In Plain Sight: Online Tracking and Profiling

July 2014

(Speaker Info)
Natasa Milic-Frayling
Principal Researcher
Microsoft Research

Microsoft Research

Microsoft

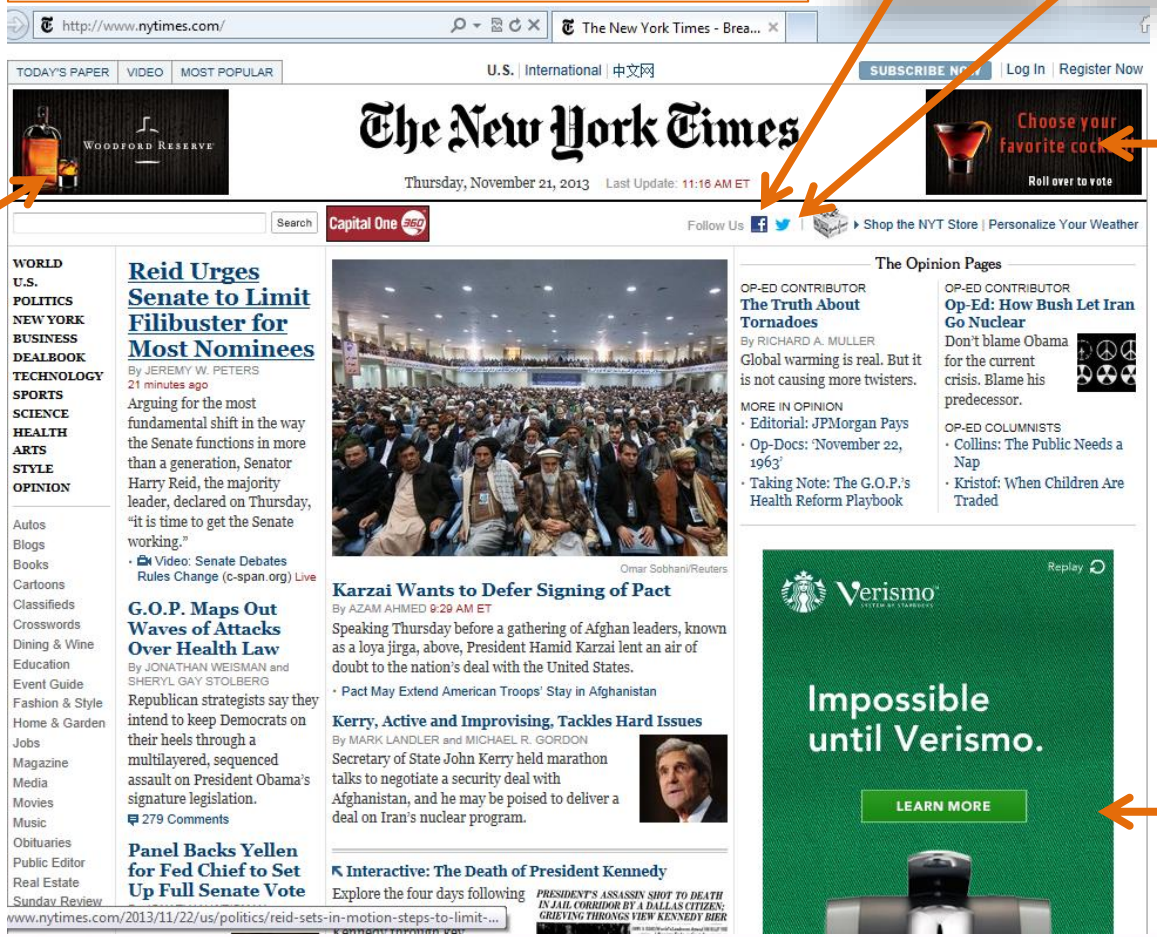
Web and third parties

User awareness of online tacking

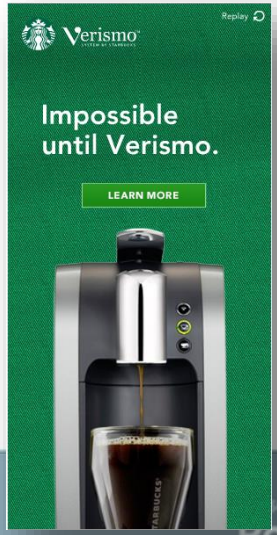
FIRST PARTY: New York Times

Third parties

FIRST PARTY: New York Times



THIRD PARTIES: providers of content, ads, analytics



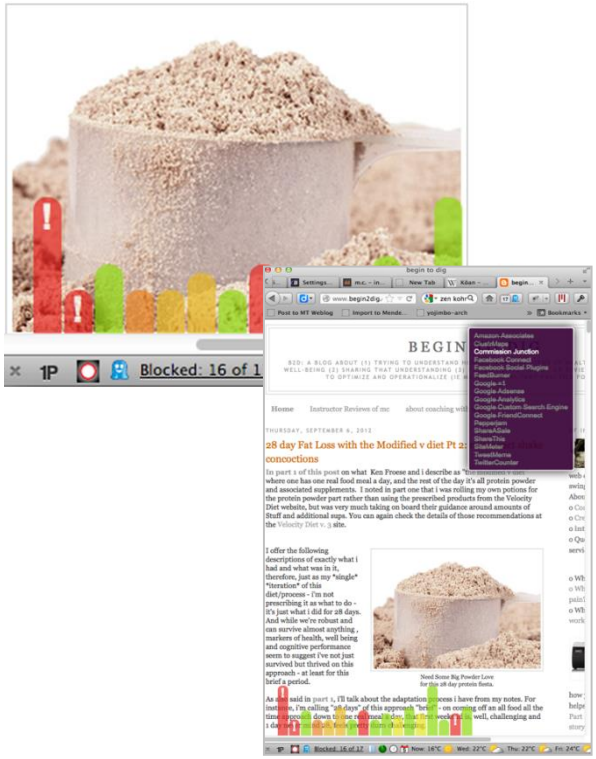
Third party tracking

Tracking user visits to first party websites

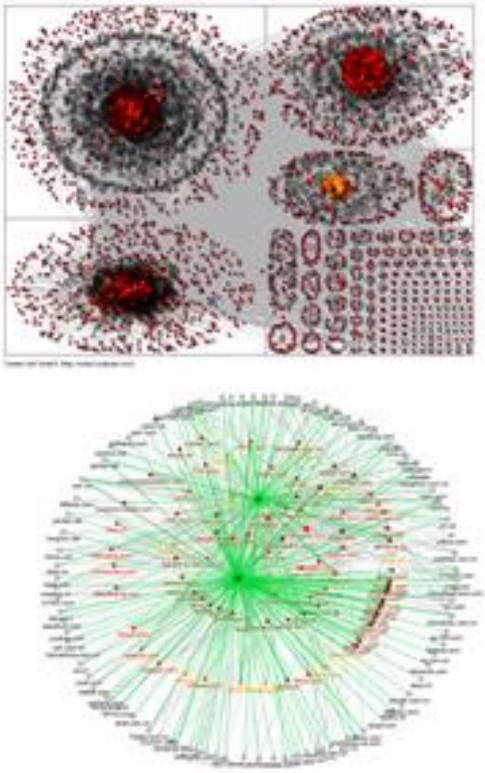
- Facilitated through co-operation with the first party
  - Websites receive income from a third-party advertising network
- Independently, by exploiting security vulnerabilities
  - cross site scripting (XSS)
  - cross site request forgery (CSRF)
- Web cookies are a common tracking mechanism
- Users can also be tracked by
  - data stored in their web browser cache
  - HTML5 local-storage
  - E-Tag data
  - Flash locally-stored objects (LSOs)
  - the long-lived unique IDs provided by many mobile devices.

Tracking through cookies

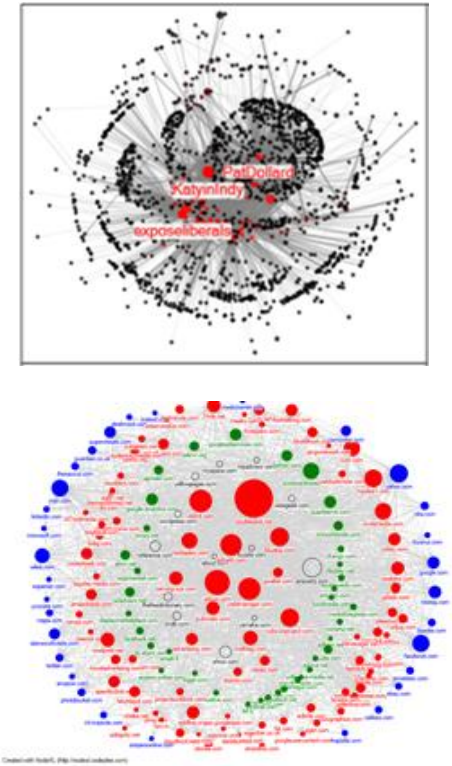
User Study:
Lifting the Lid on Cookies



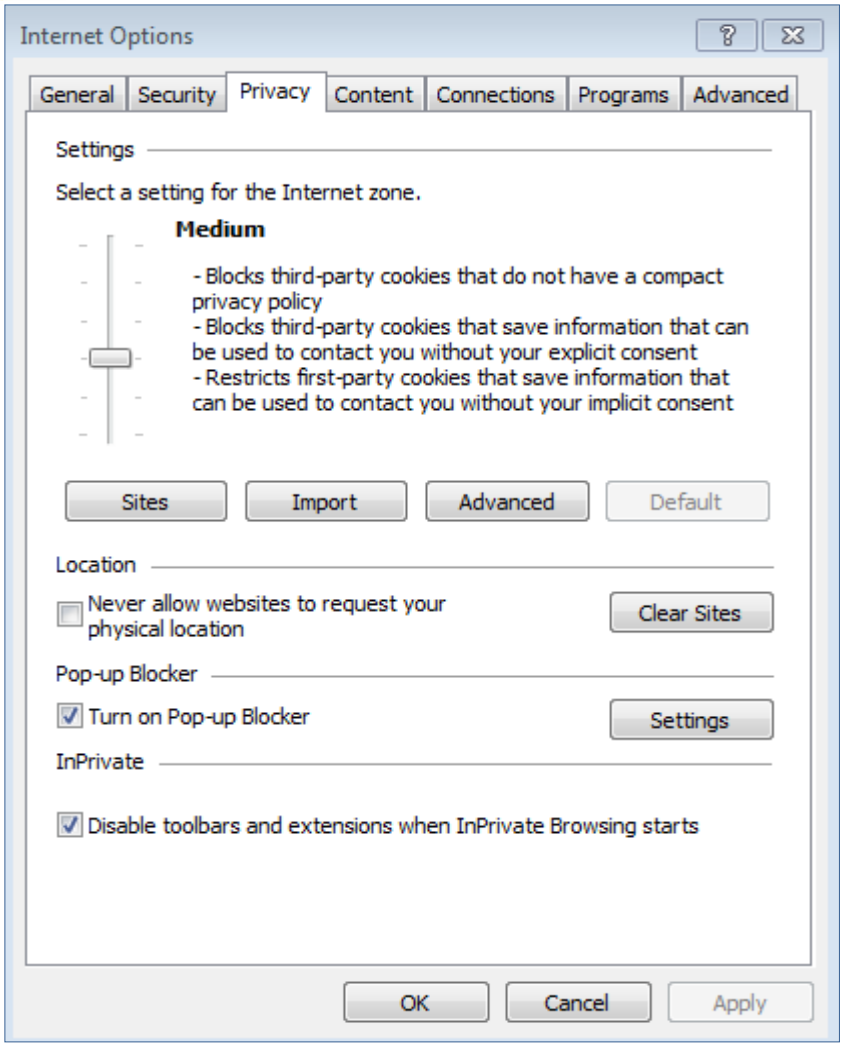
Analysis of User Tracking
Networks through Search



