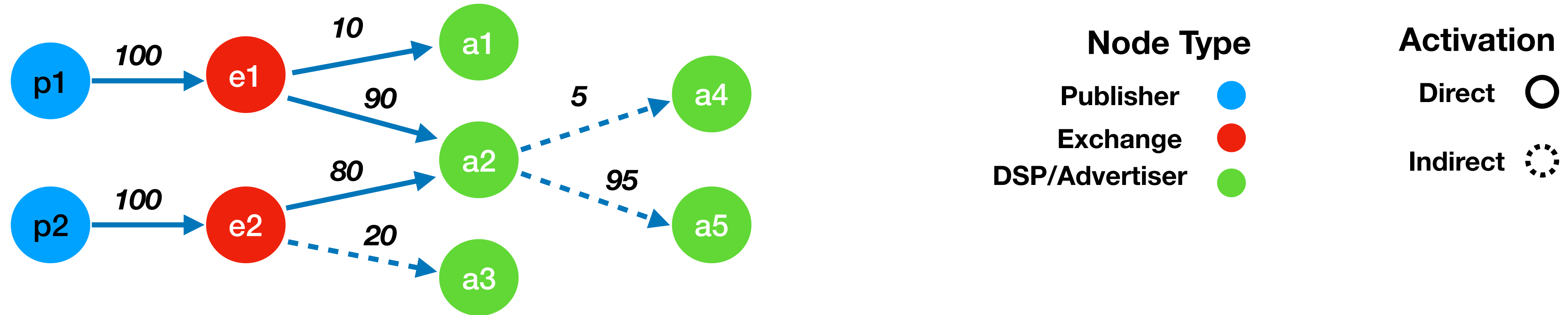
# Simulation Example (RTB Constrained)



We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

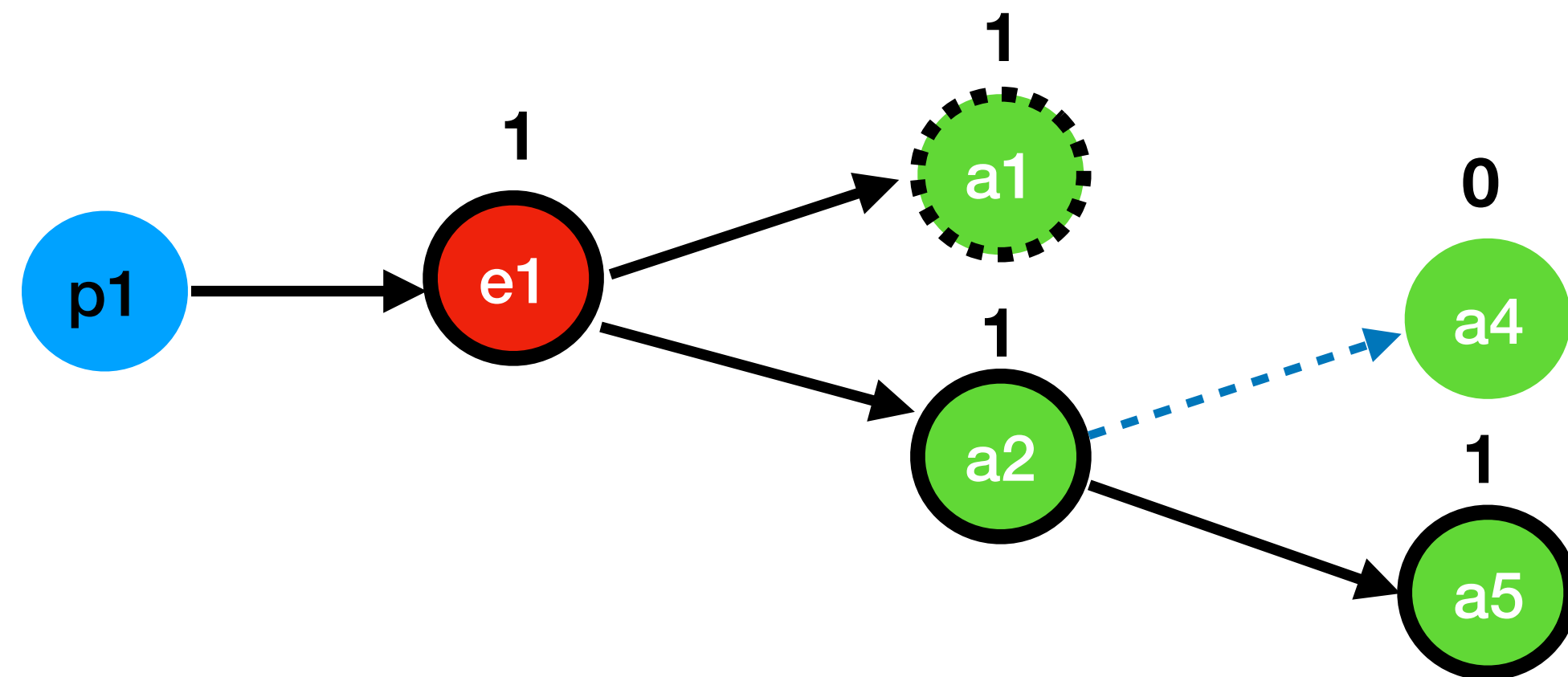
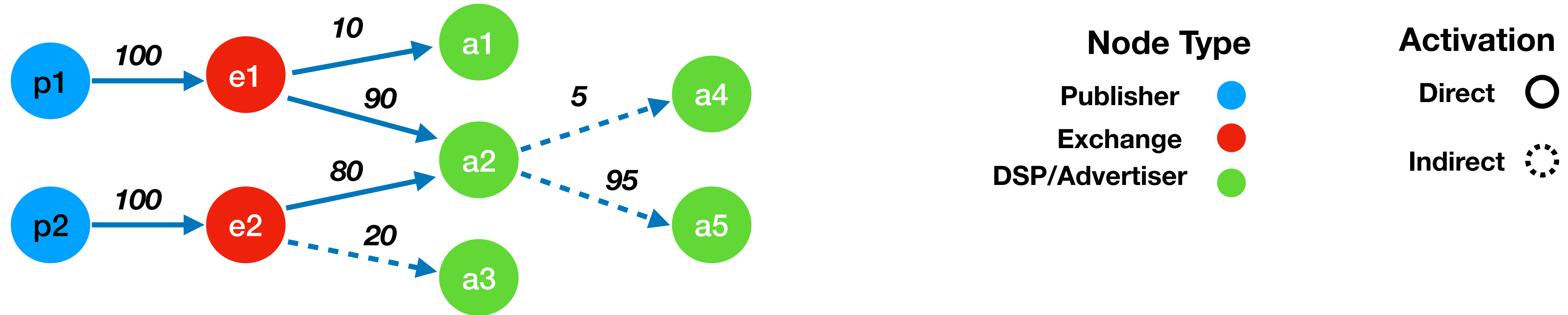


# Simulation Example (RTB Constrained)



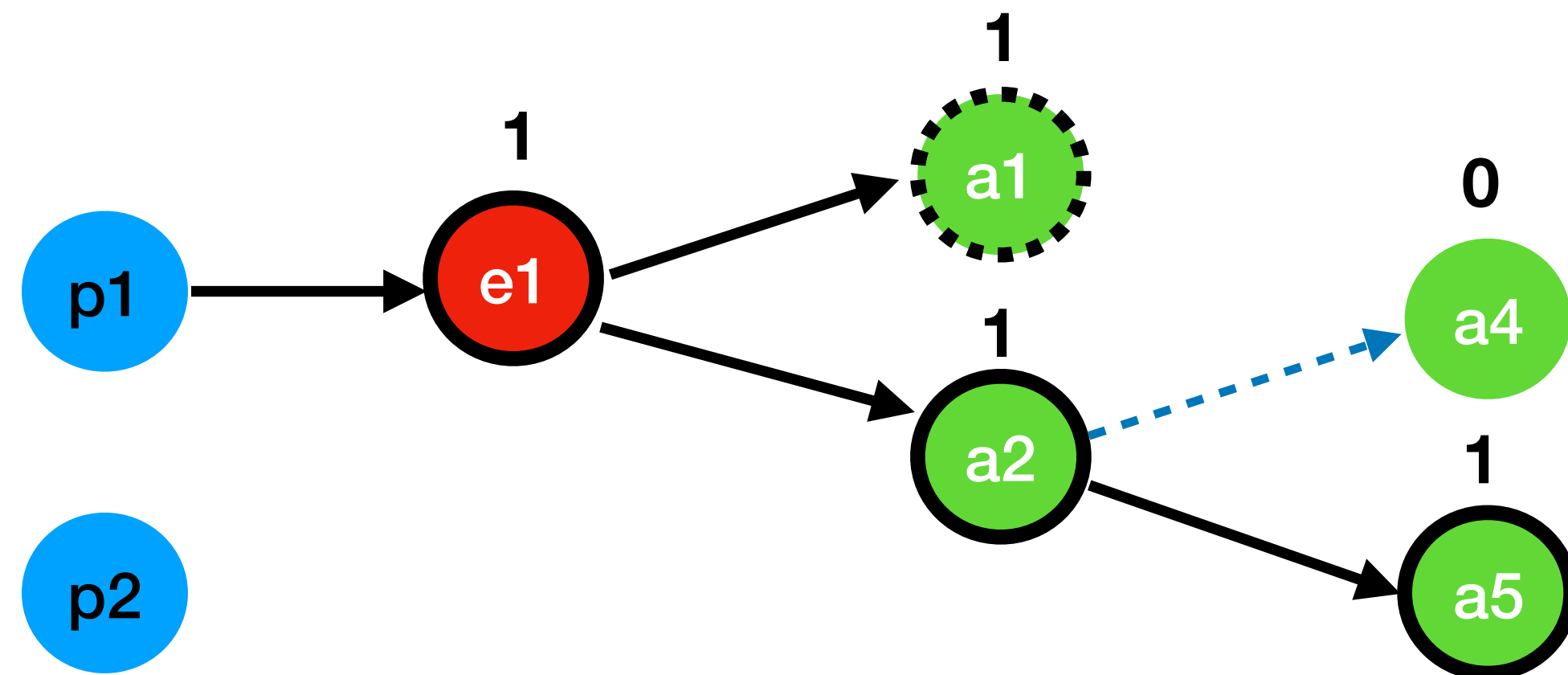
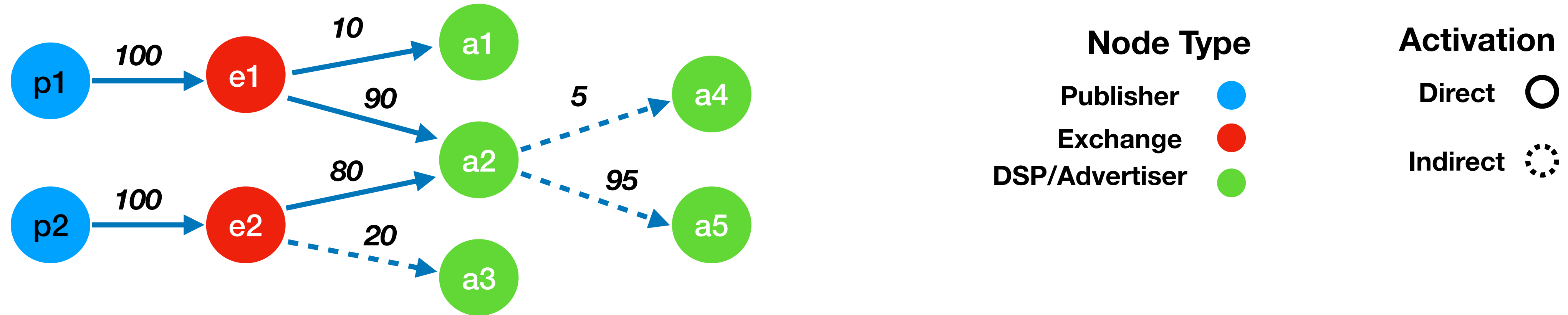
We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)



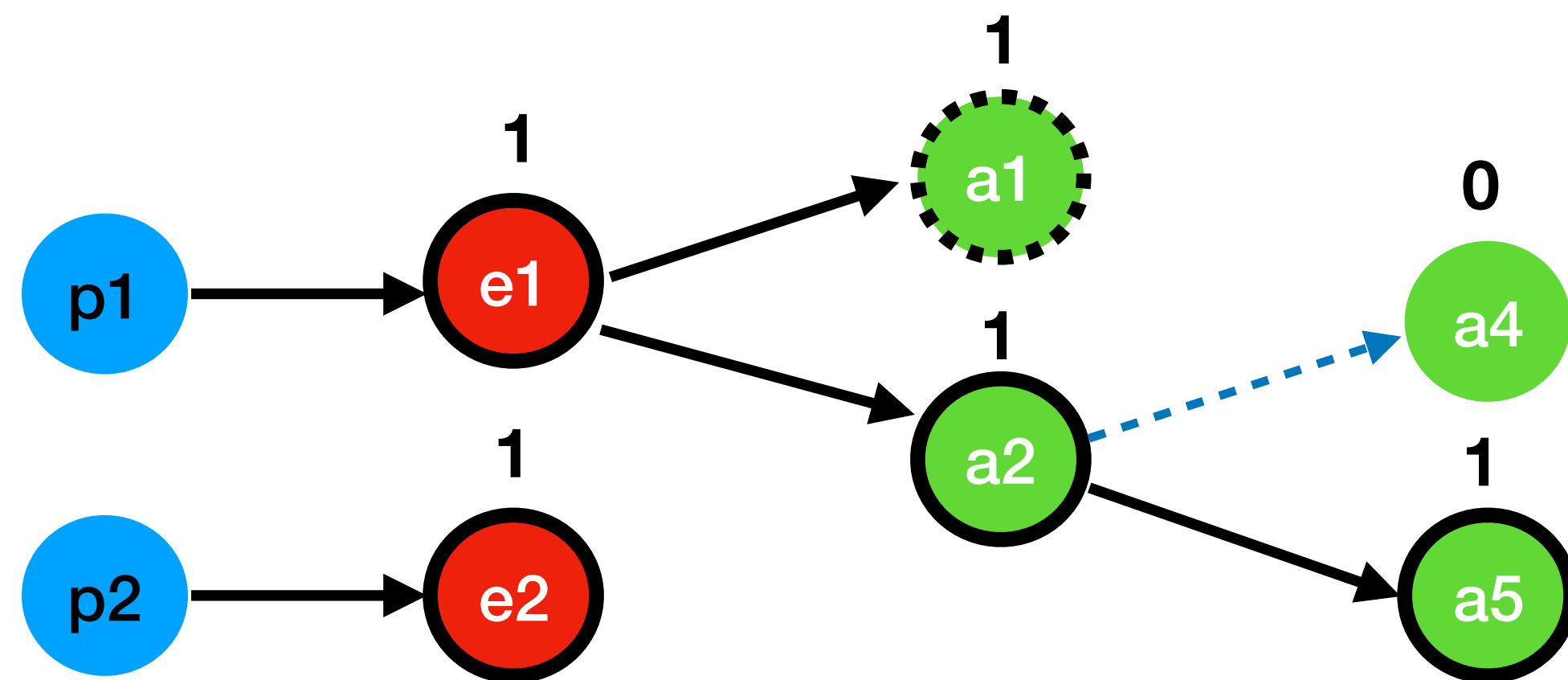
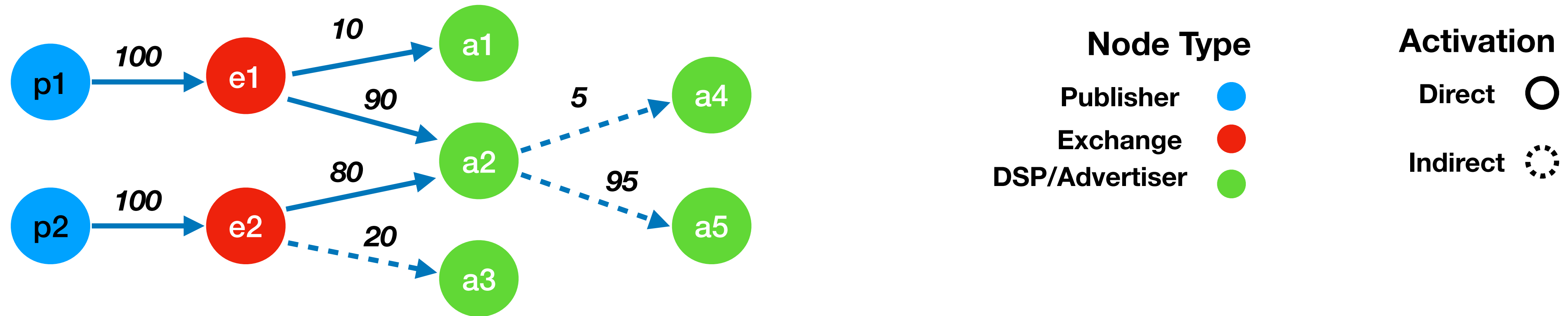
We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)