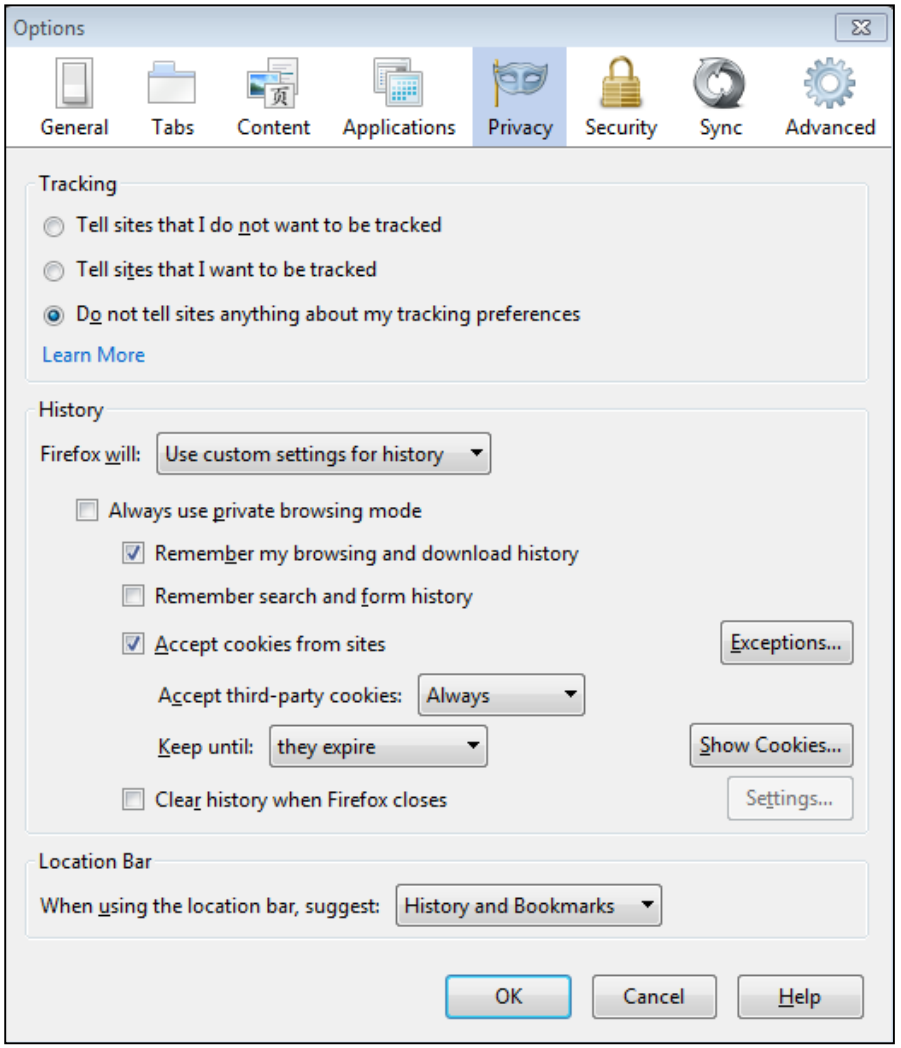
Propagation of cookies
through Social Networks



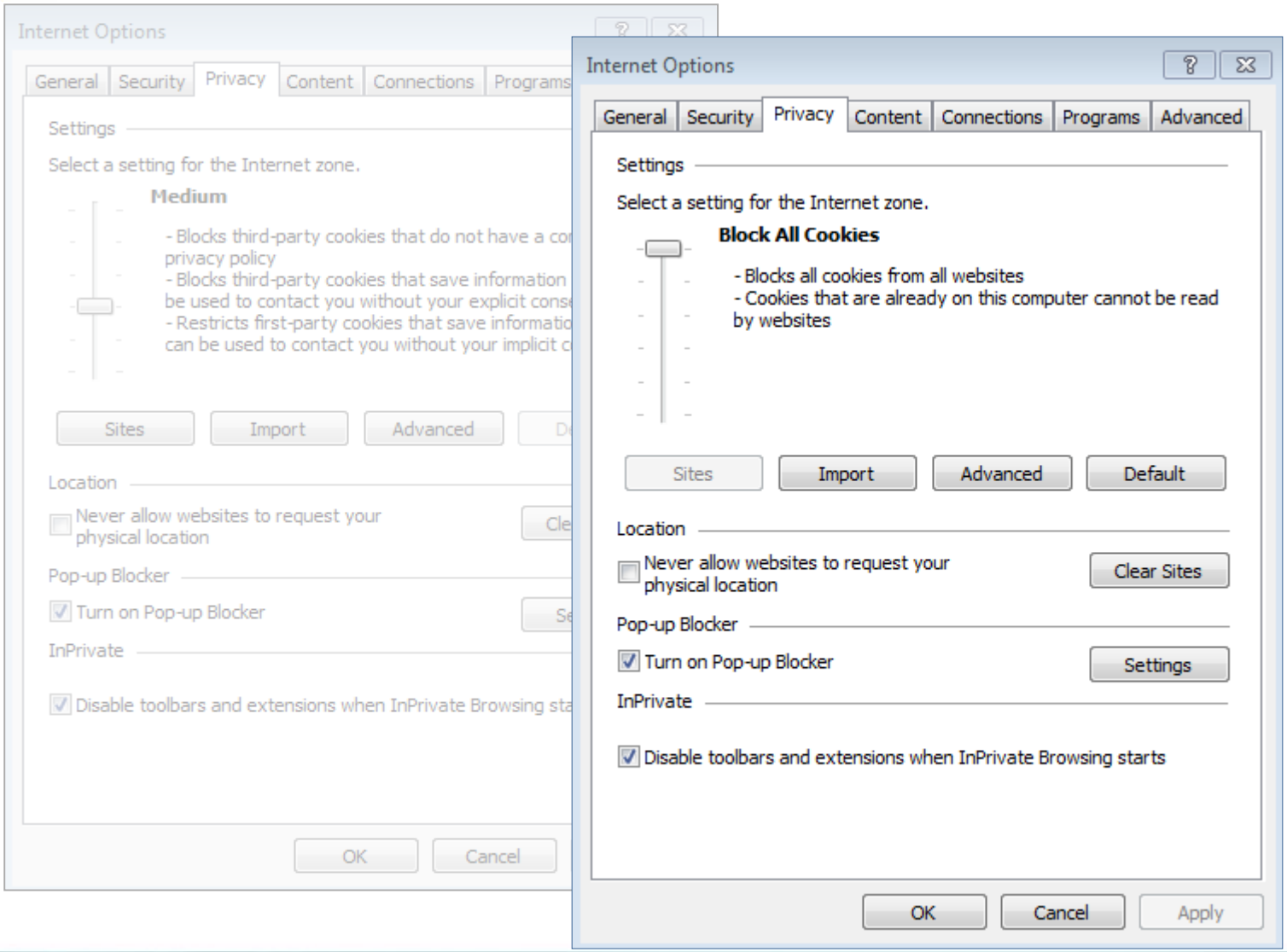
Microsoft IE



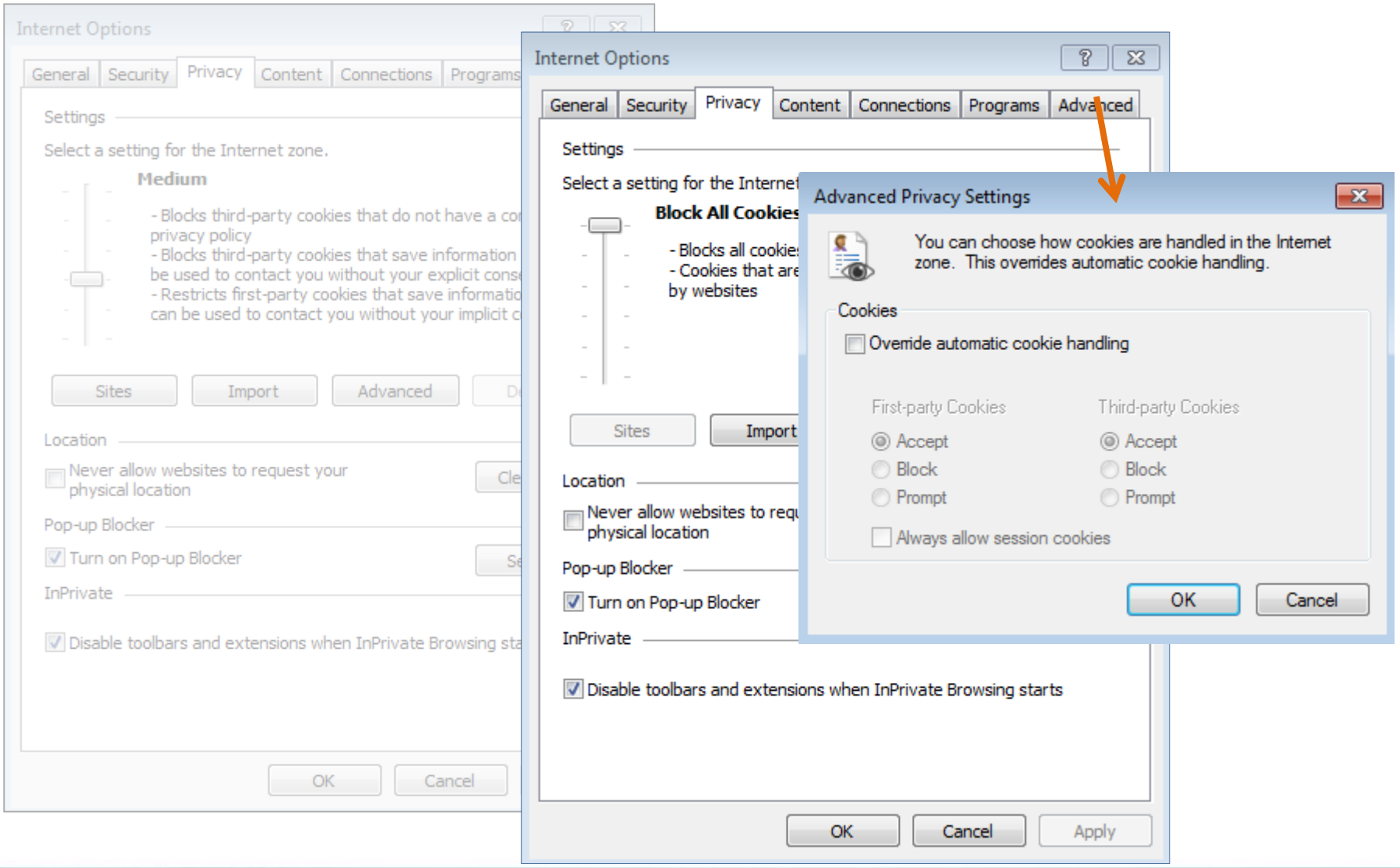
Firefox



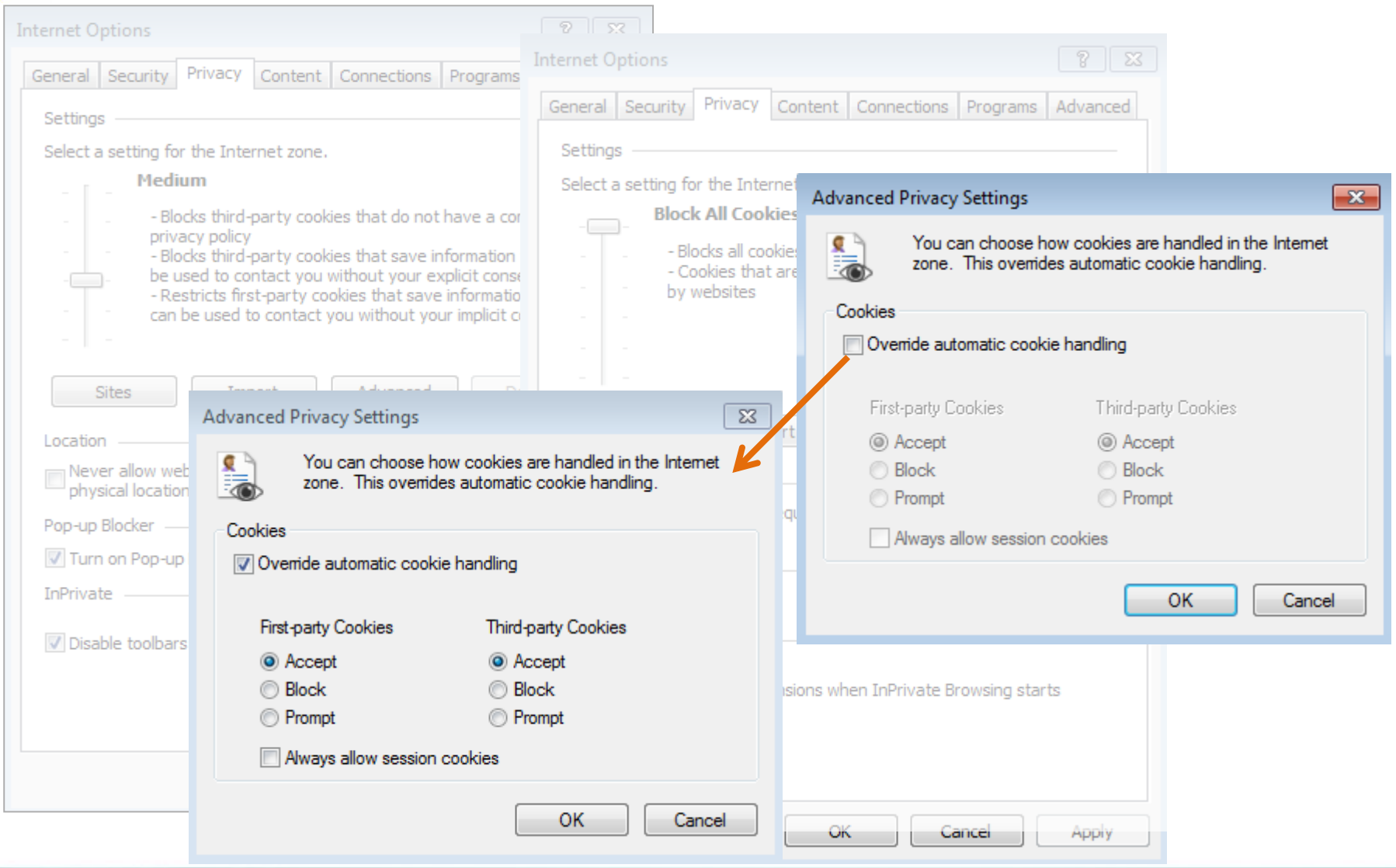
Microsoft IE



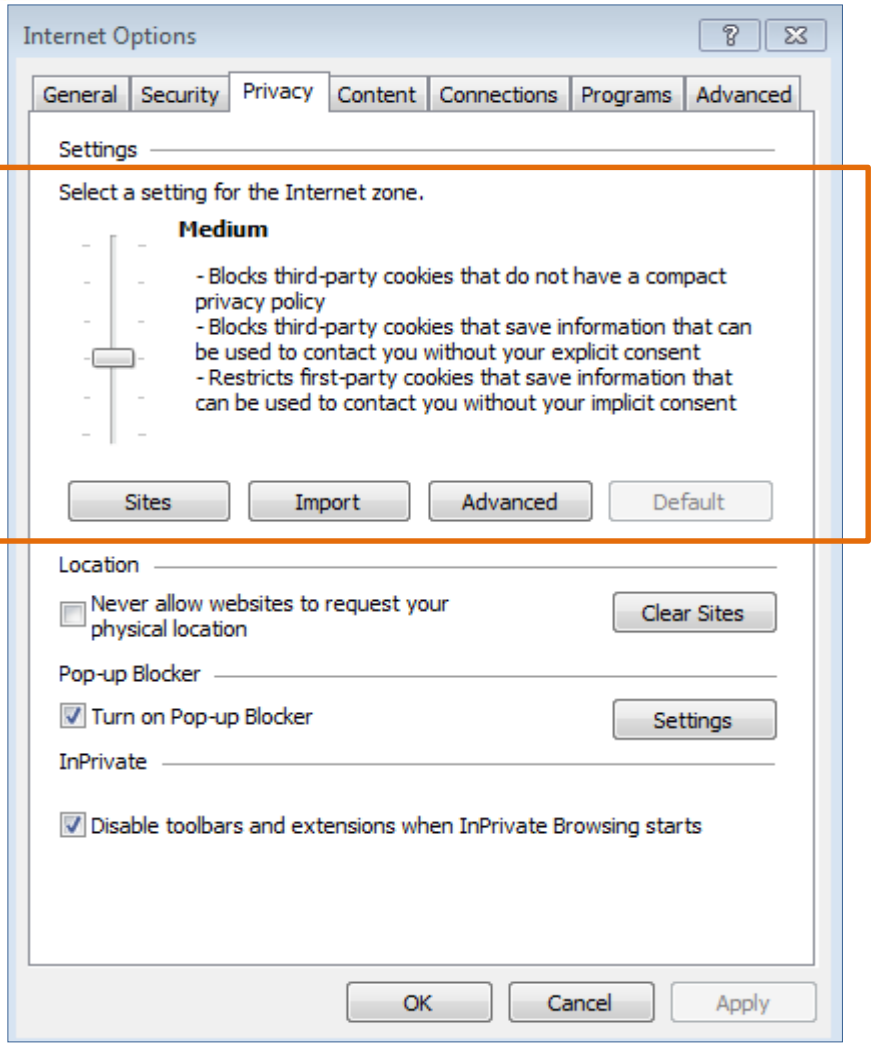
Microsoft IE



Microsoft IE

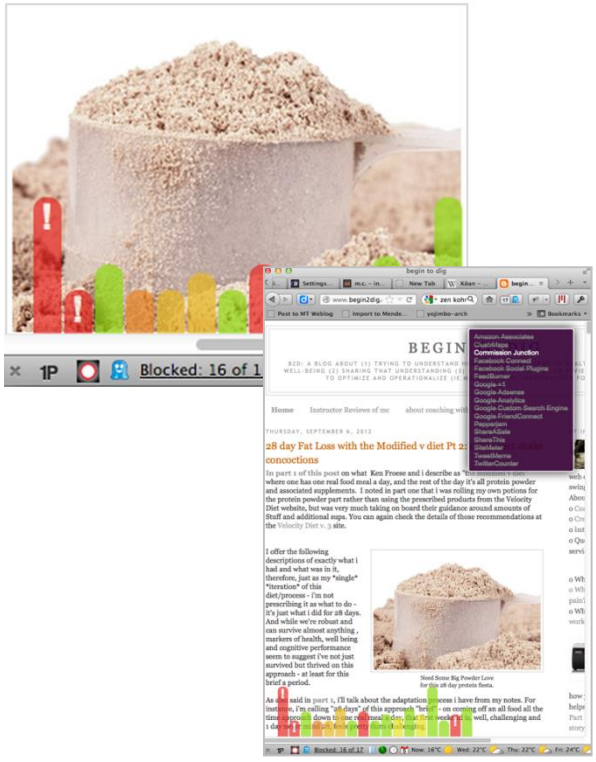


Microsoft IE



Lifting the Lid on Cookies

User Study:
Lifting the Lid on Cookies



Research topic:

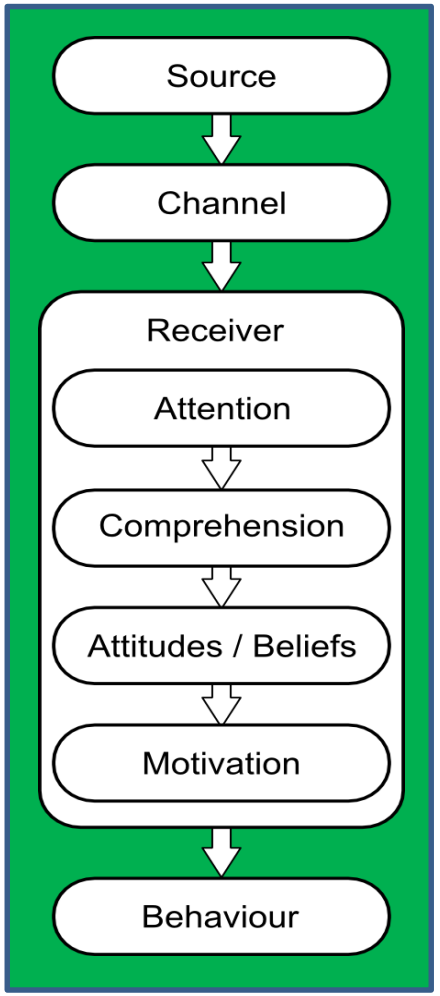
- Designs to increase the user awareness and understanding of tracking activities

Framework:

- Communication-Human Information Processing (C-HIP) model of warning effectiveness

Research method

- Technical probe: Browser extensions to Comparison of browser extensions.
- Quantitative analysis of the tracking activities.



Conzola, V.C. and Wogalter, M.S. A Communication–Human Information Processing (C–HIP) approach to warning effectiveness in the workplace. Journal of Risk Research 4, 4 (2001), 309–322.