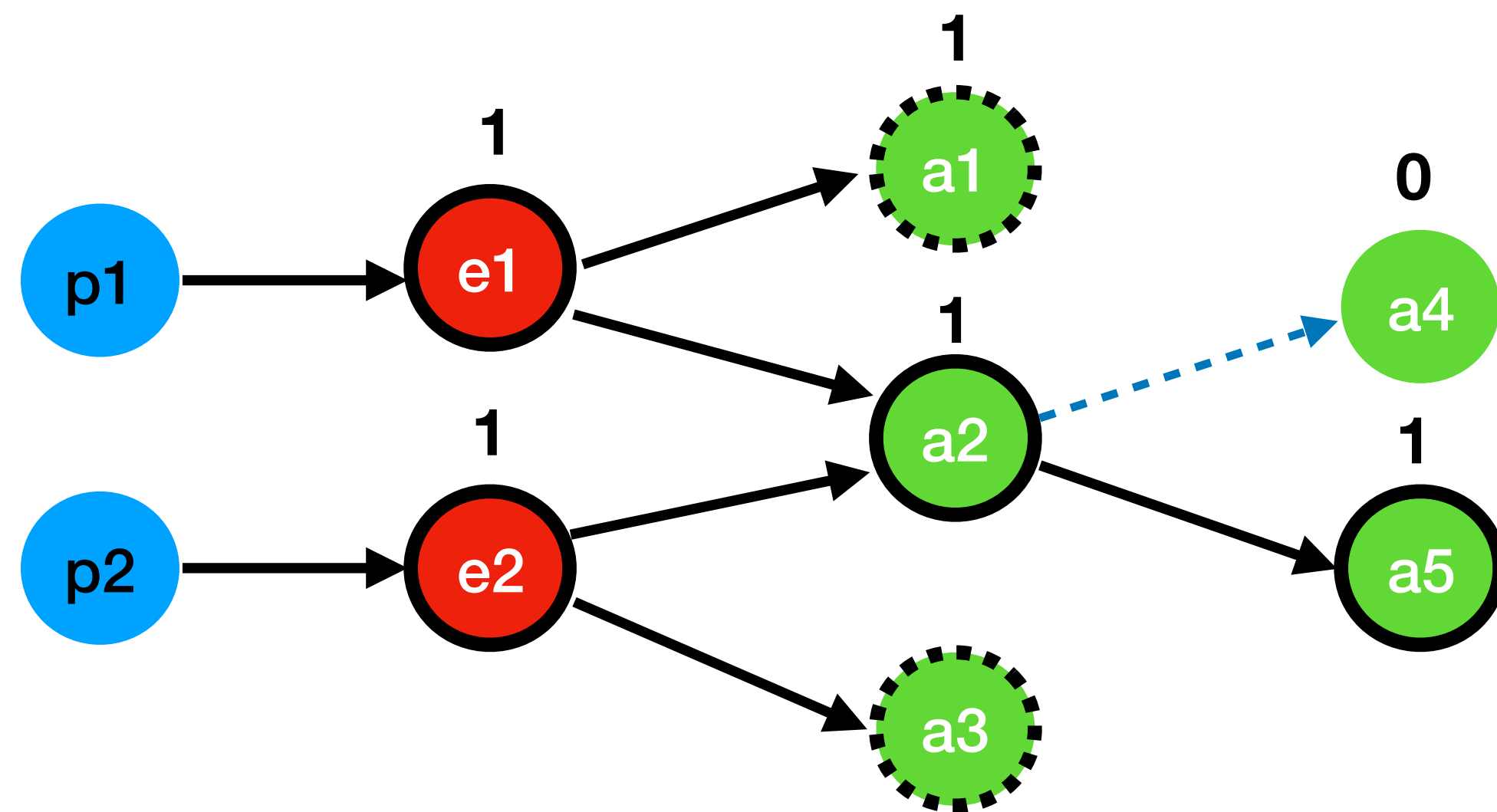
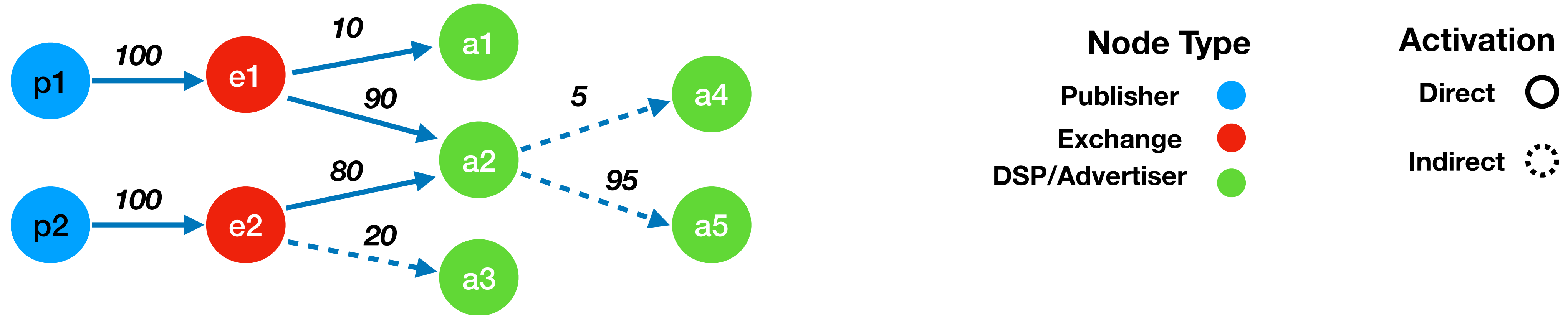
We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)



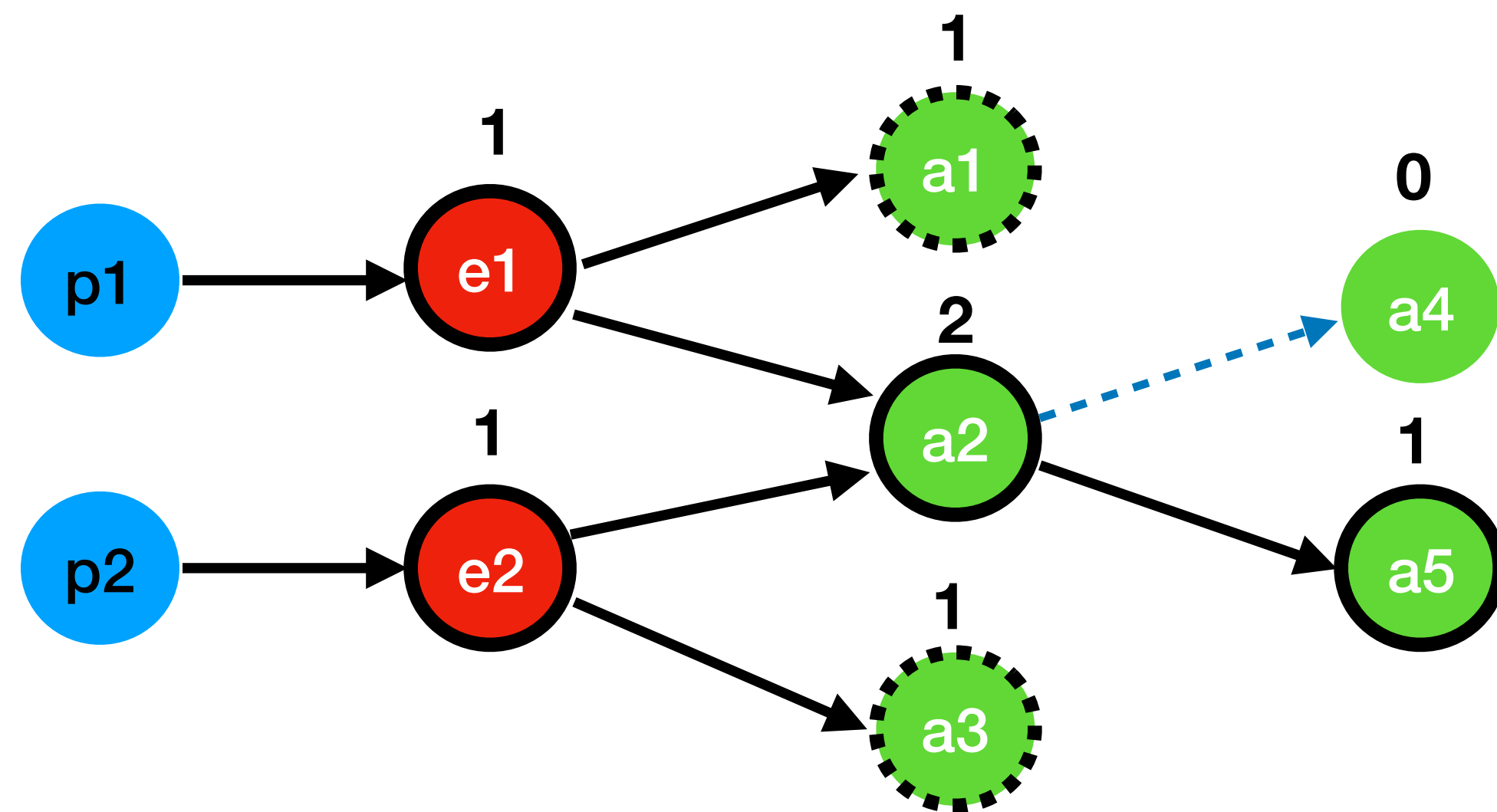
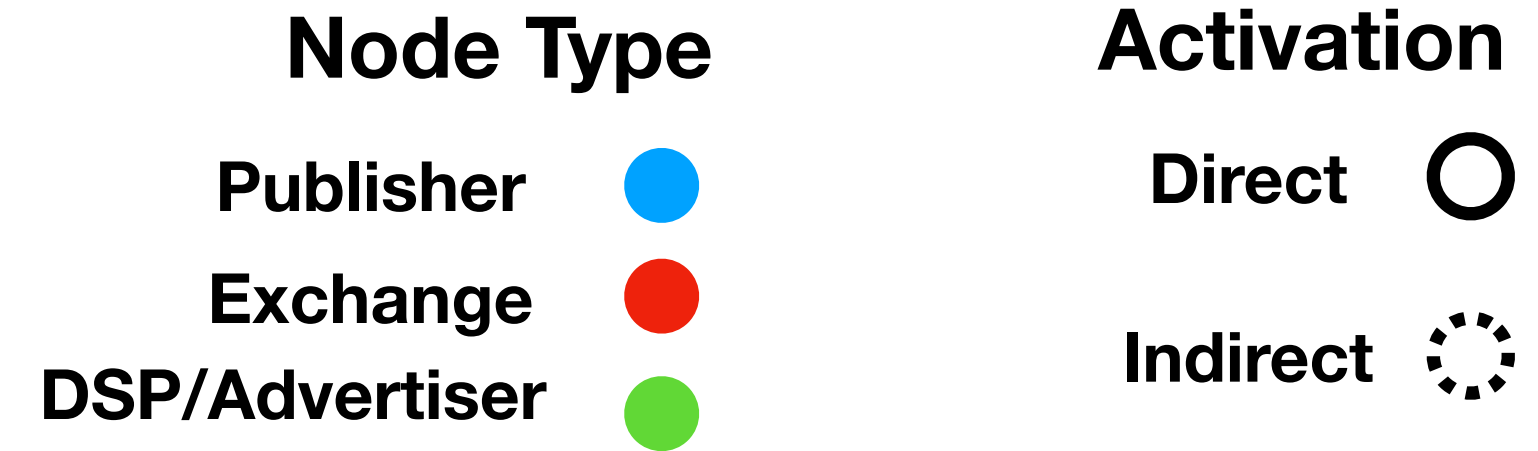
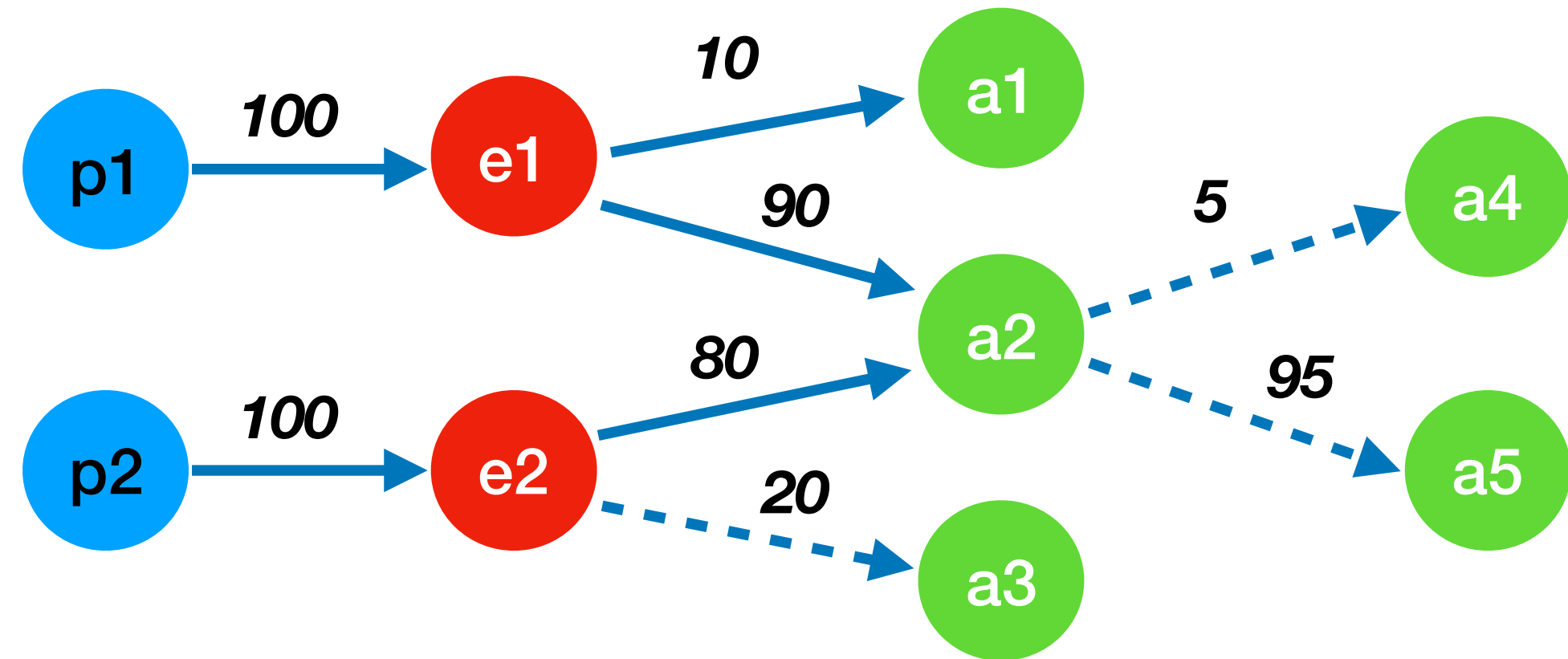
We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)



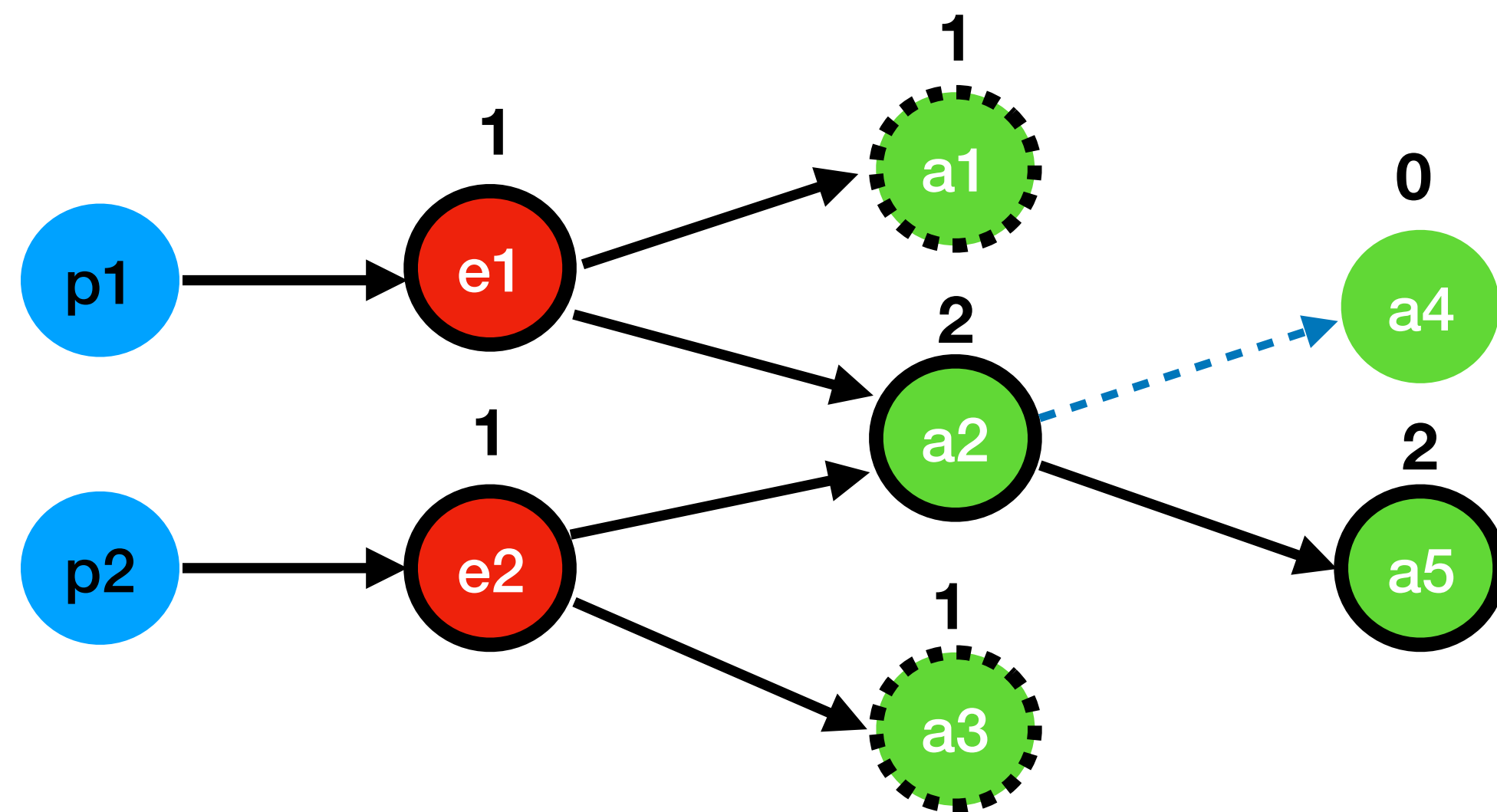
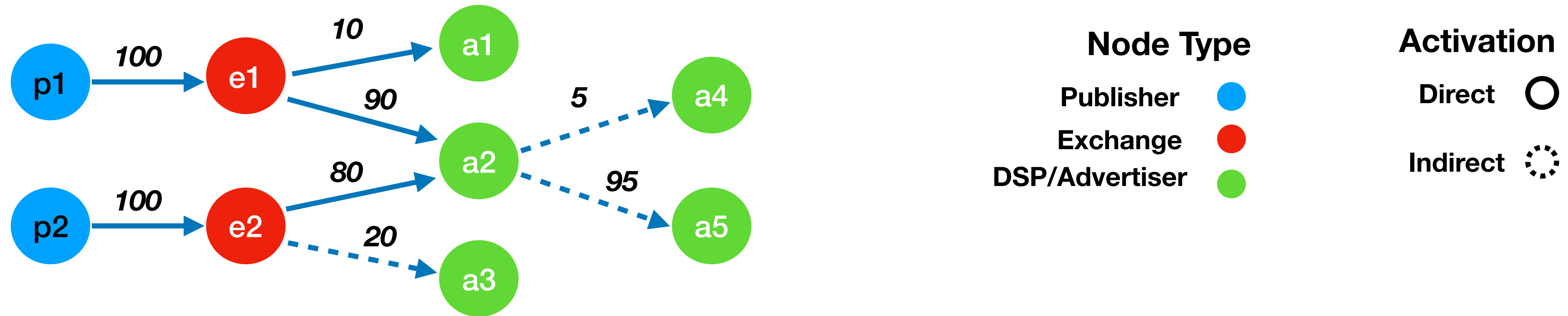
We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)