Cookies and Search

Tracking networks based on the http referral header

Search and Tracking

Observations:

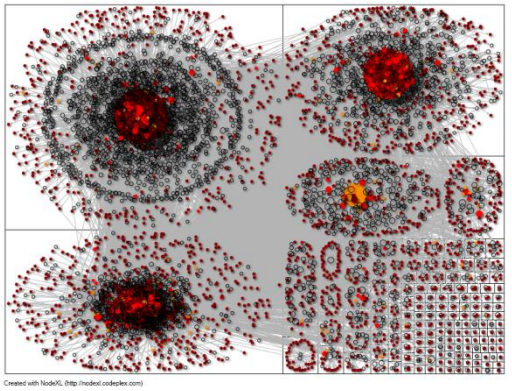
UI features facilitate the value exchange between individuals and services

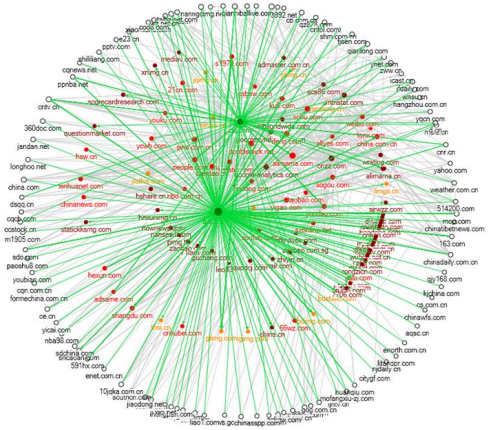
Value exchange is unclear

Research:

- Data analysis of search results and tracking companies associated with them
- Uncover the characteristics of the tracking network and model the value exchange between the consumer and services

Analysis of User Tracking Networks through Search





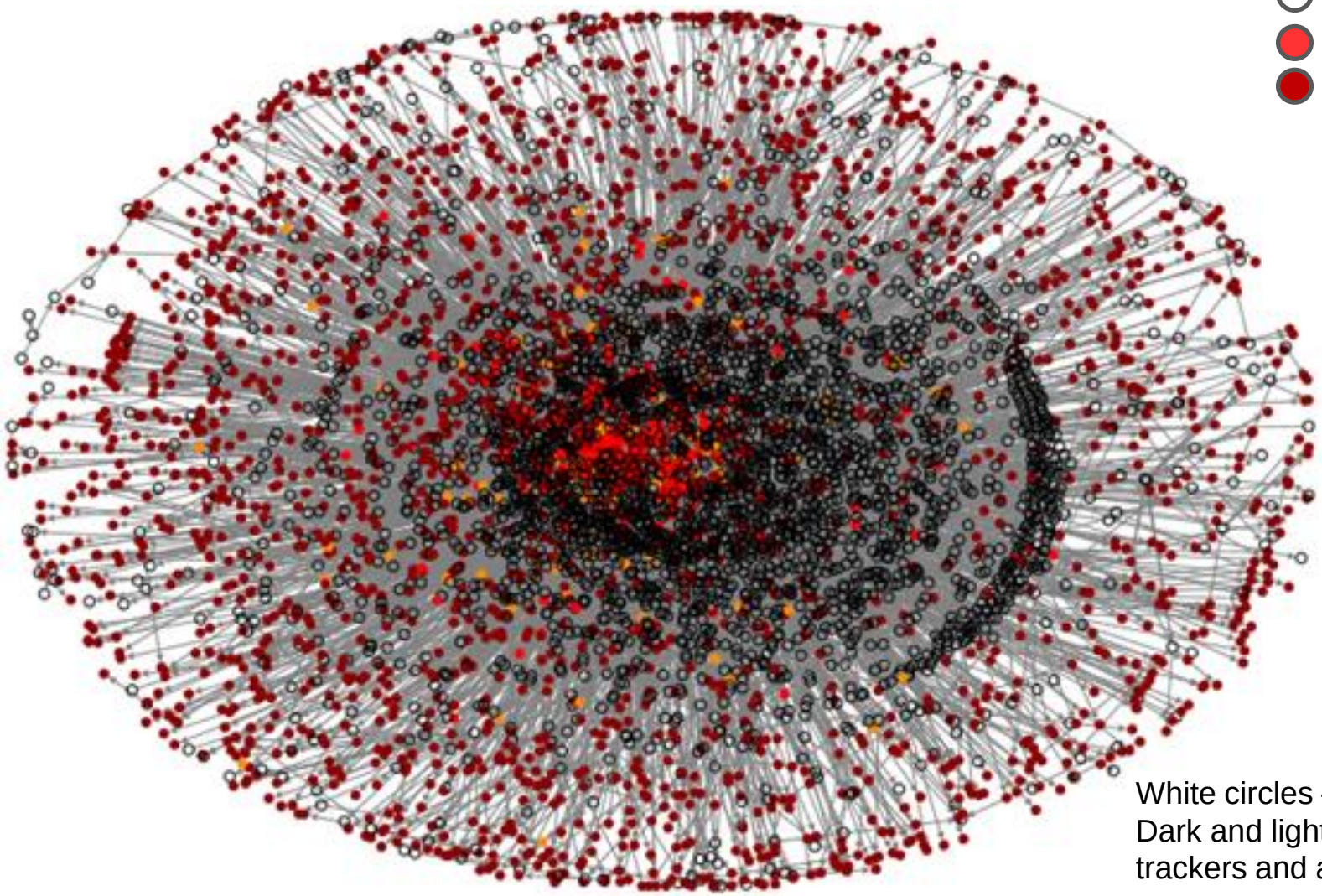
Data set – Search Queries

- KDD Cup 2005 Challenge
- 800 queries with assigned categories by three assessors
- Categorization involved three level categories:
  - Top level: Computers (8), Entertainment (9), Information (8), Living (18), Online Community (6), Shopping (6), Sports (11)
  - 67 Second and Third level categories
- Selected queries with higher label agreement among assessors: **662 queries**

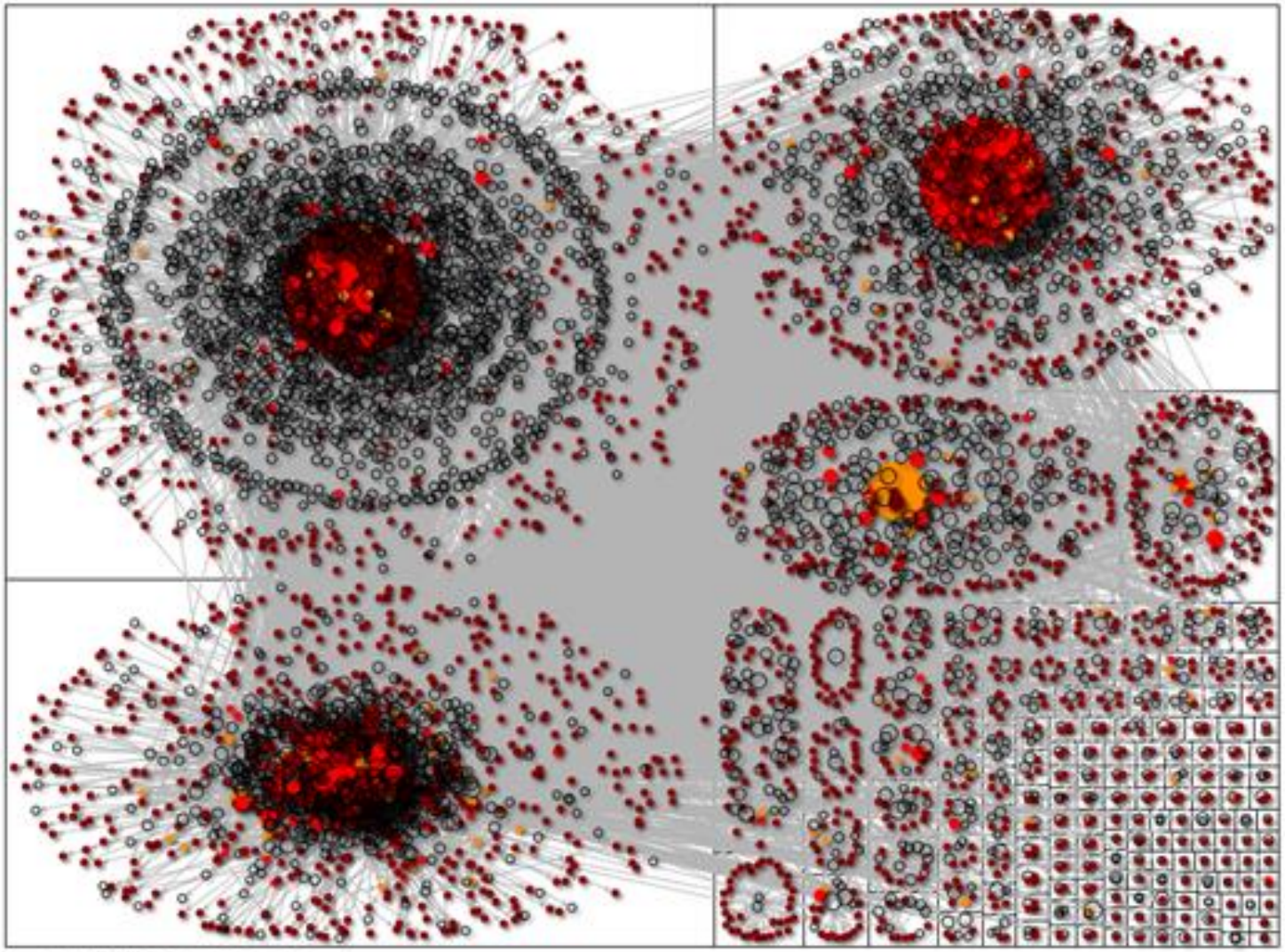
Category Label	Num of SearchQueries
Shopping\Stores & Products	101
Information\Local & Regional	95
Information\Companies & Industries	60
Living\Health & Fitness	49
Living\Car & Garage	41
Information\Law & Politics	40
Living\Travel & Vacation	39
Living\Fashion & Apparel	37
Information\Science & Technology	36
Living\Finance & Investment	34
Living\Food & Cooking	33
Information\Education	30

Data set – Retrieved Documents

Market	Bing API Identifier	Google Search Domains
India	en-IN	www.google.co.in
South Africa	en-ZA	www.google.co.za
United Kingdom	en-UK	www.google.co.uk
United States	en-US	www.google.com

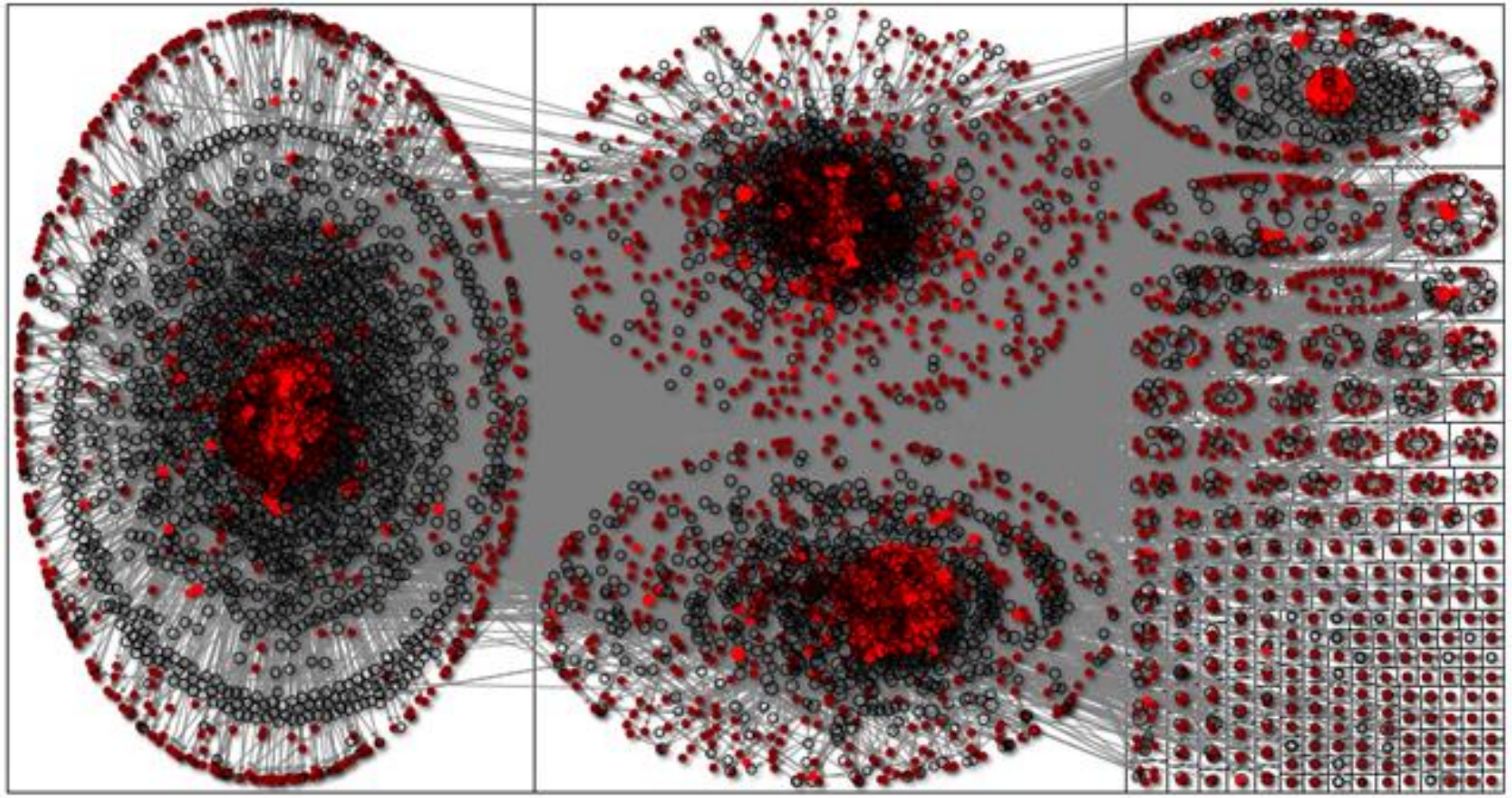


Tracking network uncovered through Google search in the India Search Market. Shows one giant connected component.



Tracking network uncovered through Google search in the India Search Market. Clustered.

White circles – Web sites
Dark and light red circles – trackers and ad brokers



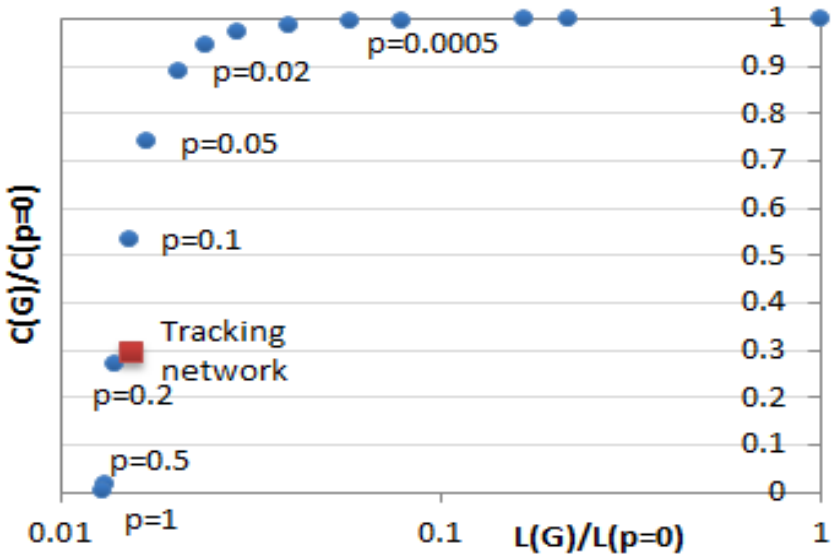
Tracking network uncovered through Google search in the US Search Market. Clustered.

White circles – Web sites
Dark and light red circles – trackers and ad brokers

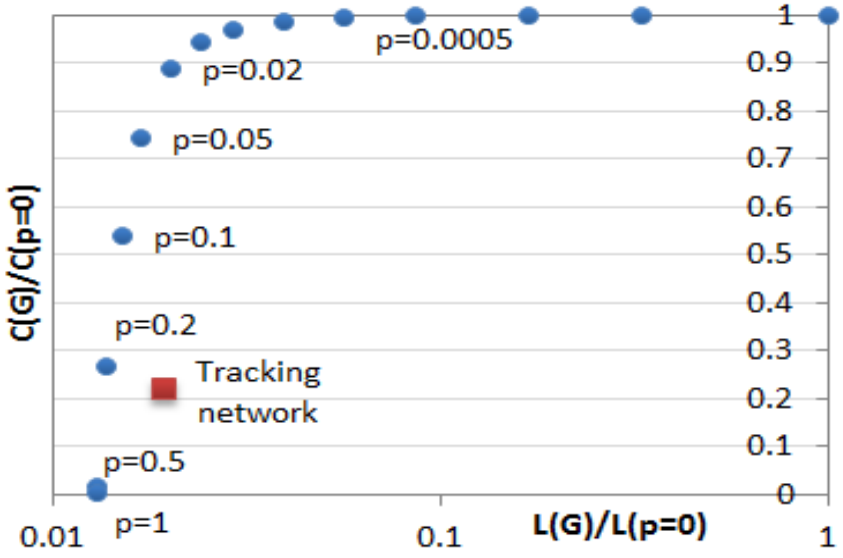
	Google				Bing				Baidu
Tracking network	**US**	**UK**	**S. Af.**	**IN**	**US**	**UK**	**S. Af.**	**IN**	**CN**
Nodes N(G)	5958	6171	5991	6000	5850	6638	5938	6321	473
Edges E(G)	67739	73374	70411	66038	79214	81015	80171	79243	4868
Unique edges E'(G)	26203	26552	25763	26058	25951	28047	26061	26625	1117
Clustering coeff.	0.1958	0.1947	0.1993	0.2078	0.2105	0.1818	0.2053	0.2082	0.1685
Avg. node degree	8.7959	8.6054	8.6006	8.6860	8.8721	8.4504	8.7777	8.4243	4.7230
Connected comp.	405	381	398	402	358	436	405	461	12
Giant component	**US**	**UK**	**S. Af.**	**IN**	**US**	**UK**	**S. Af.**	**IN**	**CN**
Nodes N(GC) /N(G)	92%	93%	93%	92%	93%	93%	92%	92%	97%
Edges E'(GC) /E'(G)	99.8%	99.8%	99.8%	99.8%	99.8%	99.8%	99.8%	99.7%	99.6%4

	Google				Bing				Baidu
Tracking network	US	UK	S. Af.	IN	US	UK	S. Af.	IN	CN
Nodes N(G)	5958	6171	5991	6000	5850	6638	5938	6321	473
Edges E(G)	67739	73374	70411	66038	79214	81015	80171	79243	4868
Unique edges E'(G)	26203	26552	25763	26058	25951	28047	26061	26625	1117
Clustering coeff.	0.1958	0.1947	0.1993	0.2078	0.2105	0.1818	0.2053	0.2082	0.1685
Avg. node degree	8.7959	8.6054	8.6006	8.6860	8.8721	8.4504	8.7777	8.4243	4.7230
Connected comp.	405	381	398	402	358	436	405	461	12
Giant component	US	UK	S. Af.	IN	US	UK	S. Af.	IN	CN
Nodes N(GC) /N(G)	92%	93%	93%	92%	93%	93%	92%	92%	97%
Edges E'(GC) /E'(G)	99.8%	99.8%	99.8%	99.8%	99.8%	99.8%	99.8%	99.7%	99.6%4

Small World Property of the Tracking Network



Full tracking network



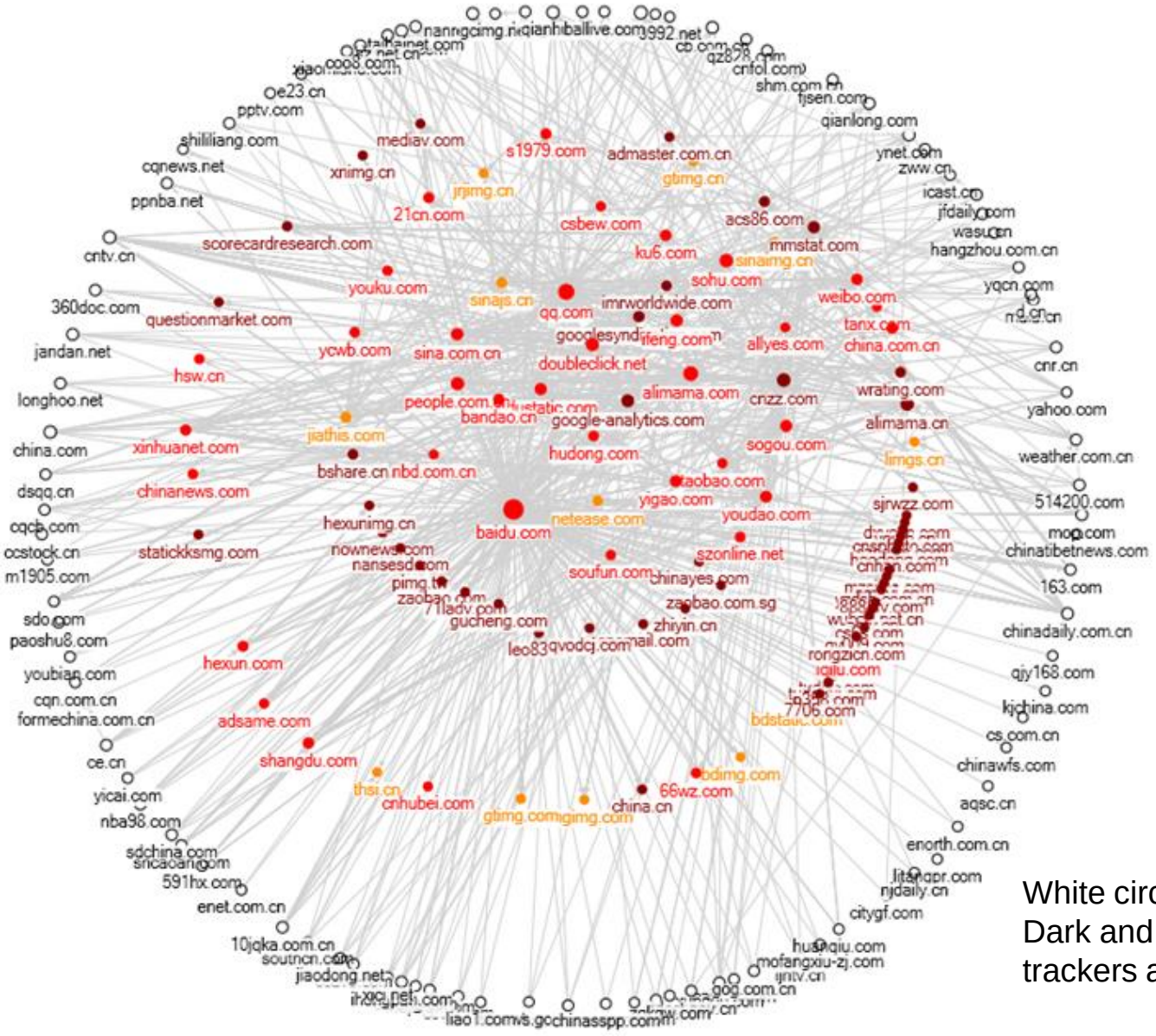
After doubleclick node is removed

Comparison of the synthetic (Watt-Strogatz random model) and the observed tracking network, based on the average path and the clustering coefficient.