We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Simulation Example (RTB Constrained)



We manually select **37 exchanges** which are allowed to forward **indirect** impressions to solicit bids during RTB

# Impressions Observed

We have 3 simulation models:

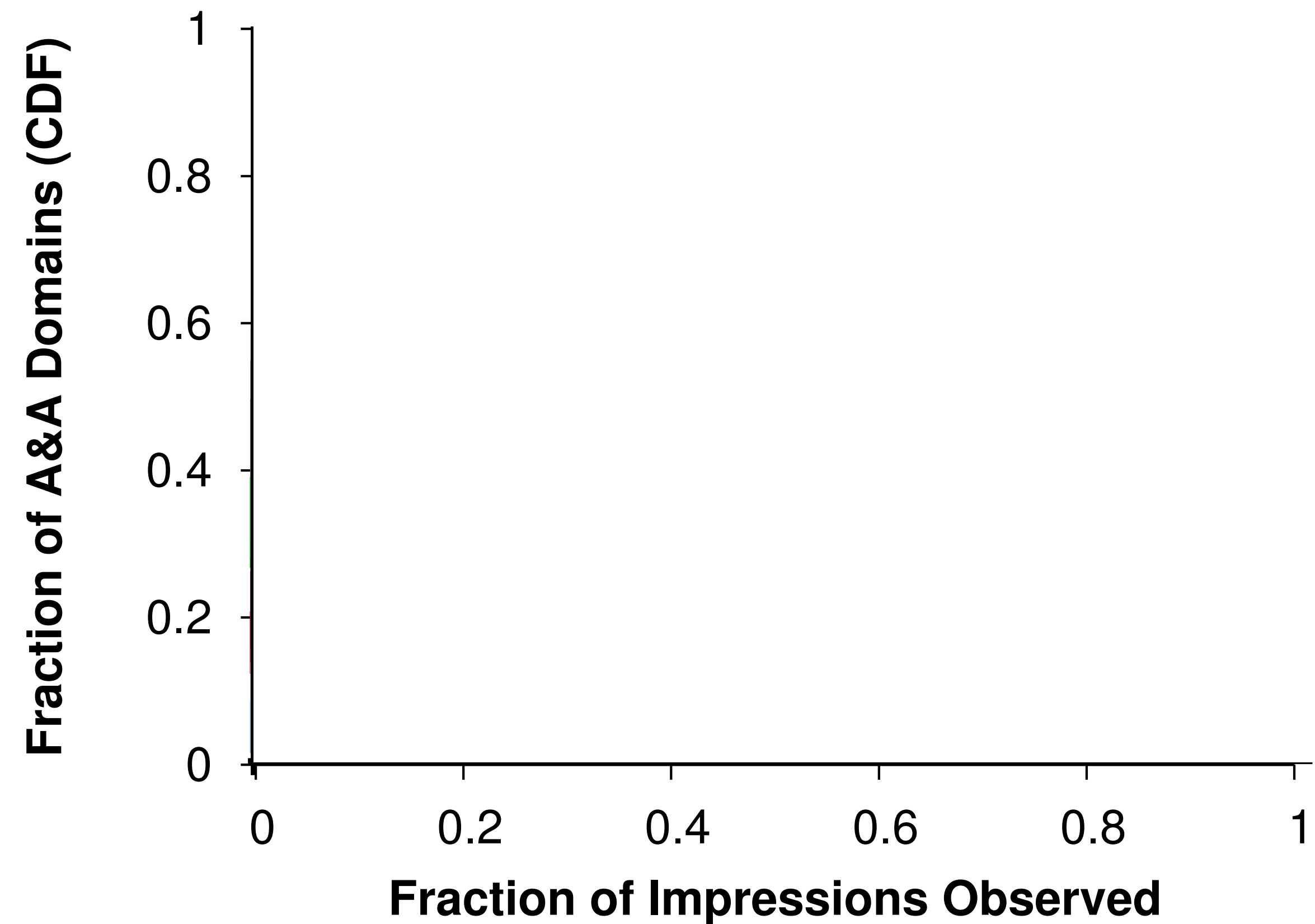
1. **RTB-Relaxed:** Upper-bound
2. **Cookie-Matching:** Lower-bound
3. **RTB-Constrained:** Realistic Scenario



# Impressions Observed

We have 3 simulation models:

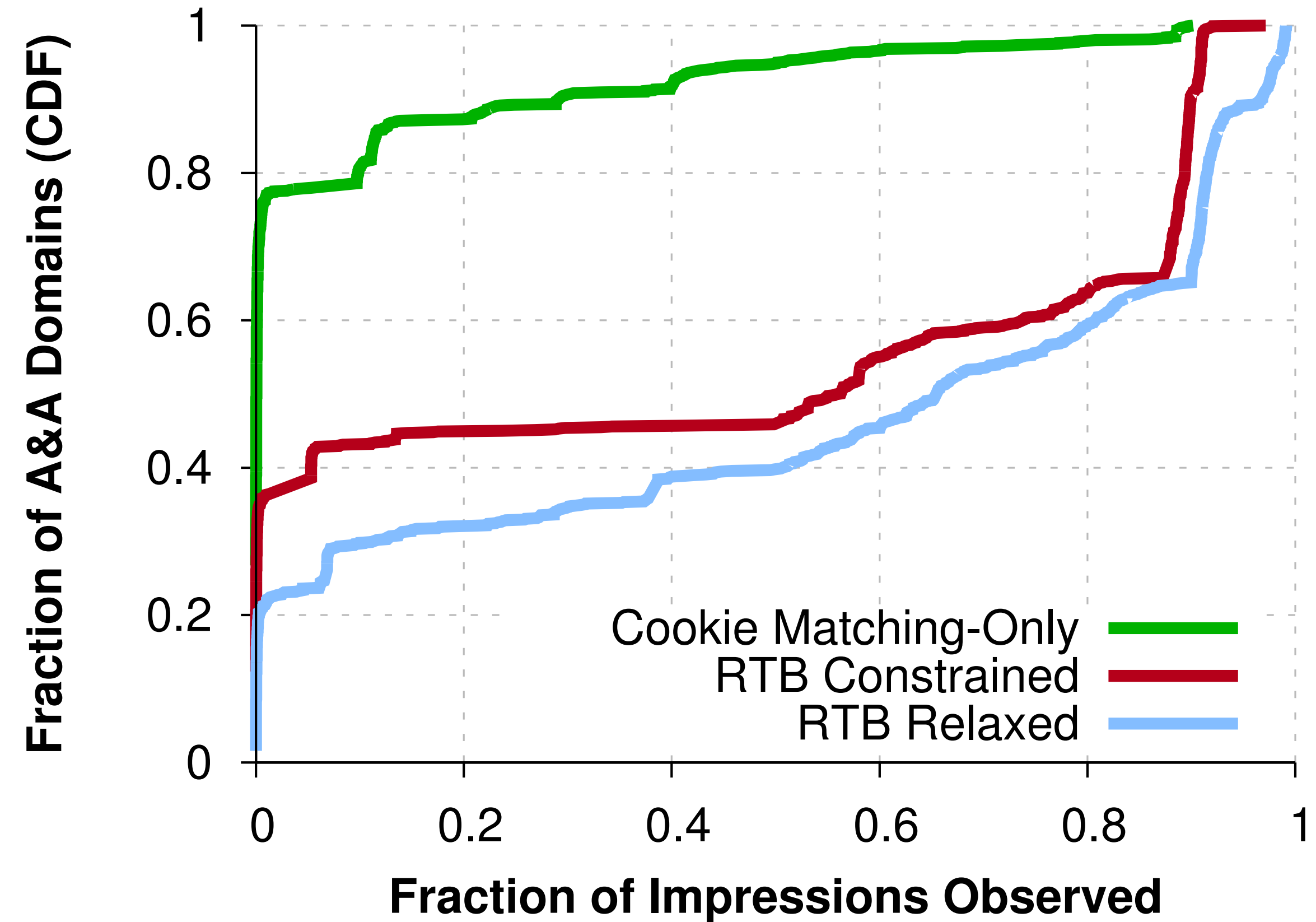
1. **RTB-Relaxed:** Upper-bound
2. **Cookie-Matching:** Lower-bound
3. **RTB-Constrained:** Realistic Scenario



# Impressions Observed

We have 3 simulation models:

1. **RTB-Relaxed:** Upper-bound
2. **Cookie-Matching:** Lower-bound
3. **RTB-Constrained:** Realistic Scenario



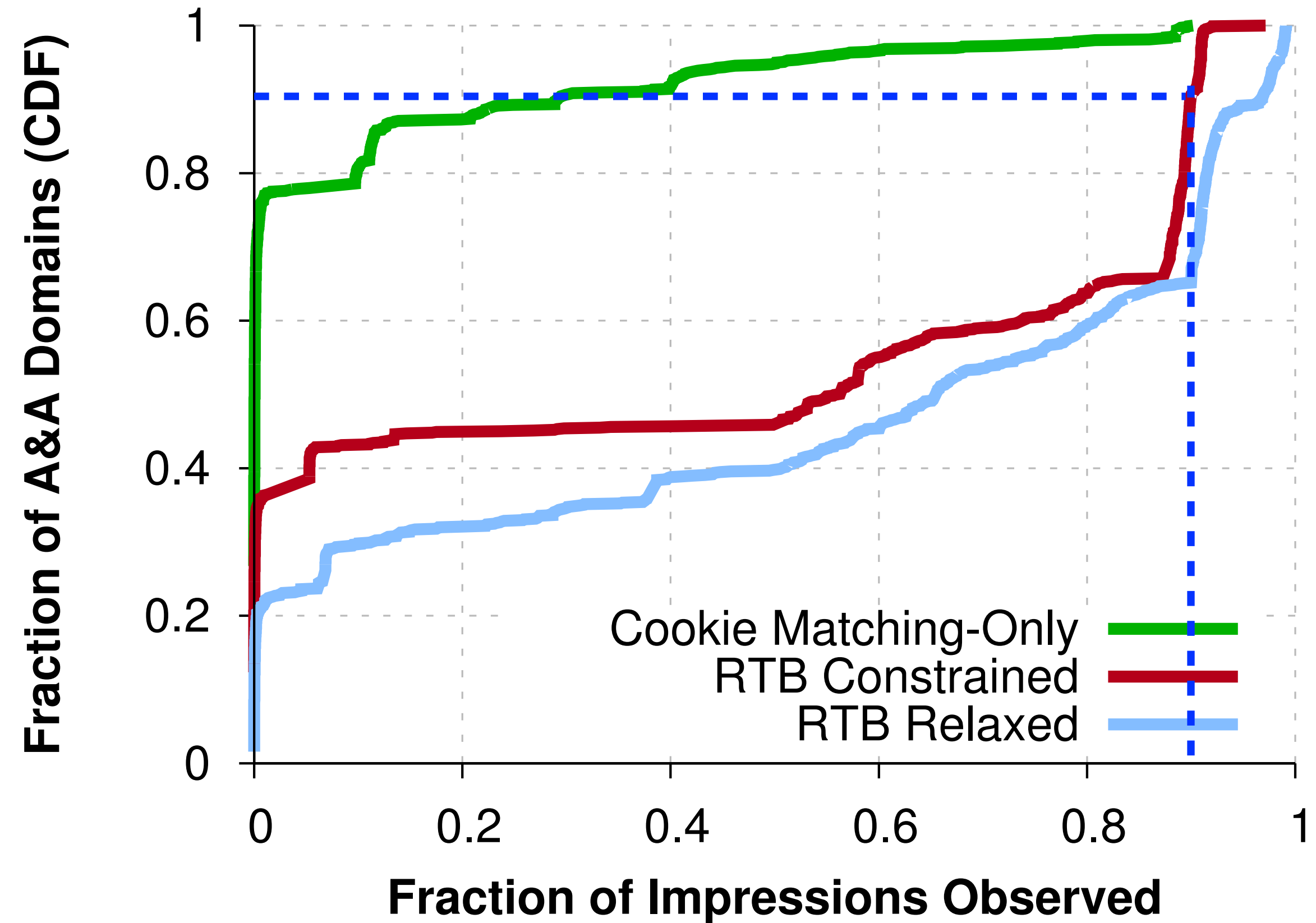
# Impressions Observed

We have 3 simulation models:

1. **RTB-Relaxed:** Upper-bound
2. **Cookie-Matching:** Lower-bound
3. **RTB-Constrained:** Realistic Scenario