Tracking network follows the small world network closely, for the rewiring probability p=0.2, (8.91 average node degree)

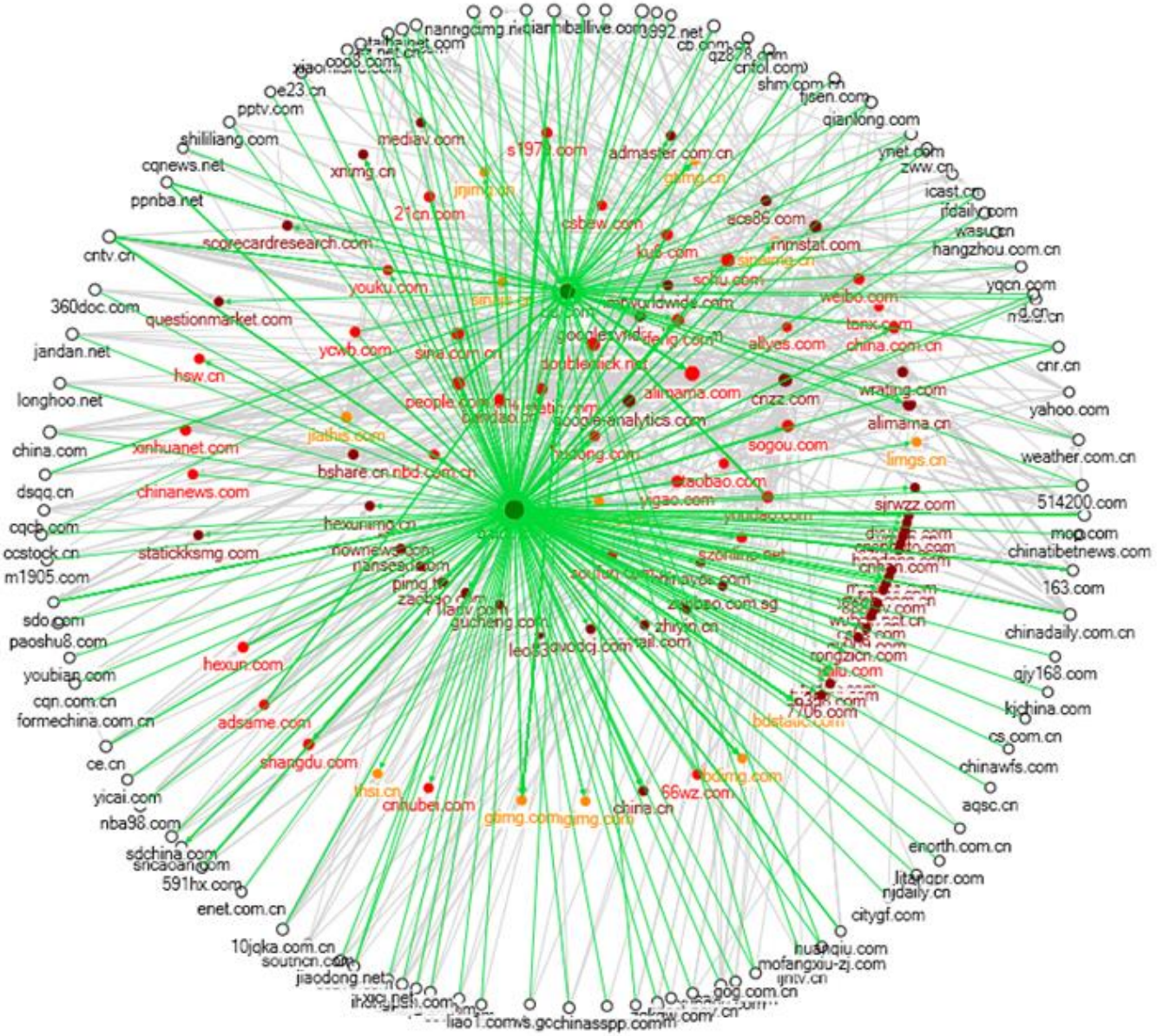
Data set – Baidu Search Queries

- 10 popular queries are published daily by Baidu
- Collected **98** popular queries
- No categories available



White circles – Web sites
Dark and light red circles – trackers and ad brokers

Tracking network uncovered through the Baidu search in China.



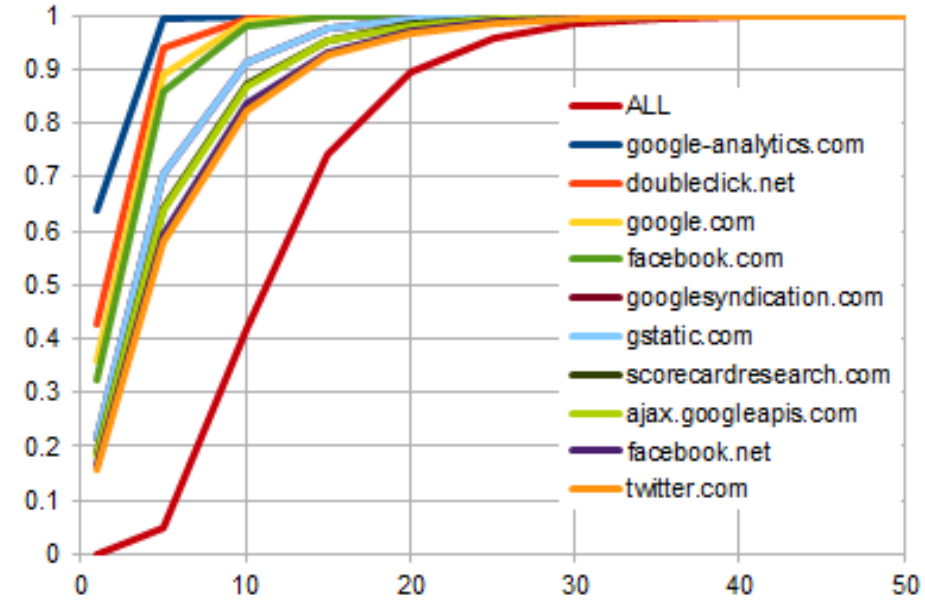
Baidu.com and qq.com cover most of the network.

Probability of Being Exposed to a Tracking Company

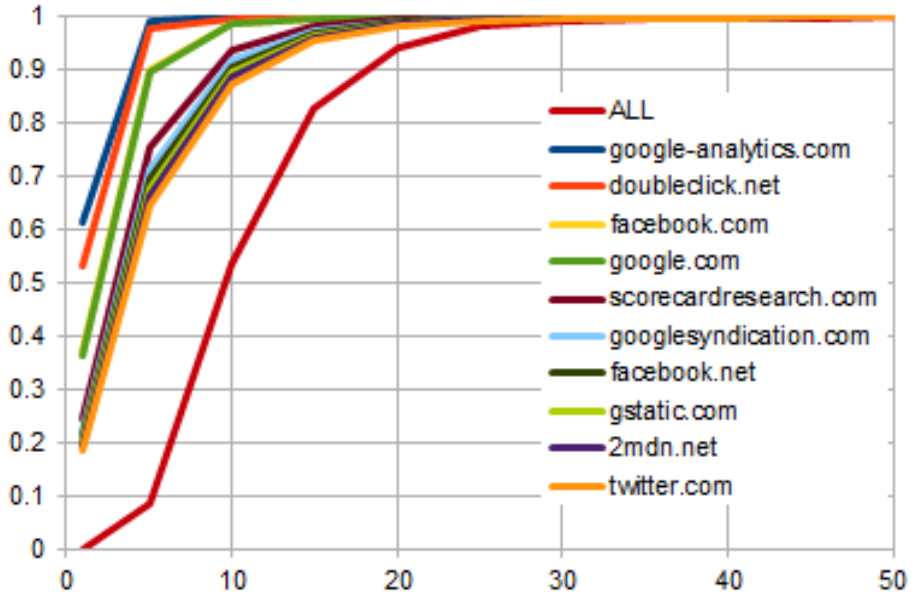
All top 10 trackers will be tracking you with:

p>90 % after clicking on 20 search results

p>99% after viewing 30 search results



Trackers in the search results of Google in the US Search market



Trackers in the search results of Bing in the US Search market

AVERAGE NUMBER OF THIRD PARTIES ASSOCIATED WITH THE FIRST SEARCH RESULT (STD. DEV.)

Label	Number of Logs	TPs w/ Cookies	TPs w/ No Cookies
Shopping\Stores & Products	785	2.79 (3.57)	3.65 (3.37)
Information\Local & Regional	726	2.16 (4.26)	3.44 (4.27)
Info\Companies & Industries	459	2.88 (4.02)	3.79 (4.28)
Living\Health & Fitness	362	2.33 (3.54)	3.42 (3.42)
Living\Car & Garage	286	3.11 (4.05)	3.84 (3.98)
Information\Law & Politics	298	0.44 (1.26)	1.23 (1.42)
Living\Travel & Vacation	301	3.10 (4.85)	3.19 (2.93)
Living\Fashion & Apparel	289	3.37 (3.87)	3.91 (3.31)
Information\Science & Tech..	271	1.77 (3.16)	2.07 (2.35)
Living\Finance & Investment	245	3.08 (3.68)	3.99 (4.28)

Spread of tracking through Twitter

Tracking network uncovered by analyzing URL sharing in Twitter

Re-defining the Social Contribution

Observations:

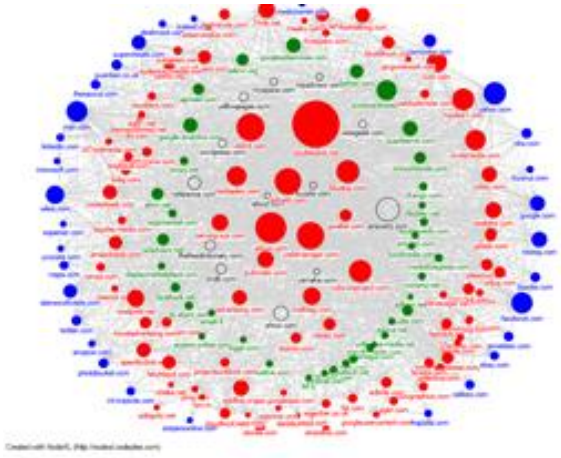
Design of specific social media encourages interaction and rewards specific behaviours.

However, the contribution needs to be assessed by considering the entire ecosystem, not the individual service alone.

Research:

- Sharing URLs in Twitter
- How to measure the social contribution when individuals' actions can affect exposure to tracking of others.