## Take Away

1. RTB-Constrained is very close to RTB-Relaxed
2. 10% A&A see more than 90% of impressions in RTB-Constrained

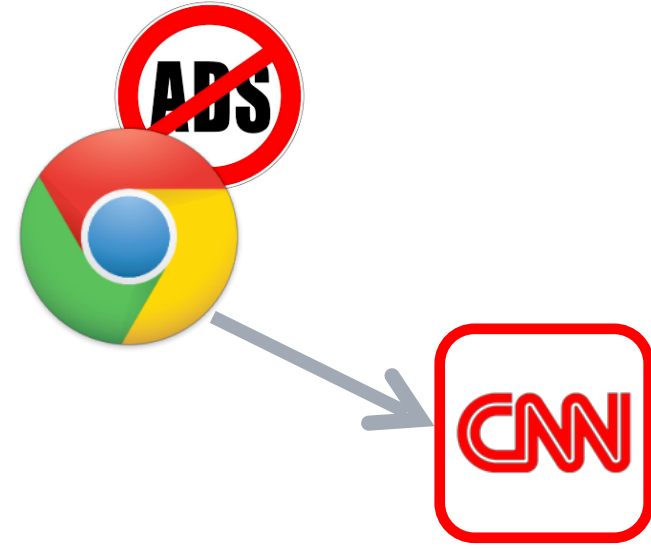


# Effect of Blocking Extensions



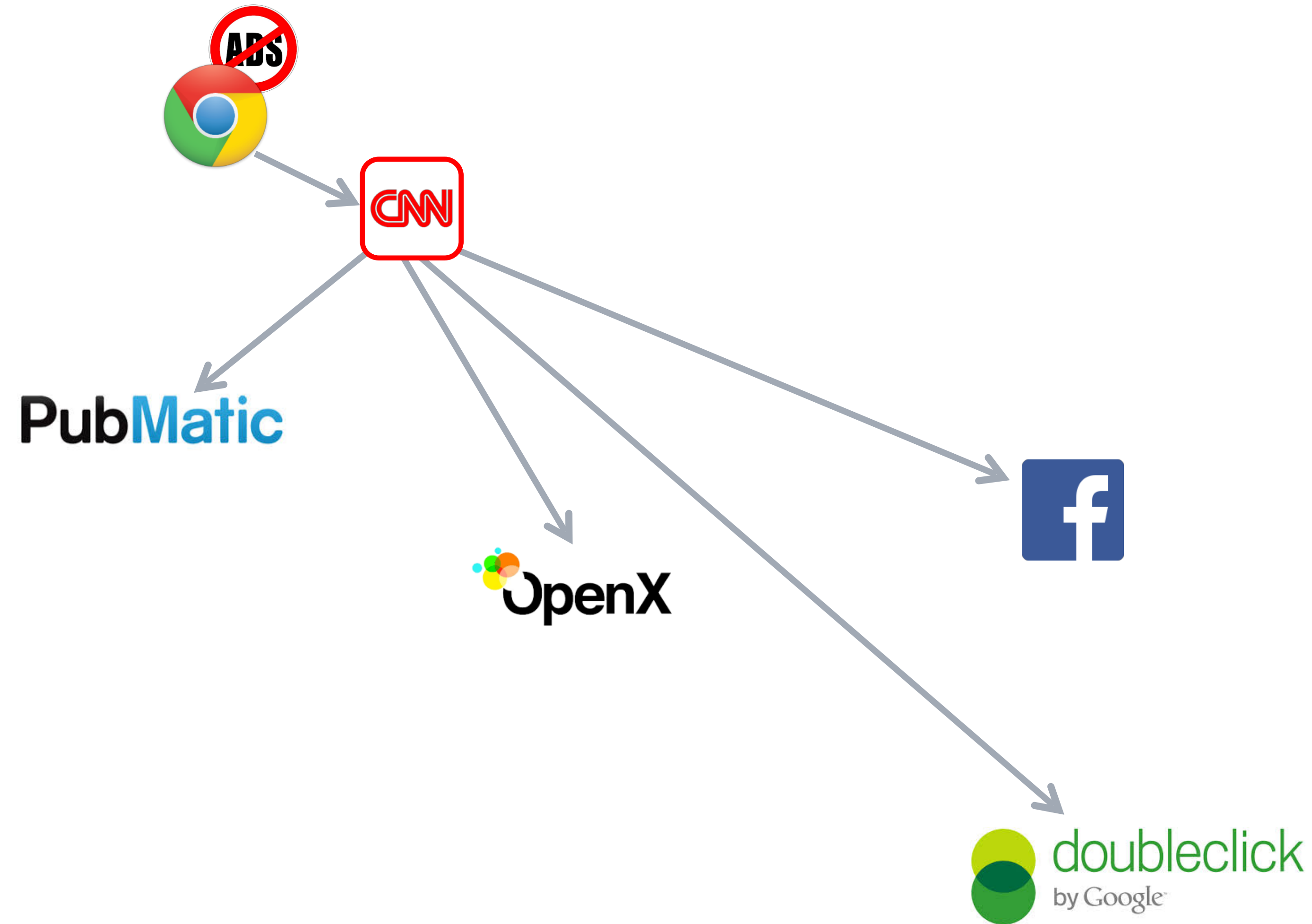
DoubleClick  
OpenX  
PubMatic

# Effect of Blocking Extensions



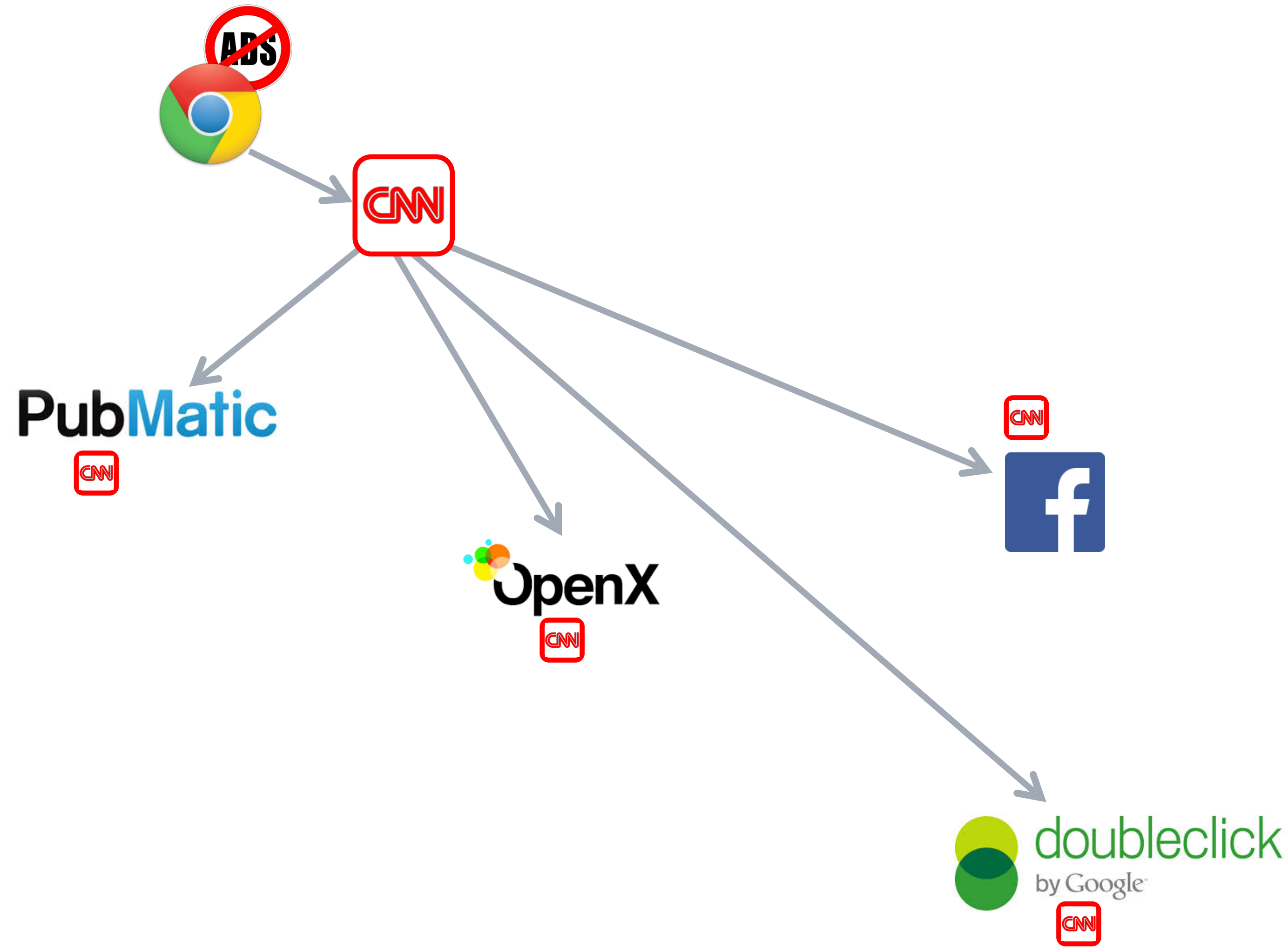
DoubleClick  
OpenX  
PubMatic

# Effect of Blocking Extensions



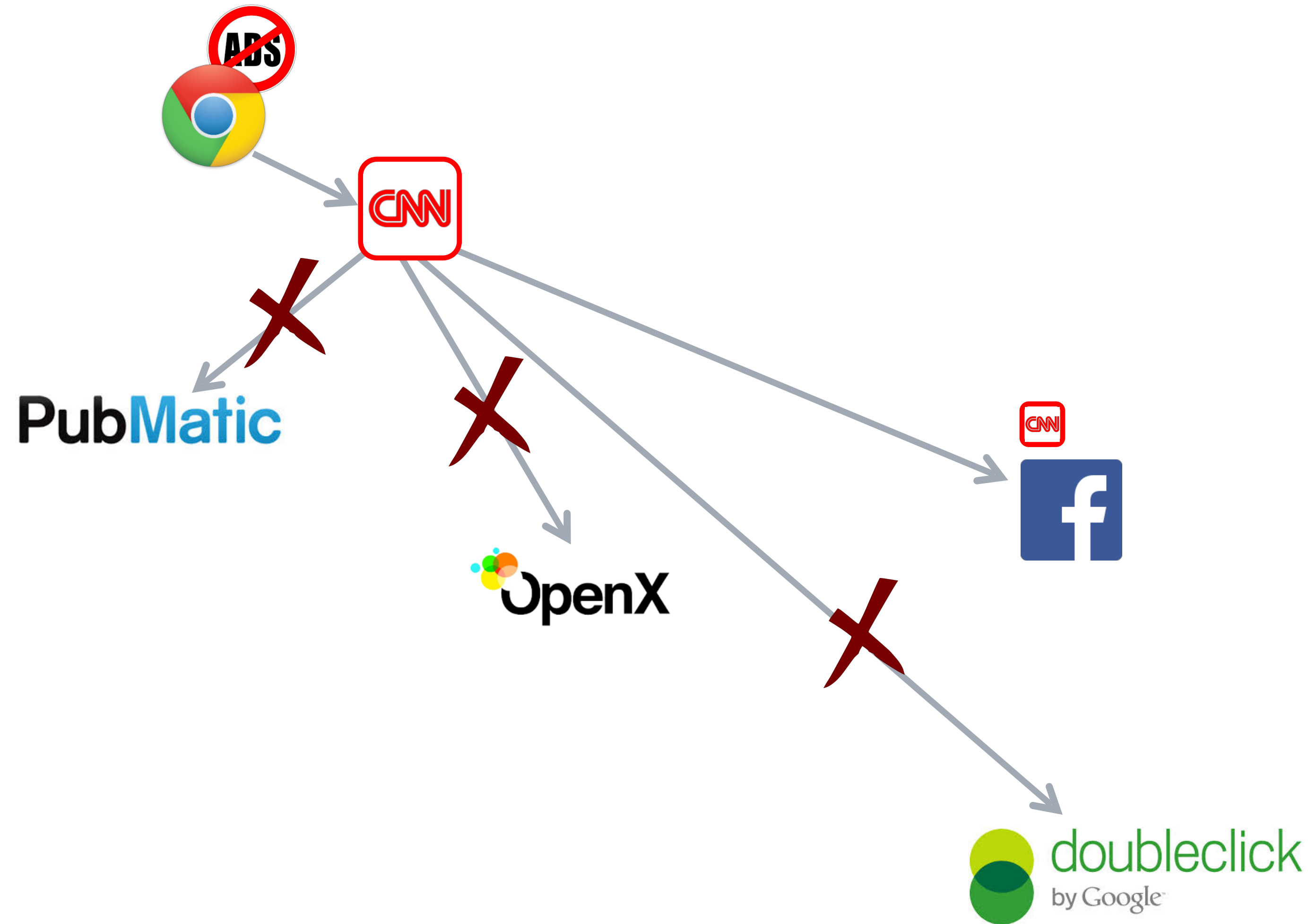
DoubleClick  
OpenX  
PubMatic

# Effect of Blocking Extensions



~~ADS~~  
DoubleClick  
OpenX  
PubMatic

# Effect of Blocking Extensions

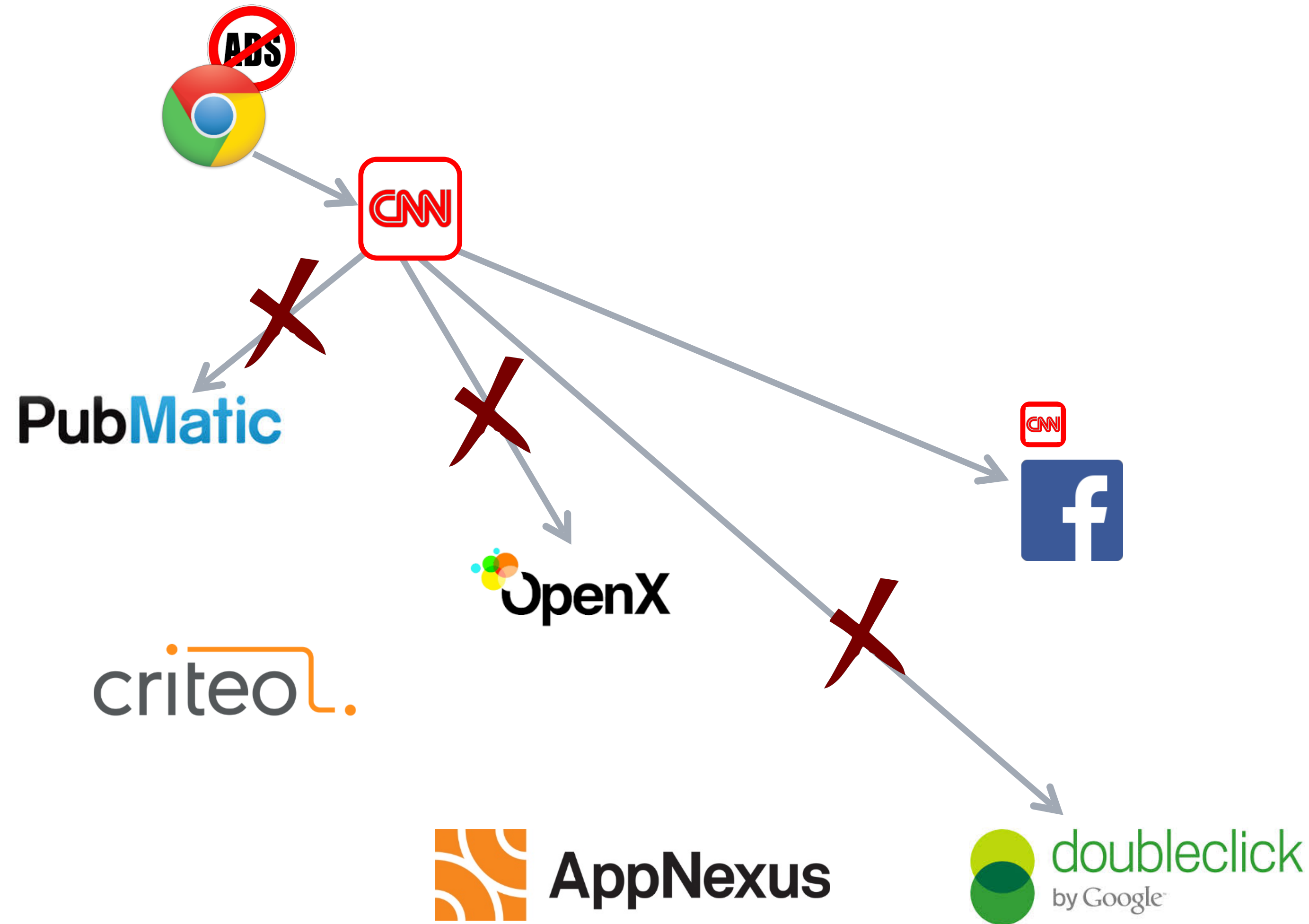


**ADS**

- DoubleClick
- OpenX
- PubMatic



# Effect of Blocking Extensions

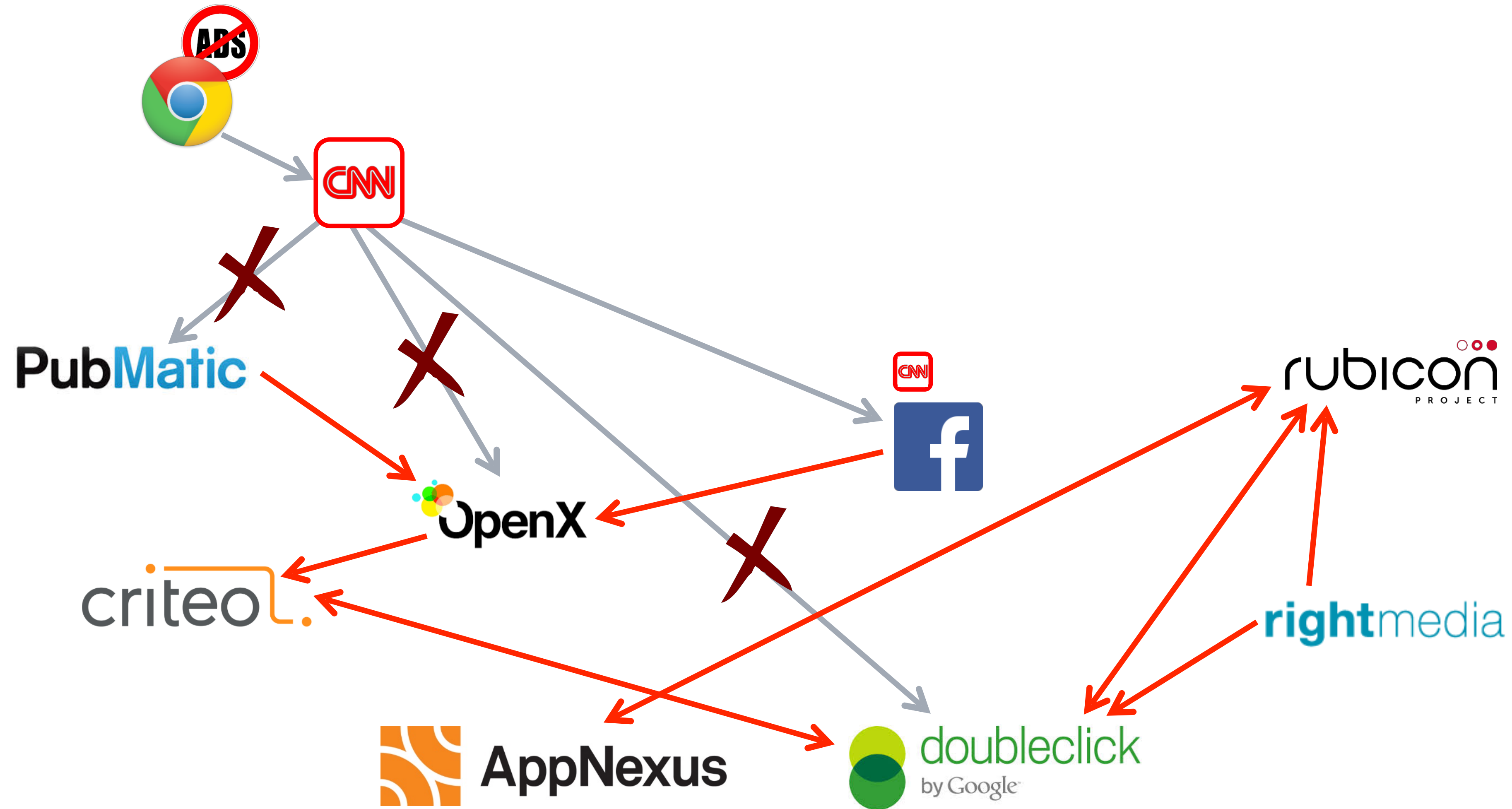


  
DoubleClick  
OpenX  
PubMatic

rubicon  
PROJECT

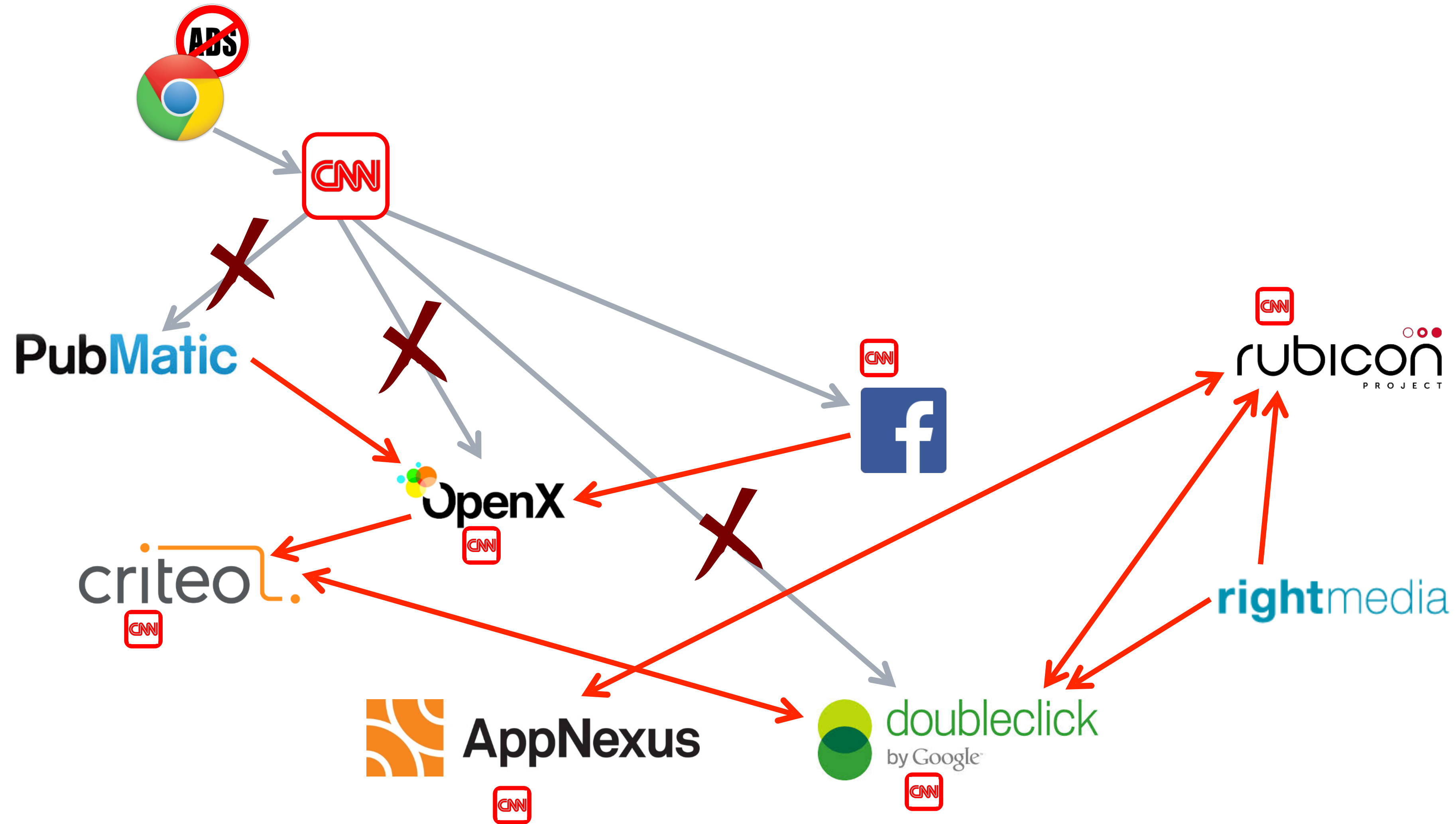
rightmedia

# Effect of Blocking Extensions



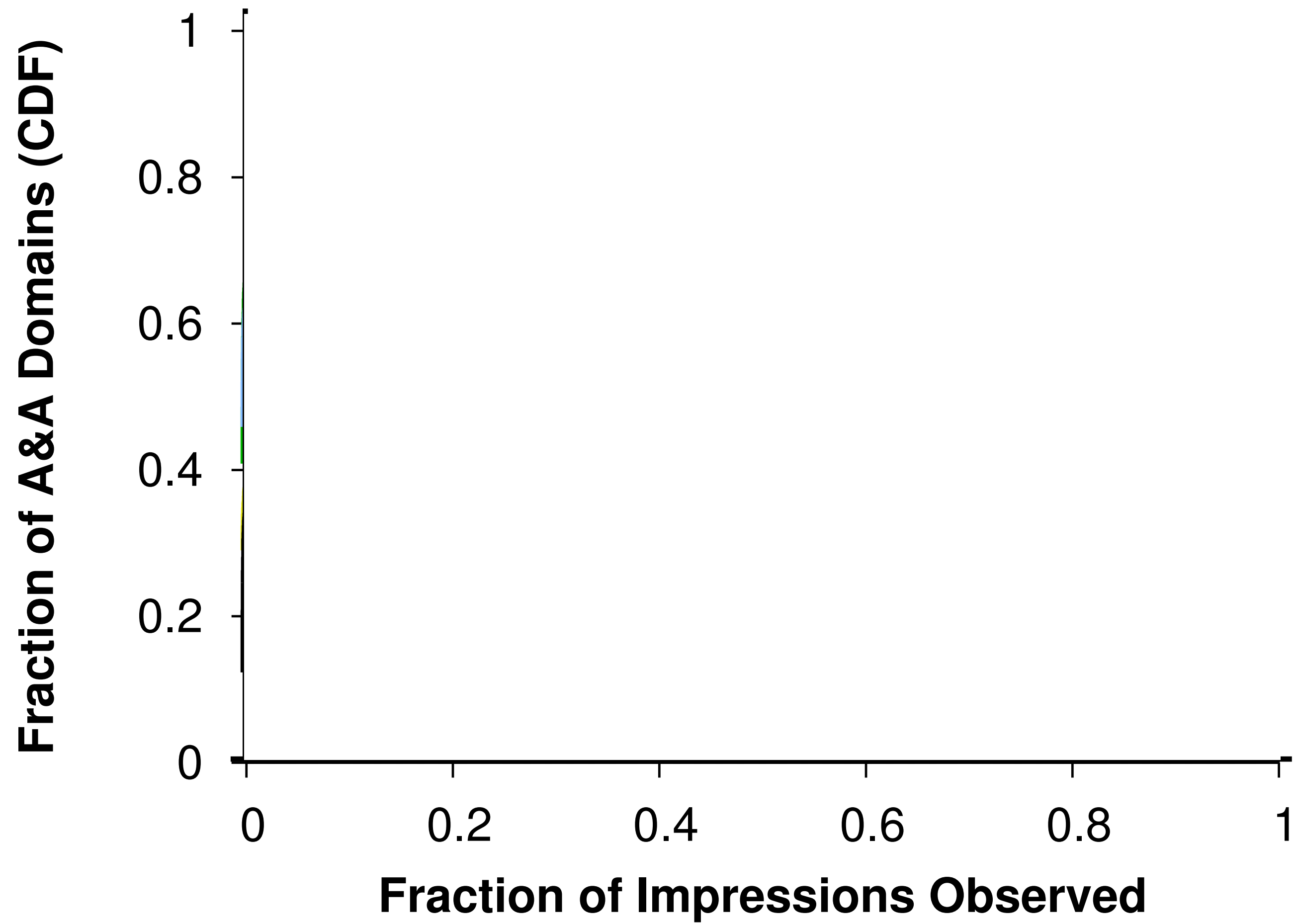
DoubleClick  
OpenX  
PubMatic

# Effect of Blocking Extensions

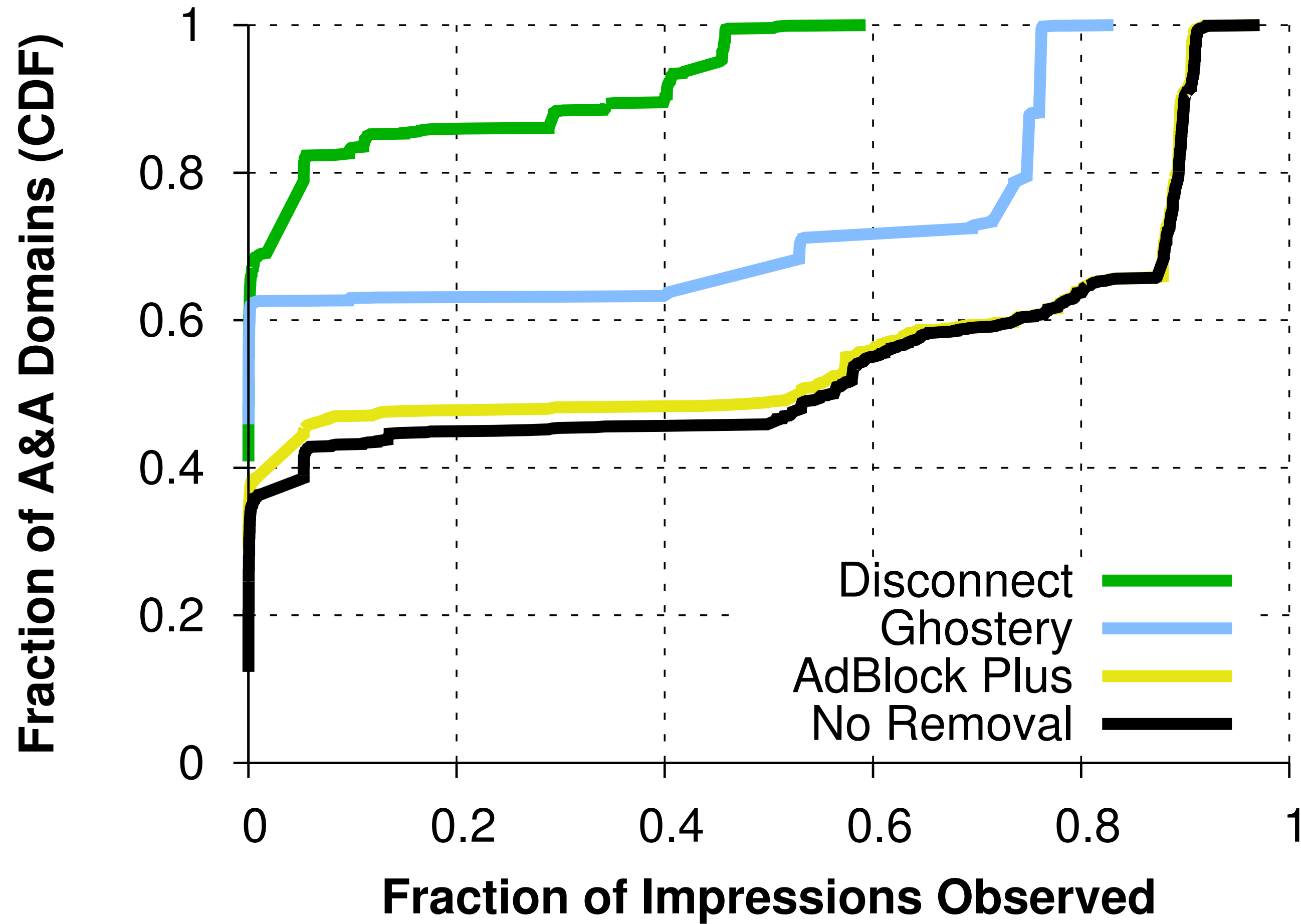


DoubleClick  
OpenX  
PubMatic

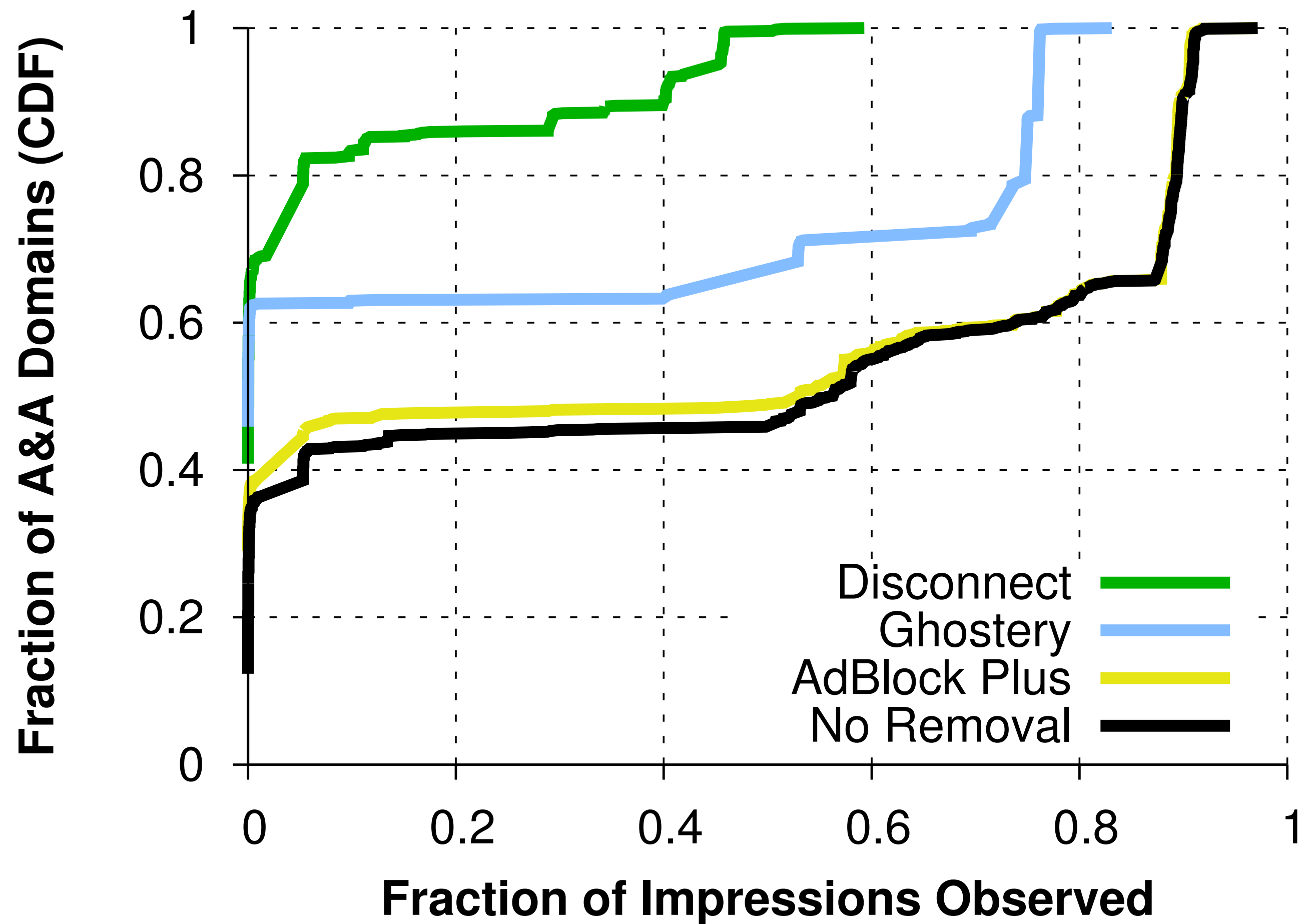
# Impressions With Blocking



# Impressions With Blocking



# Impressions With Blocking

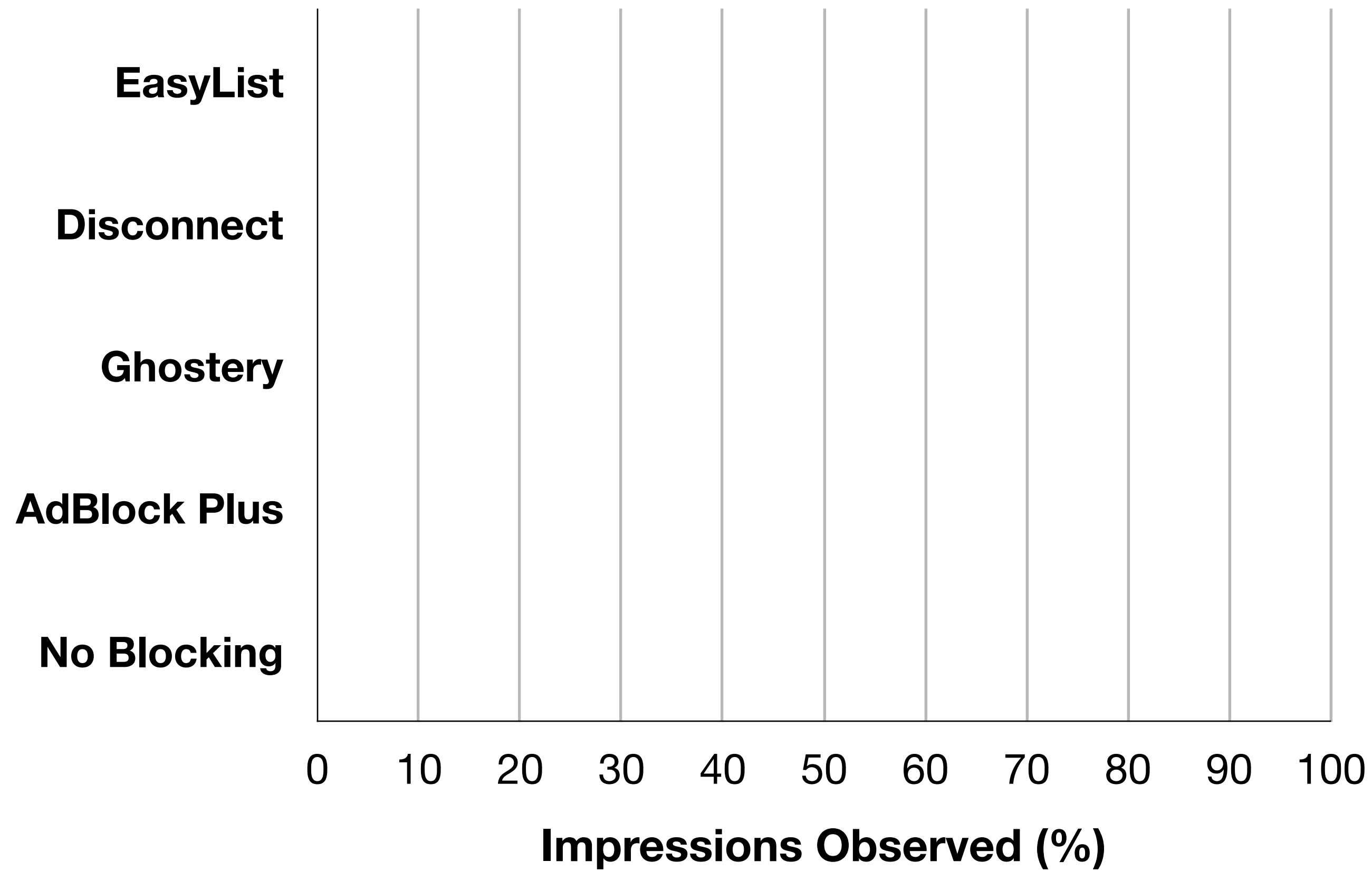


## Take Away

- Disconnect list is most effective.
- ABP is not effective at all due to Acceptable Ads program.
- Due to RTB, impressions are leaked to A&A domains even with blocking extensions.