Propagation of cookies through Social Networks



Twitter DATA

hashtags.org on 05/01/2013

TOPICS
U.S. Politics
TV/Entertainment
Music
General
Business
Tech
Education
Environment
Social Change
Astrology

twitaholic.com/top100/followers/ on 05/01/2013

TWITTER USERS	
BarackObama	KimKardashian
britneyspears	ladygaga
BrunoMars	NICKIMINAJ
Cristiano	Oprah
instagram	rihanna
JLo	shakira
jtimberlake	taylorswift13
justinbieber	TheEllenShow
KAKA	twitter
katyperry	YouTube

Dataset	Total tweets	Original Tweets with URLs
TOPICS	5,364,905	499,228
TOP USERS	7,914,188	153,029

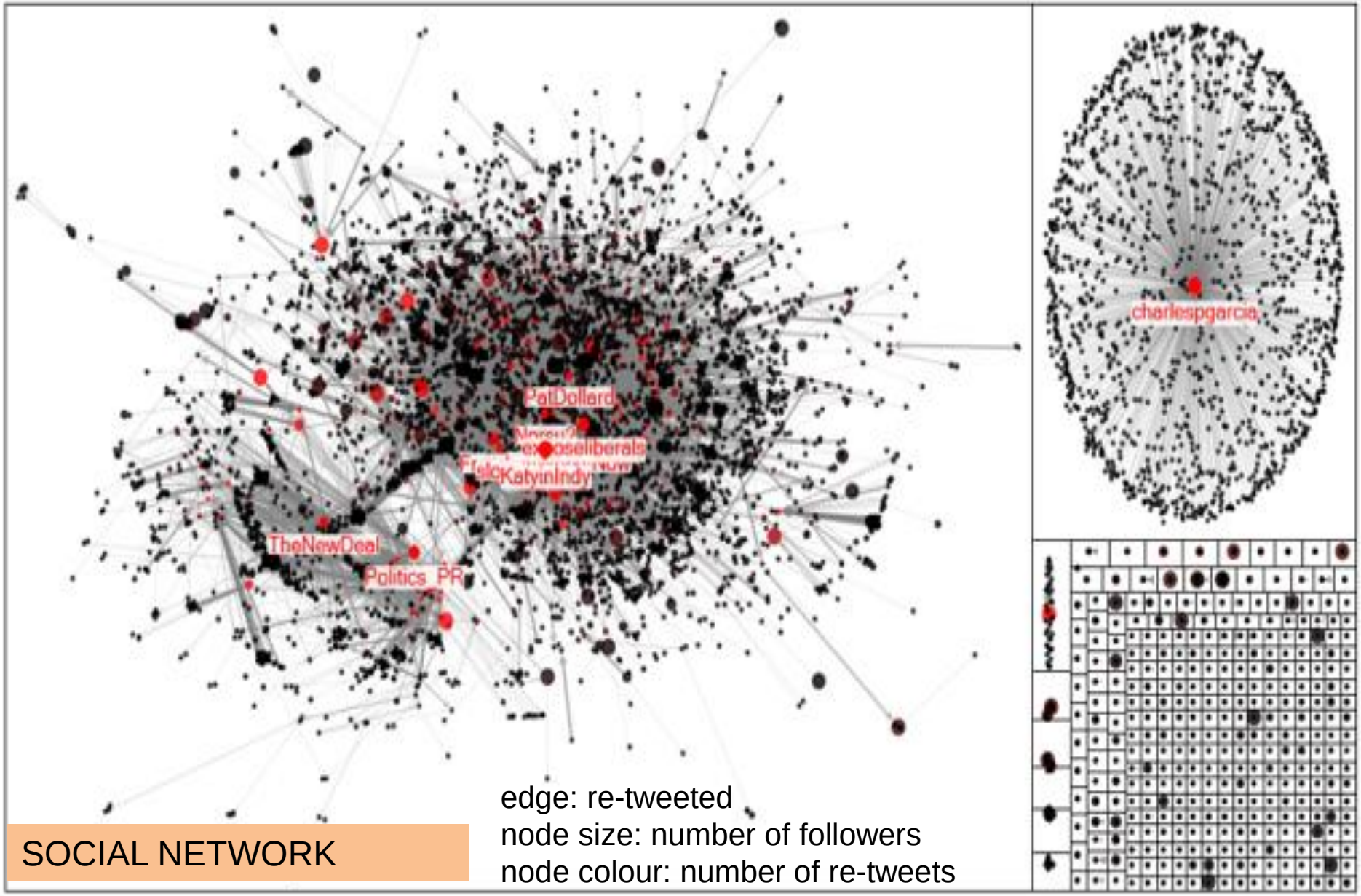
URLs Distribution in the Top User Dataset

Domain	URLs	%	# Third Parties
youtube.com	13,918	18.84%	12
tumblr.com	8,927	17.02%	246
youtu.be	5,929	10.15%	1
instagr.am	5,343	6.40%	0
peopleschoice.com	1,511	3.25%	37
twitpic.com	1,493	2.49%	17
twitlonger.com	1,442	2.41%	23
twitter.com	1,085	1.89%	27
facebook.com	896	1.76%	24
tl.gd	744	1.56%	0

Most tweeted Web domains, measured by the appearance of URLs in tweets

Network Analysis

Network	Dataset	# Nodes	# Edges
SOCIAL	TOPICS	151,624	214,327
	TOP USERS	286,389	300,697



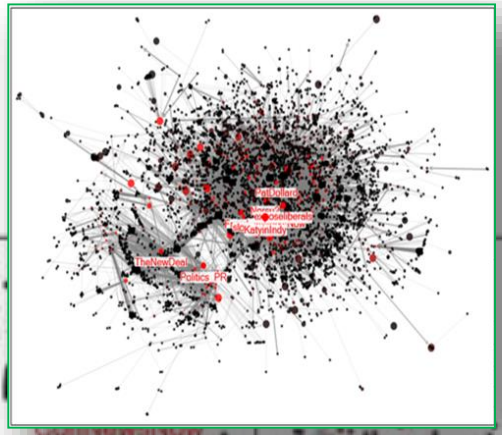
SOCIAL NETWORK

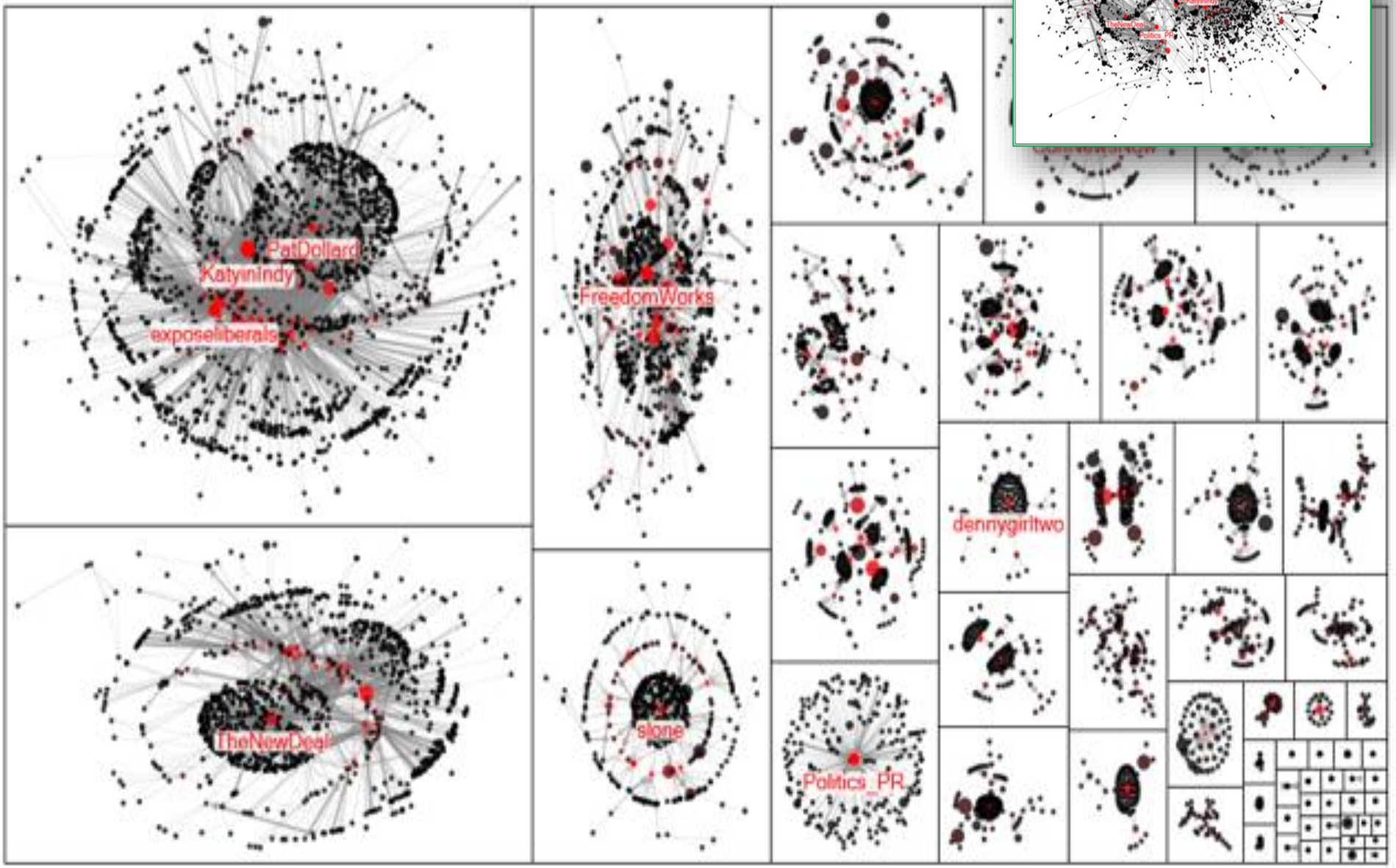
edge: re-tweeted
node size: number of followers
node colour: number of re-tweets

Connected components, showing a dominant sub-graph of the Twitter community that emerged in the Politics topic