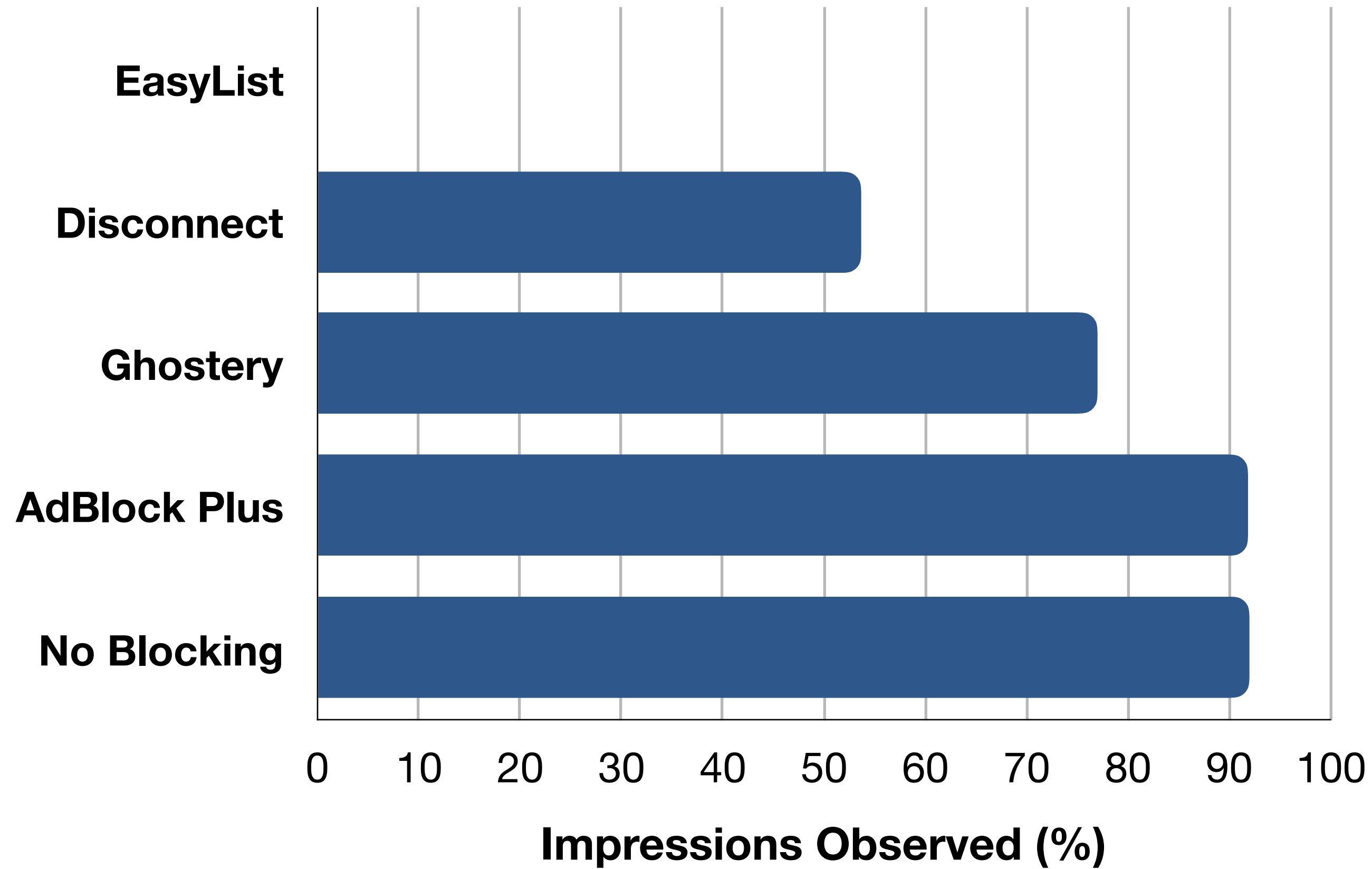
# Top 10 Domains

# Top 10 Domains



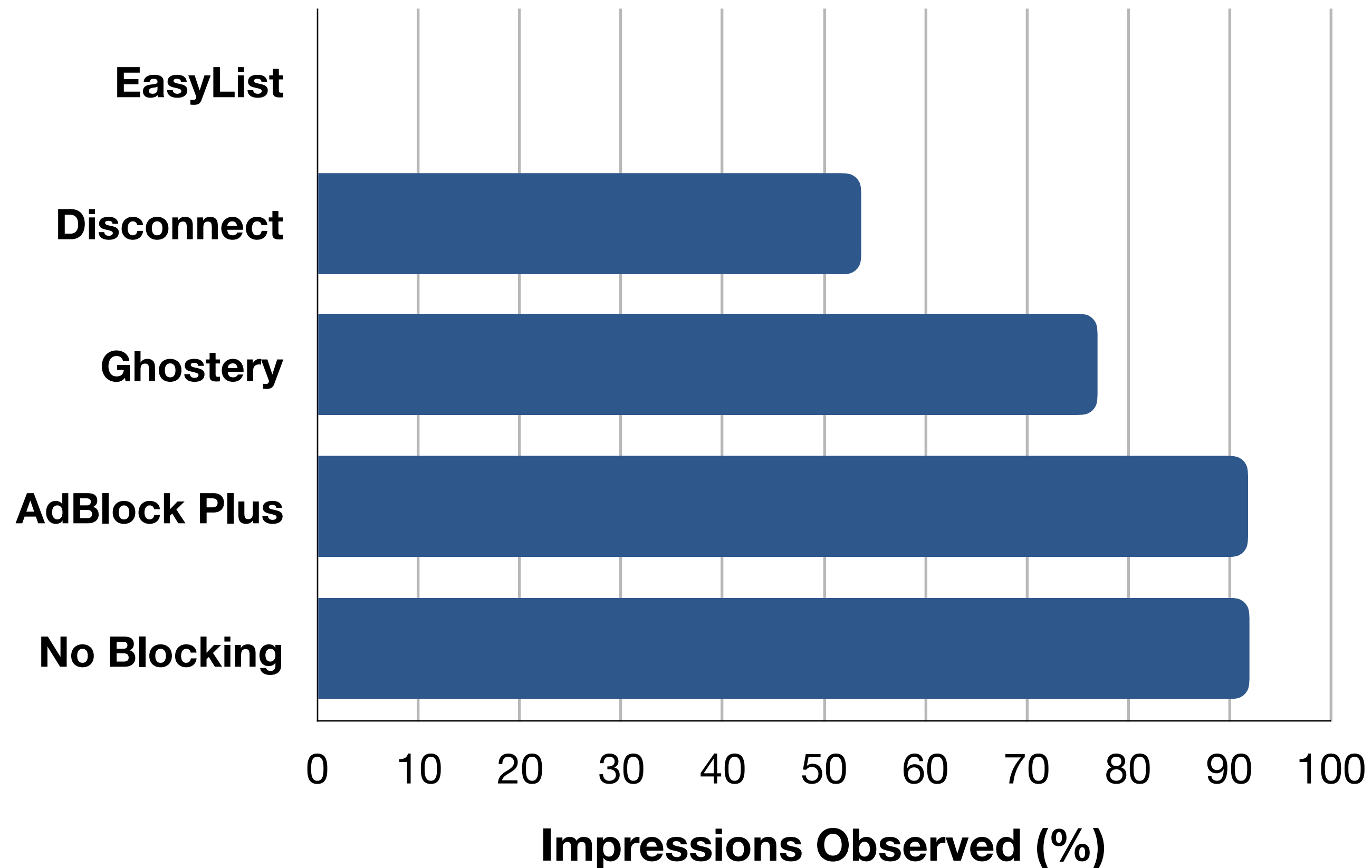


# Top 10 Domains



**Top 10 companies can view majority of user impressions even with (most) blocking extensions installed**

# Top 10 Domains



Domain	Impression %
google-analytics	97.0
youtube	91.7
quantserve	91.6
scorecardresearch	91.6
skimresources	91.3
twitter	91.1
pinterest	91.0
addthis	90.0
criteo	90.0
bluekai	90.8

**Top 10 domains with most observed impressions under AdBlock Plus**

**Top 10 companies can view majority of user impressions even with (most) blocking extensions installed**

# Simulation Limitations

- Our simulation models provide approximations
- Different users might have different browsing behaviors
  - We only simulate with respect to popular publishers
- The ecosystem could have changed from when the dataset was collected (December 2015)
- Not representative of mobile advertising ecosystem

# Summary

- We are the first to provide a model to study the impact of Real Time Bidding (RTB) on user privacy.
- Ad Exchanges share user impressions to facilitate RTB
  - More than 10% A&A domains view up to 90% of user impressions under realistic conditions.
- Due to RTB, impressions can leak to A&A domains even with blocking extensions
  - AdBlock Plus is not effective at all due to Acceptable Ads program
  - Disconnect performed the best in terms of protecting privacy

# Summary

- We are the first to provide a model to study the impact of Real Time Bidding (RTB) on user privacy.
- Ad Exchanges share user impressions to facilitate RTB
  - More than 10% A&A domains view up to 90% of user impressions under realistic conditions.
- Due to RTB, impressions can leak to A&A domains even with blocking extensions
  - AdBlock Plus is not effective at all due to Acceptable Ads program
  - Disconnect performed the best in terms of protecting privacy

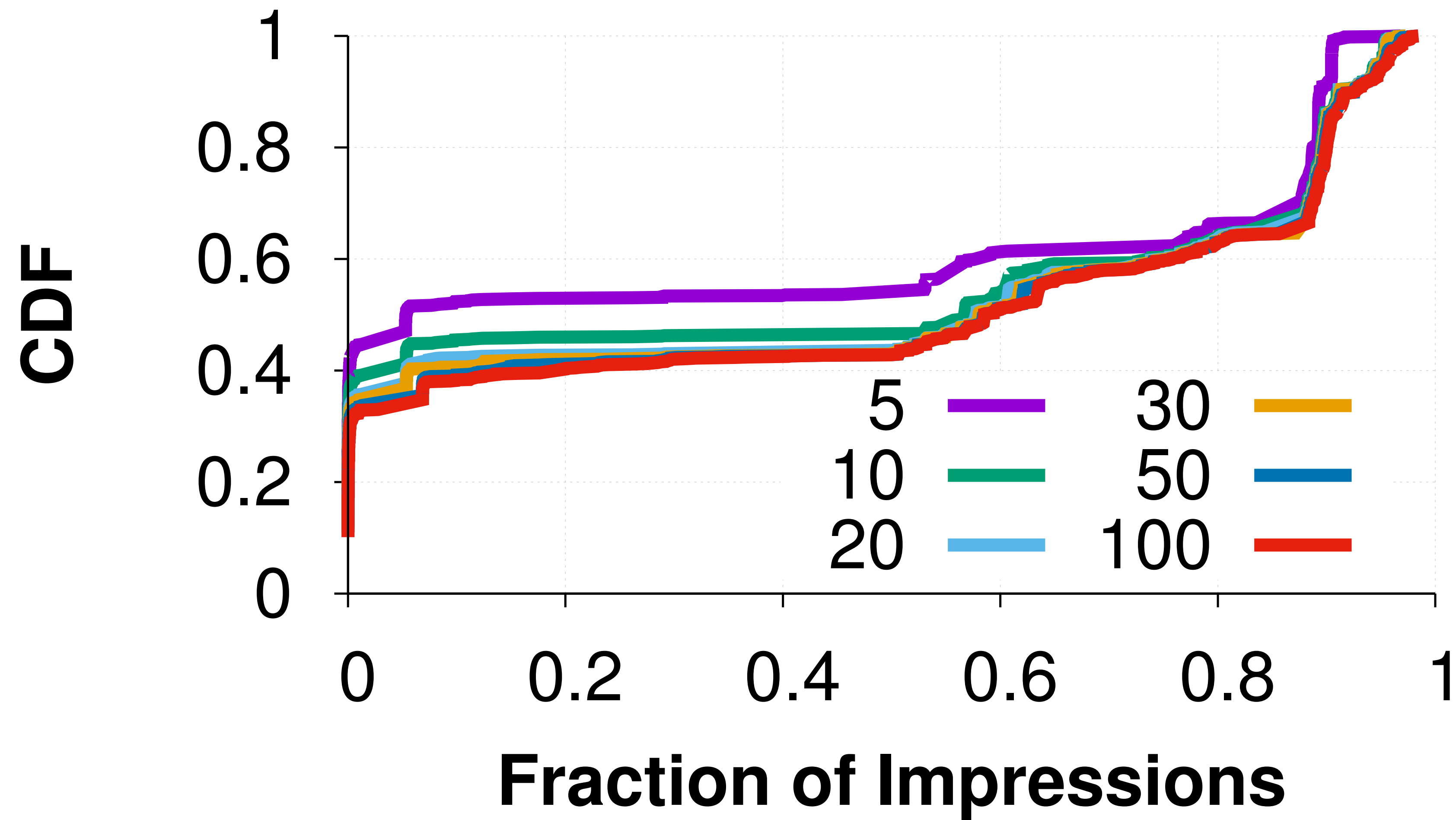
Questions?

ahmad@ccs.neu.edu

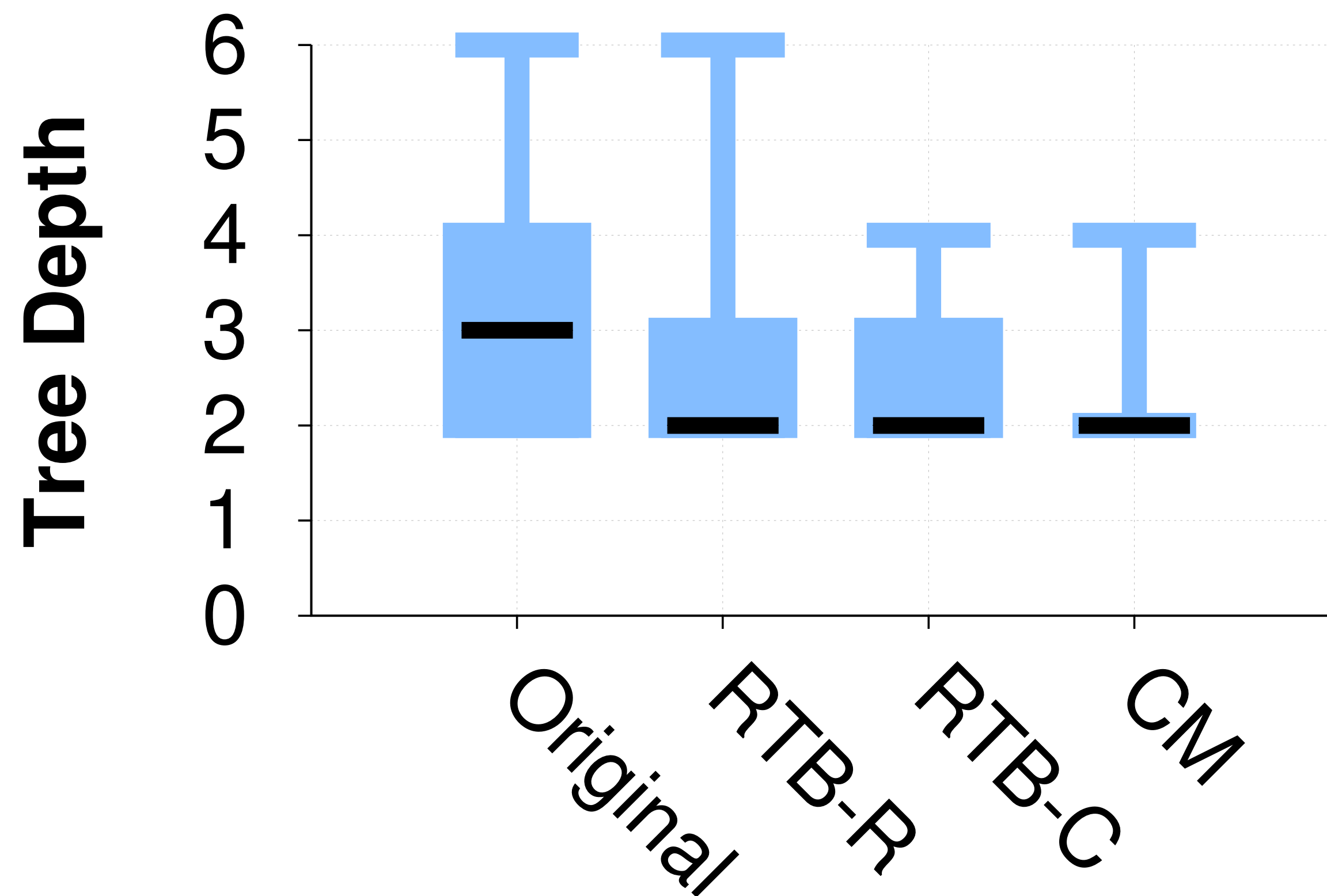
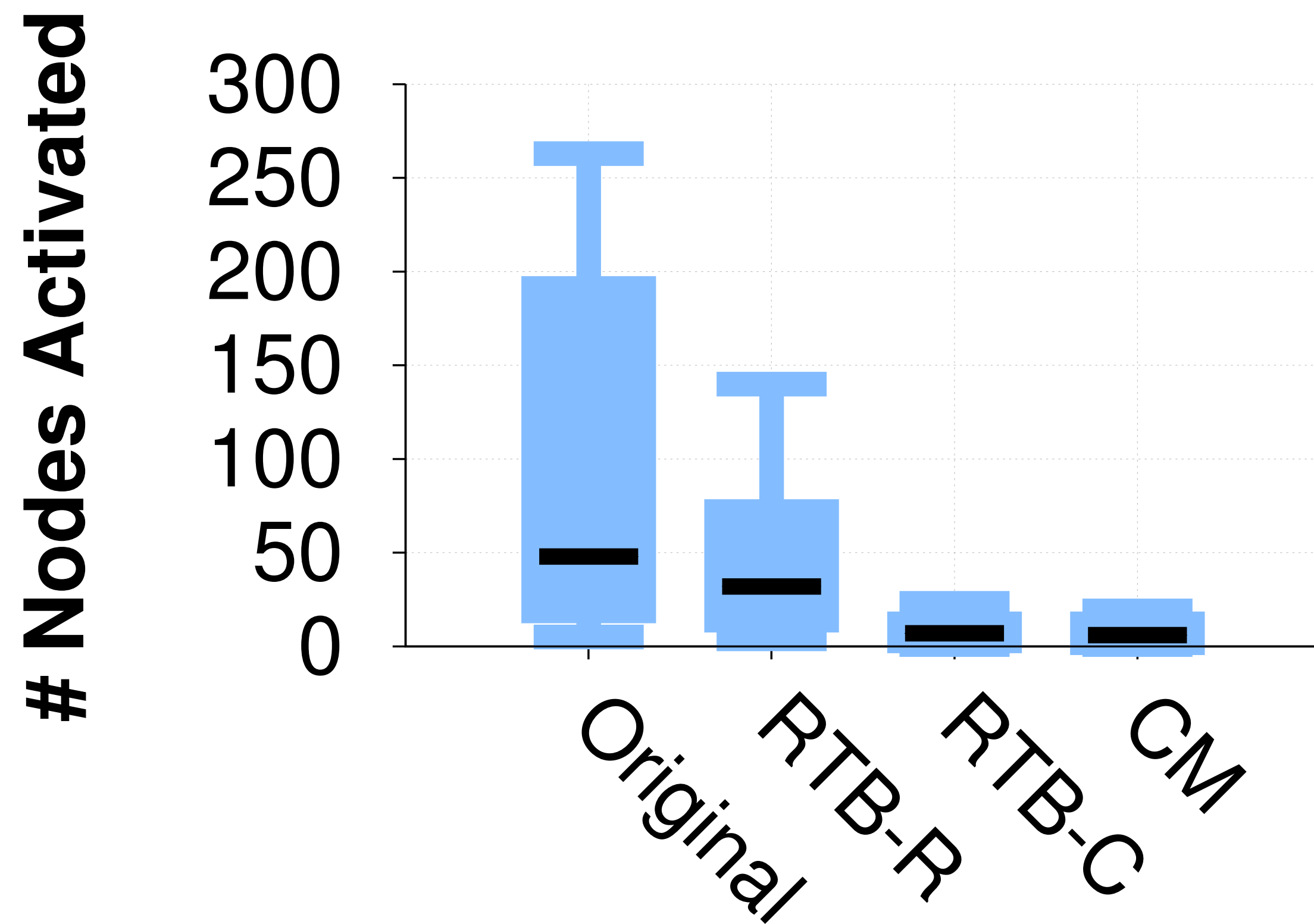
<http://personalization.ccs.neu.edu/Projects/AdGraphs/>

# Backup Slides

# Number of Exchanges for RTB-C



# Model Validation — Per Publisher





# Inclusion Chain from DOM

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**

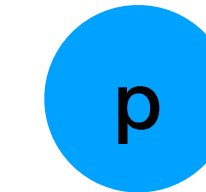
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



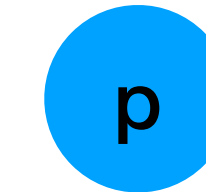
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    → <script src="a1.com/cookie-match.js" </script>
      <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
      

      <iframe src="a3.com/banner.html">
        <script src="a4.com/ad.js"> </script>
      </iframe>
    </body>
  </html>
```

**DOM Tree for <http://p.com/index.html>**



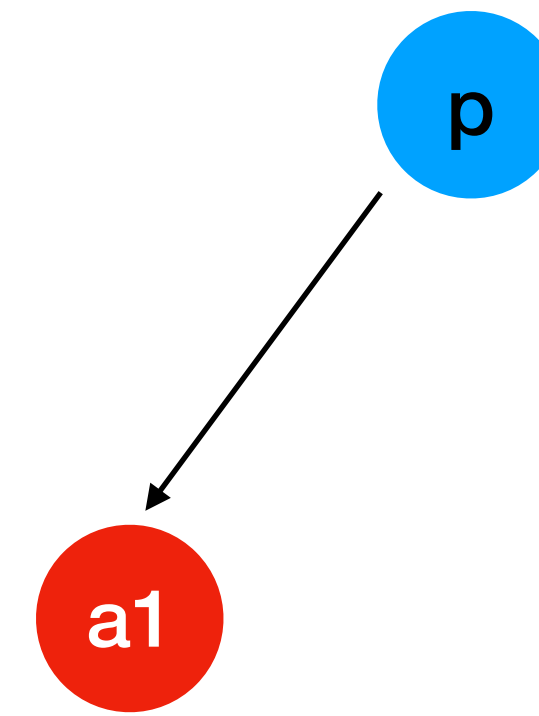
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    → <script src="a1.com/cookie-match.js" </script>
      <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
      

      <iframe src="a3.com/banner.html">
        <script src="a4.com/ad.js"> </script>
      </iframe>
    </body>
  </html>
```

**DOM Tree for <http://p.com/index.html>**



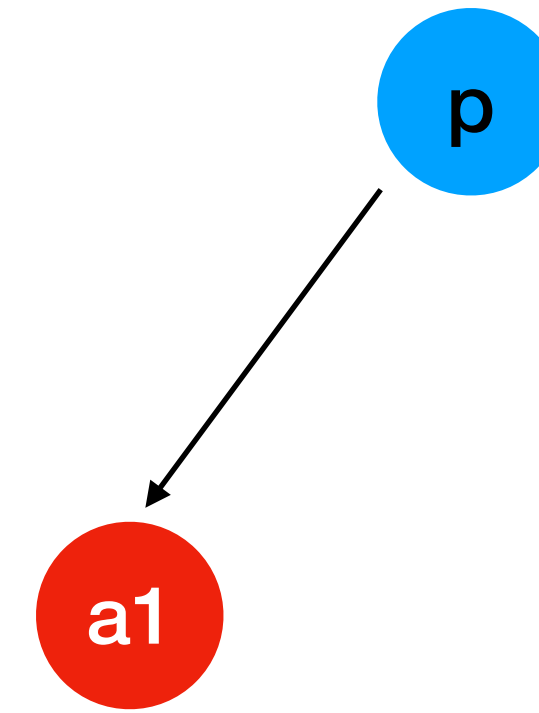
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    → 

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



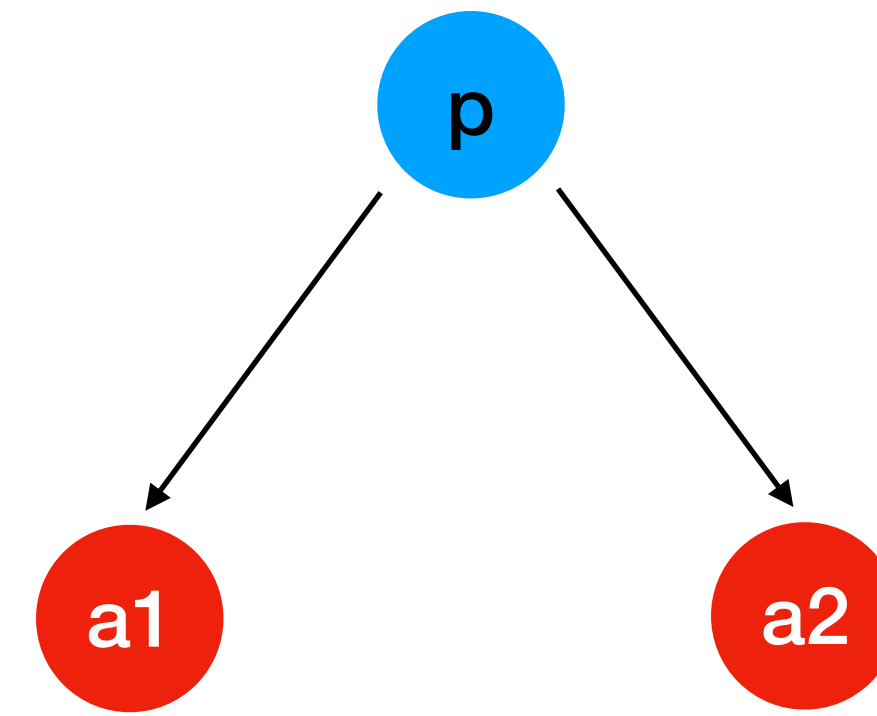
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    → 

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



**Inclusion of resources**

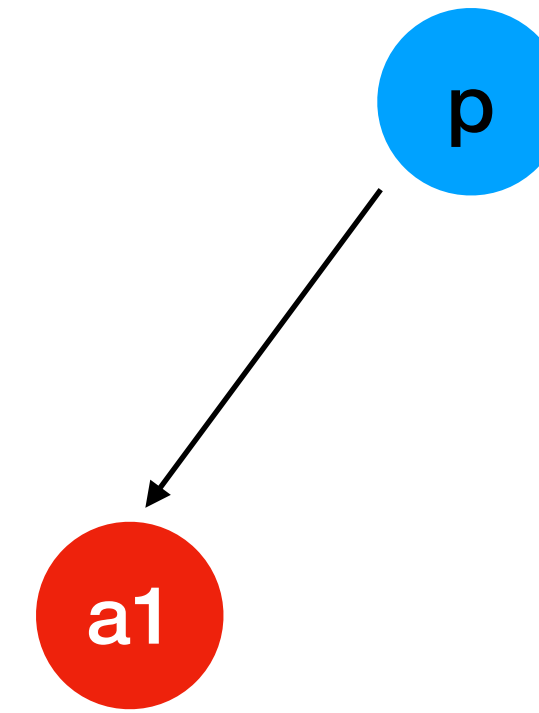


# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    → 

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



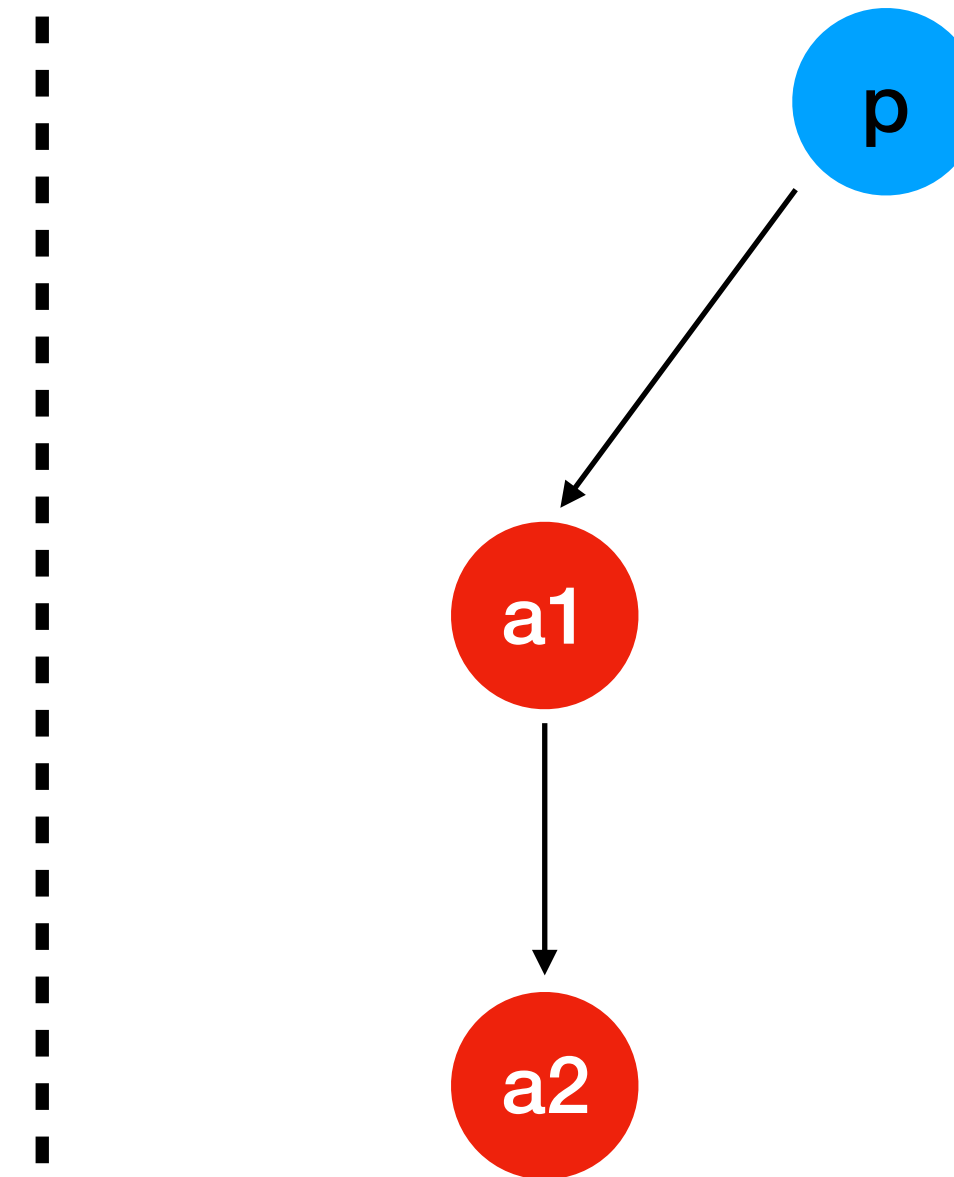
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    → 

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**

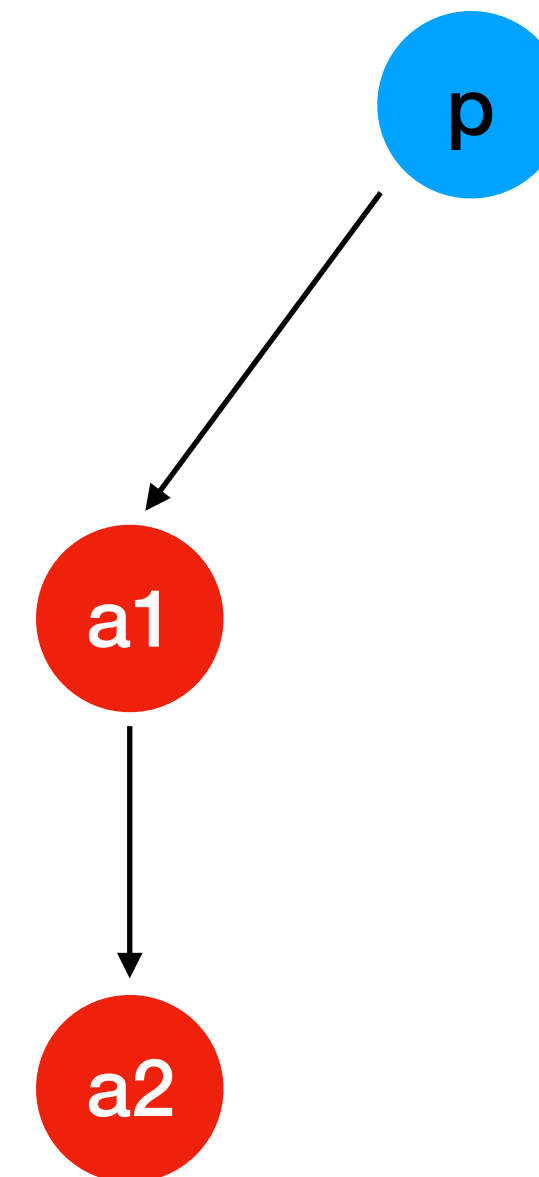


**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    
    <!--
    
    <!--
    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**

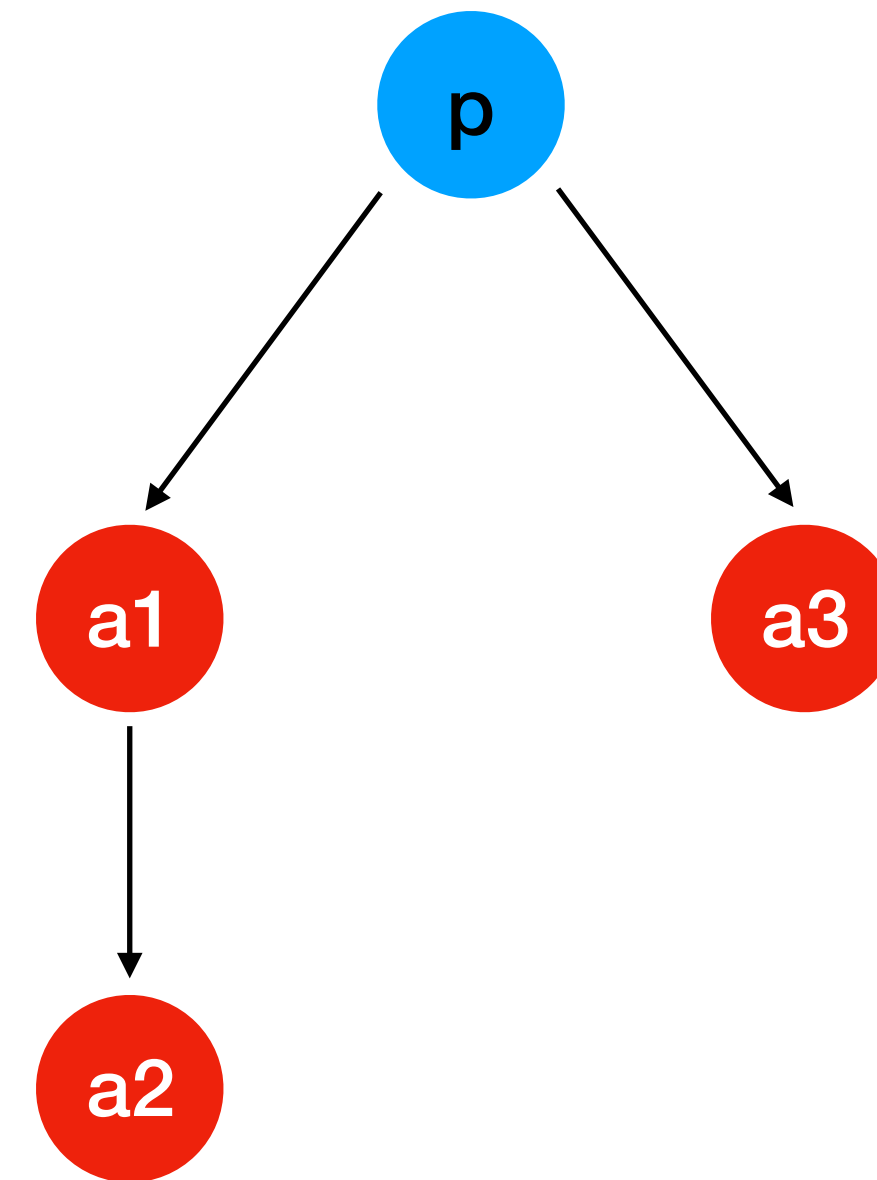


**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
        by cookie-match.js -->
    
    → <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js"> </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



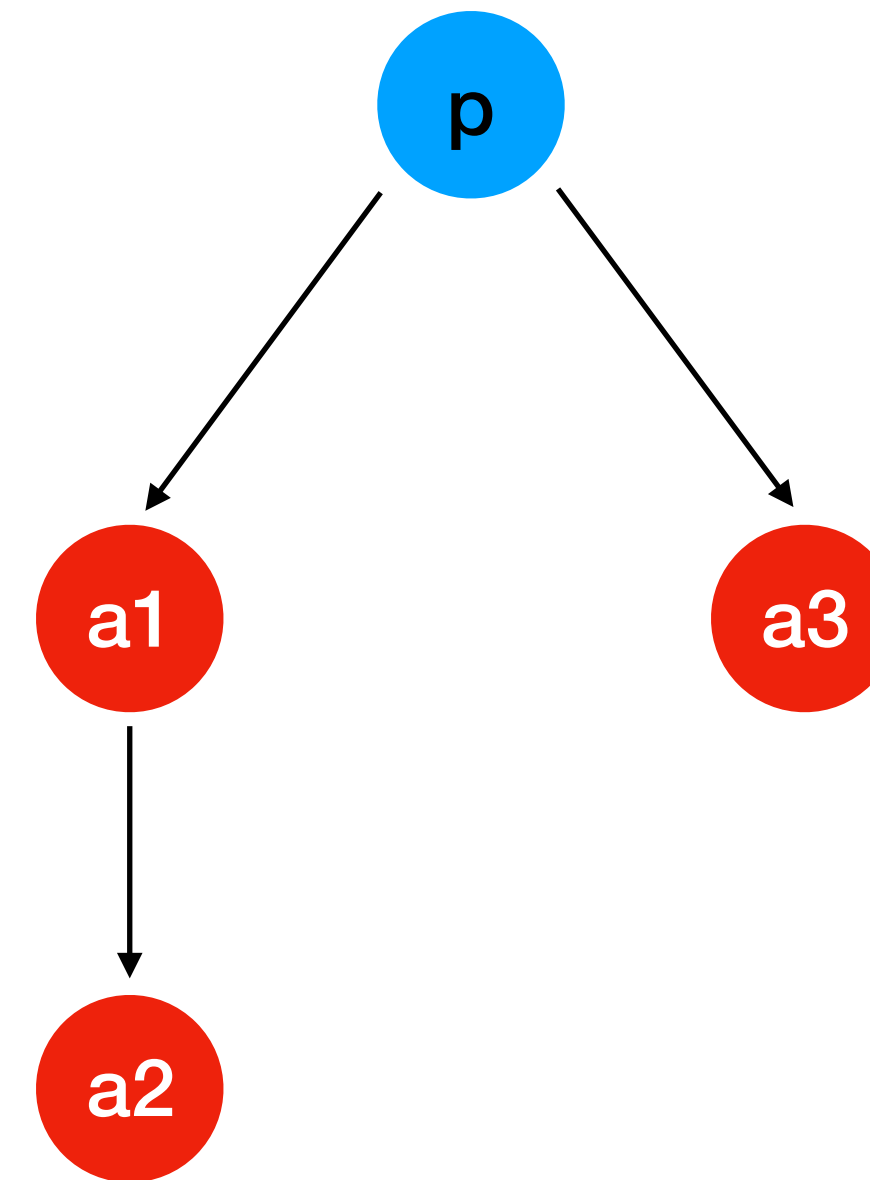
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js" </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



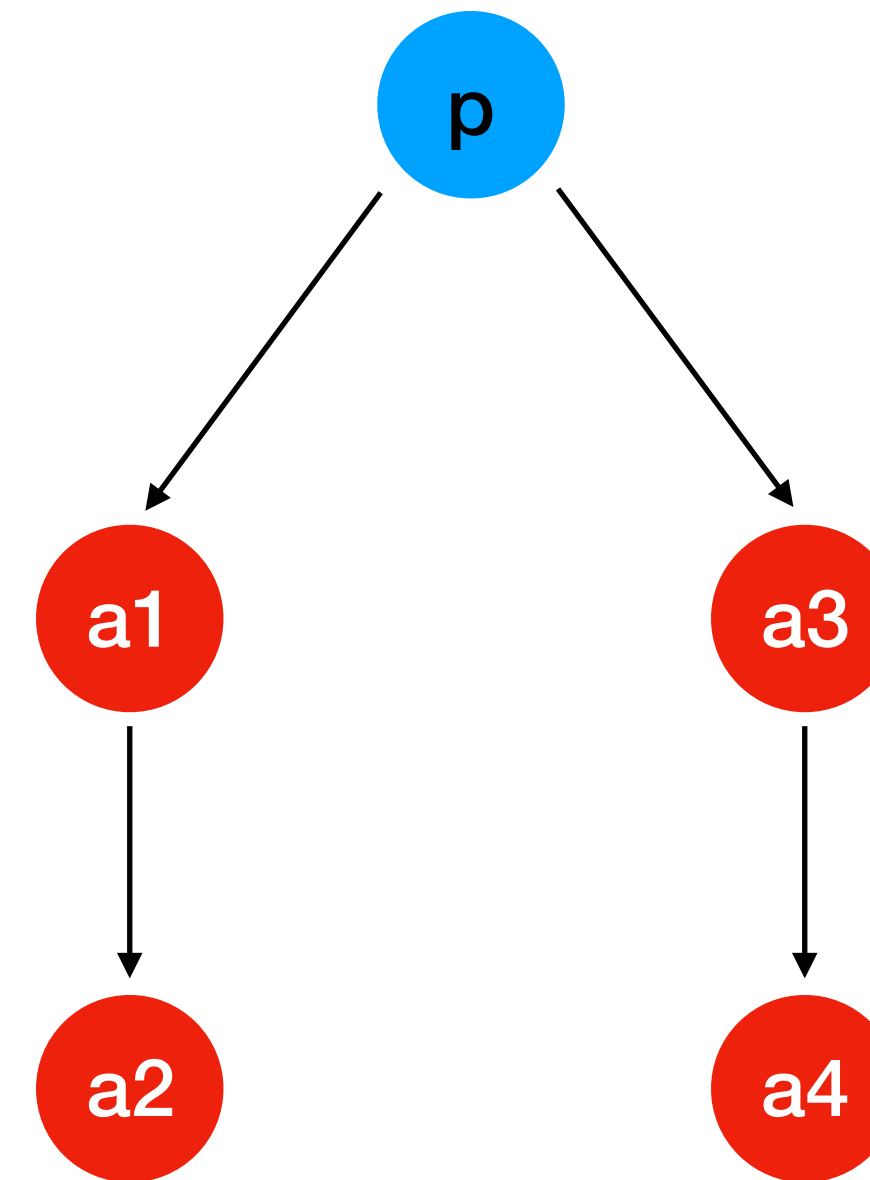
**Inclusion of resources**

# Inclusion Chain from DOM

```
<html>
  <body>
    <script src="a1.com/cookie-match.js" </script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ad.js" </script>
    </iframe>
  </body>
</html>
```

**DOM Tree for <http://p.com/index.html>**



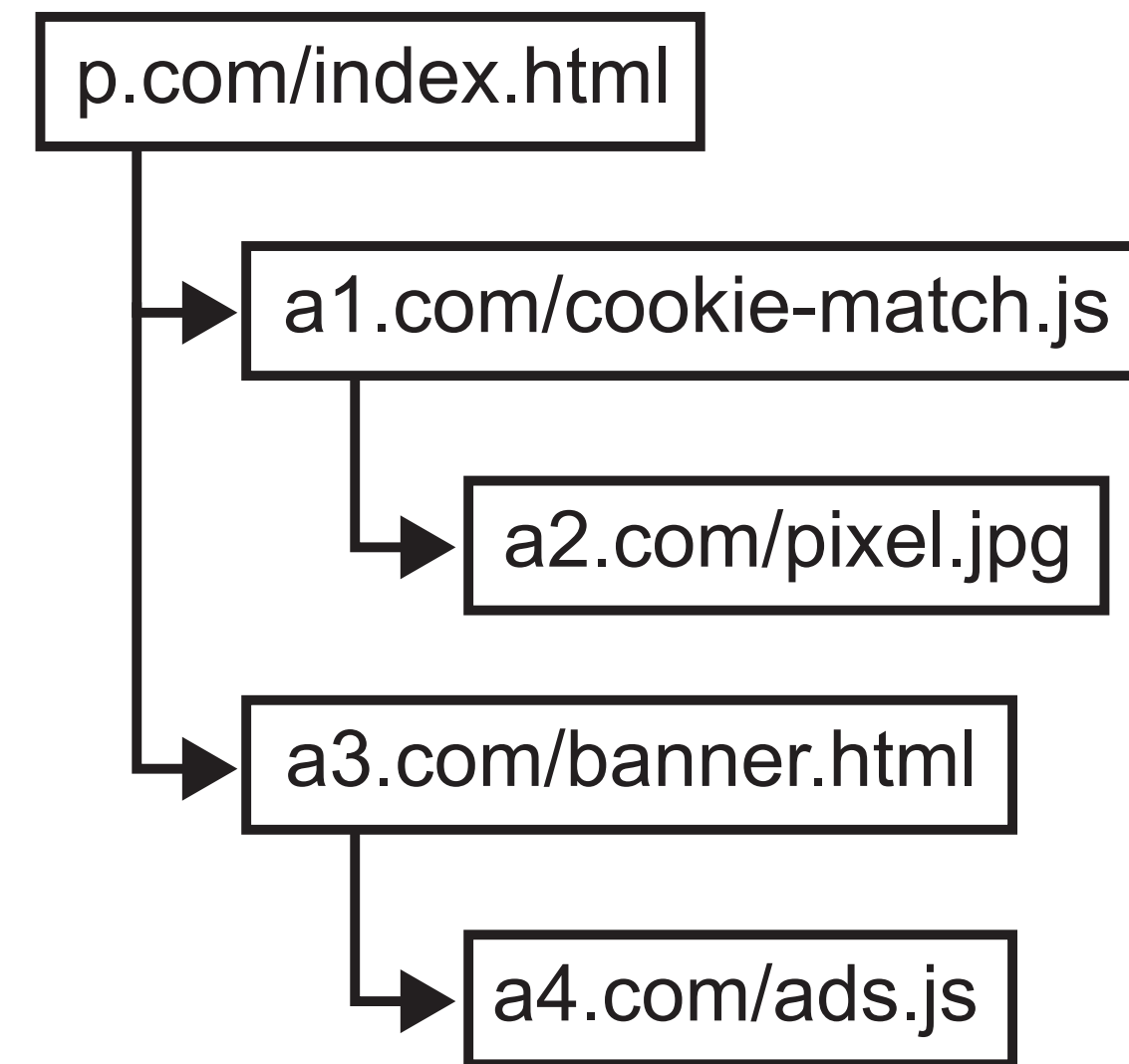
**Inclusion of resources**

# DOM -> Inclusion & Referrer Graph

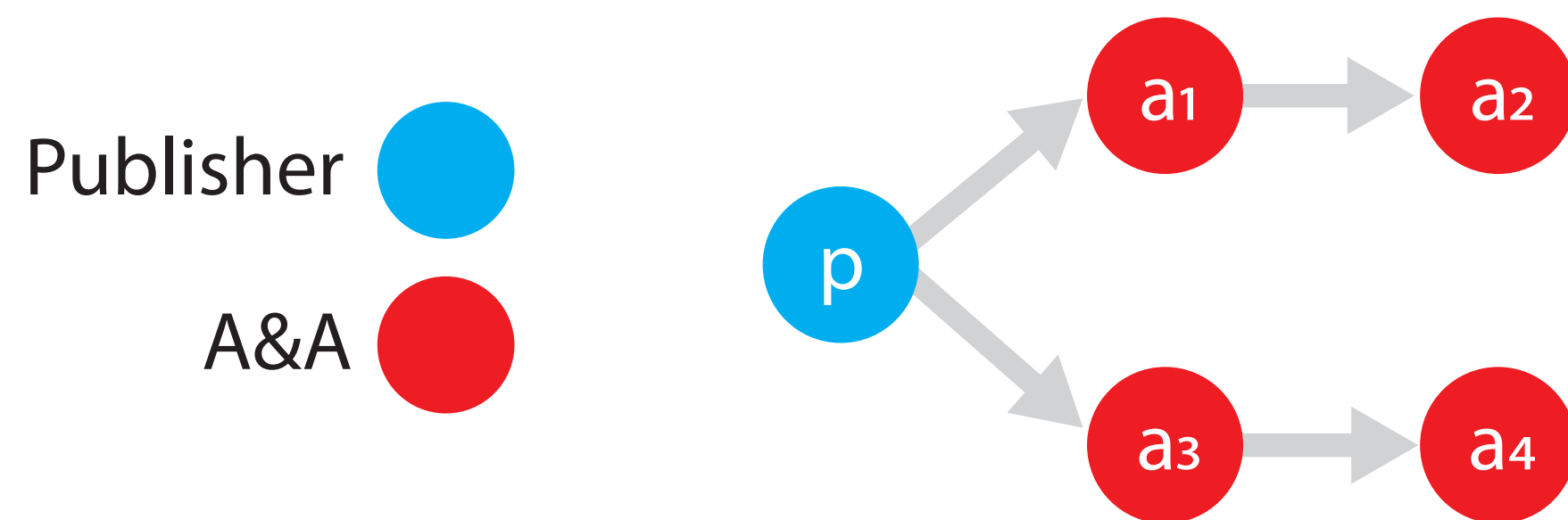
```
<html>
  <body>
    <script src="a1.com/cookie-match.js"></script>
    <!-- Tracking pixel inserted dynamically
         by cookie-match.js -->
    

    <iframe src="a3.com/banner.html">
      <script src="a4.com/ads.js"></script>
    </iframe>
  </body>
</html>
```

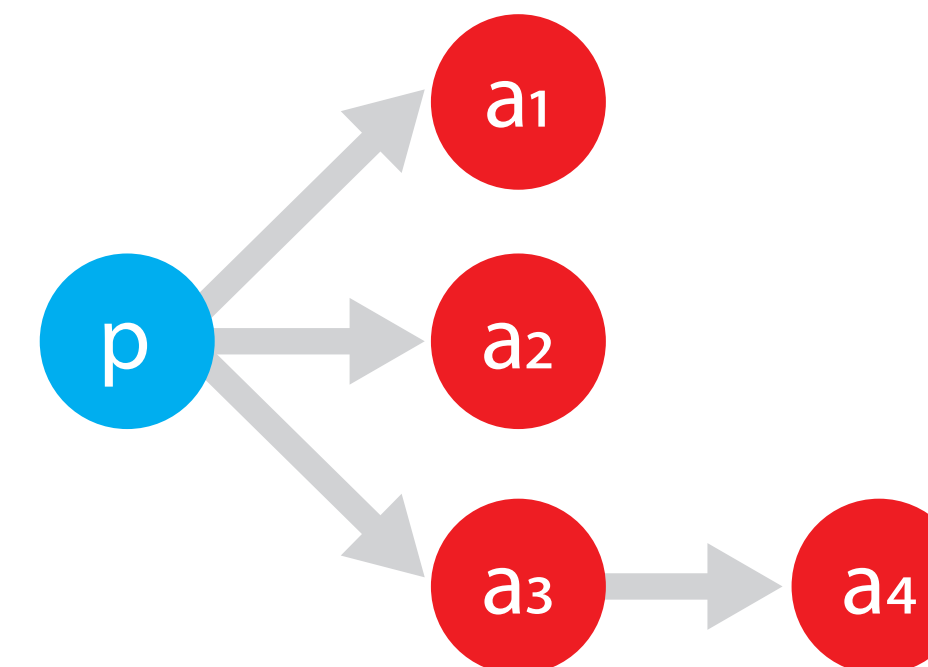
(a) DOM Tree for *http://p.com/index.html*



(b) Inclusion Tree



(c) Inclusion Graph



(d) Referrer Graph

# Comparison with Random Model