SOCIAL NETWORK

edge: re-tweeted
node size: number of followers
node colour: number of re-tweets



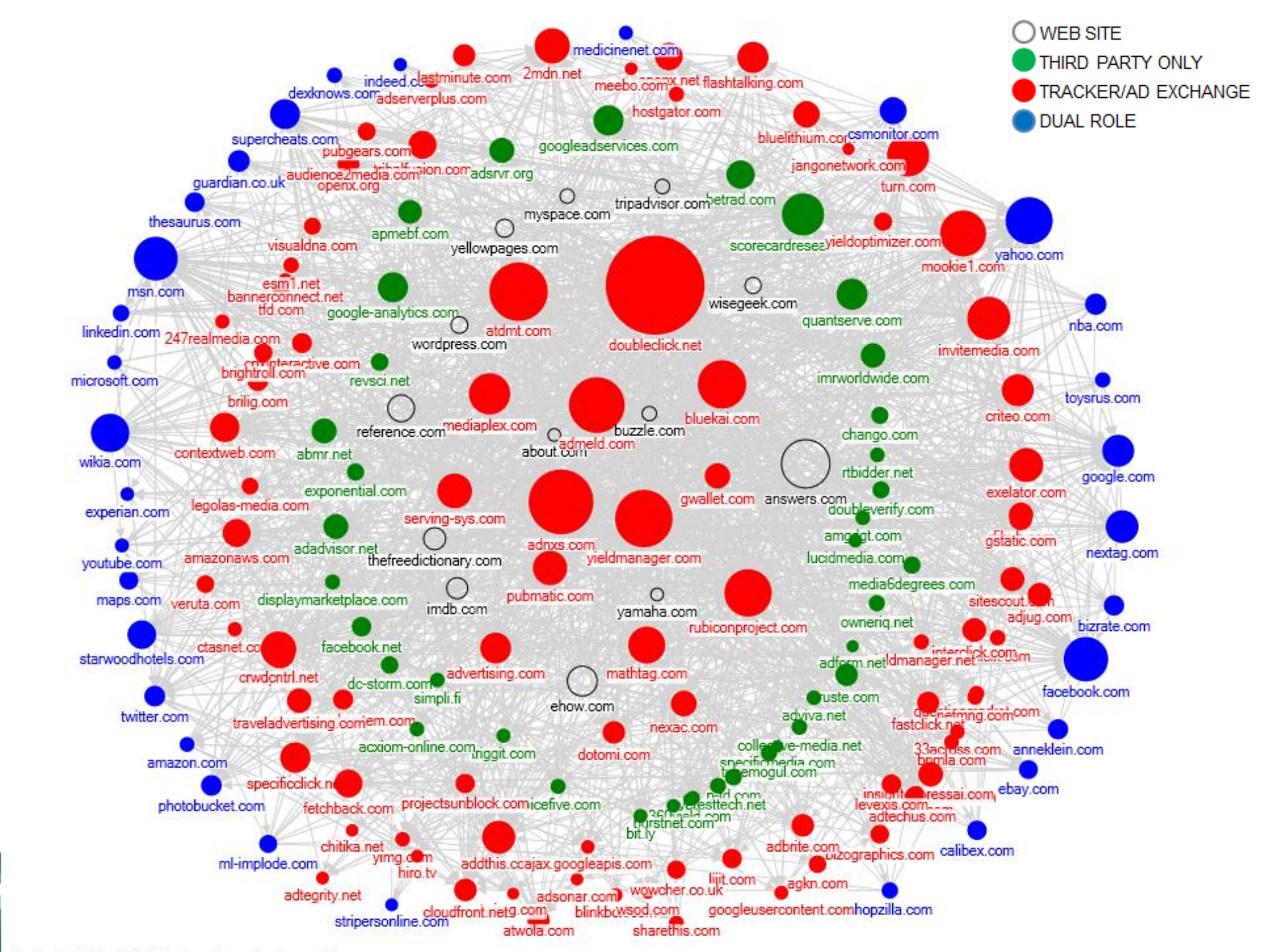


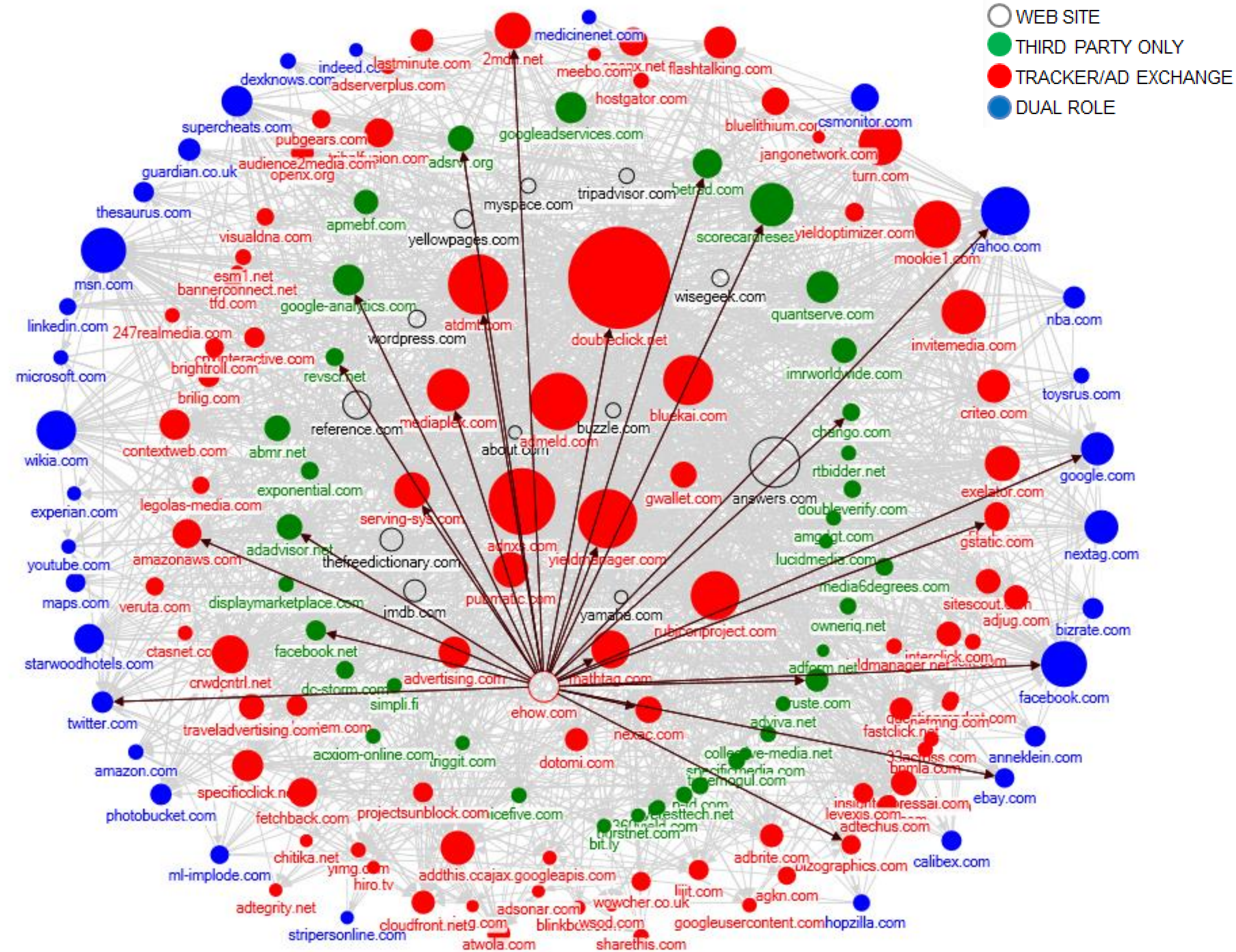
Third Party Domain	% of URLs with Third Parties	Third Party Domain	% of Users who Tweeted
google-analytics.com	65.46%	google-analytics.com	66.19%
facebook.com	56.14%	doubleclick.net	58.02%
google.com	53.59%	google.com	57.83%
twitter.com	50.24%	gstatic.com	57.08%
gstatic.com	48.58%	googlesyndication.com	51.52%
chartbeat.net	44.62%	googleadservices.com	48.92%
chartbeat.com	40.31%	facebook.com	45.18%
youtube.com	38.32%	googleusercontent.com	43.80%
doubleclick.net	38.32%	youtube.com	36.24%
facebook.net	36.17%	youtube-nocookie.com	35.86%

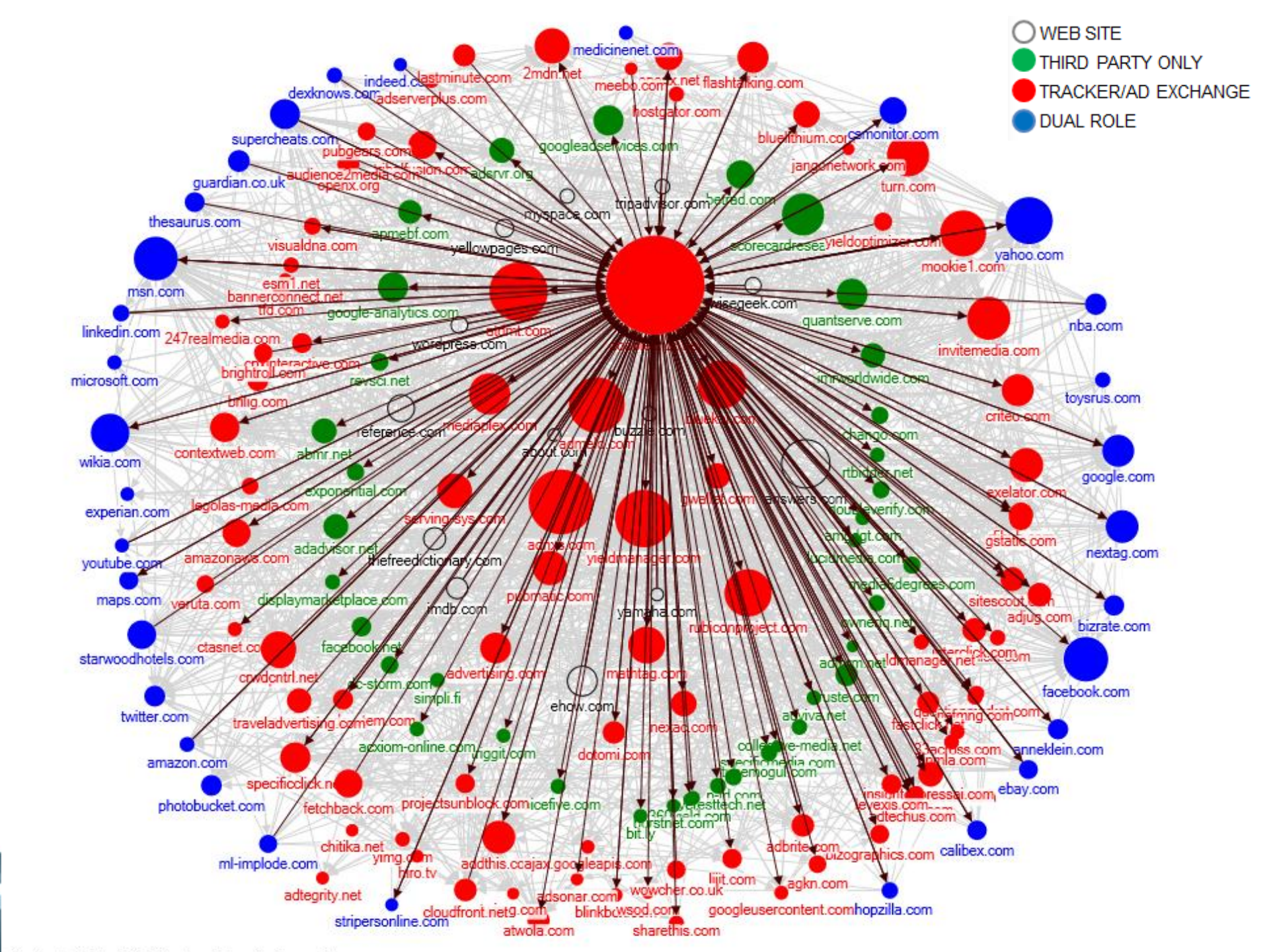
LEFT: Percentage of URLs that are associated with, i.e., 'refer to' the specific third party

RIGHT: Percentage of users that tweeted about Web sites, i.e., URLs who are associated with the specific third parties.

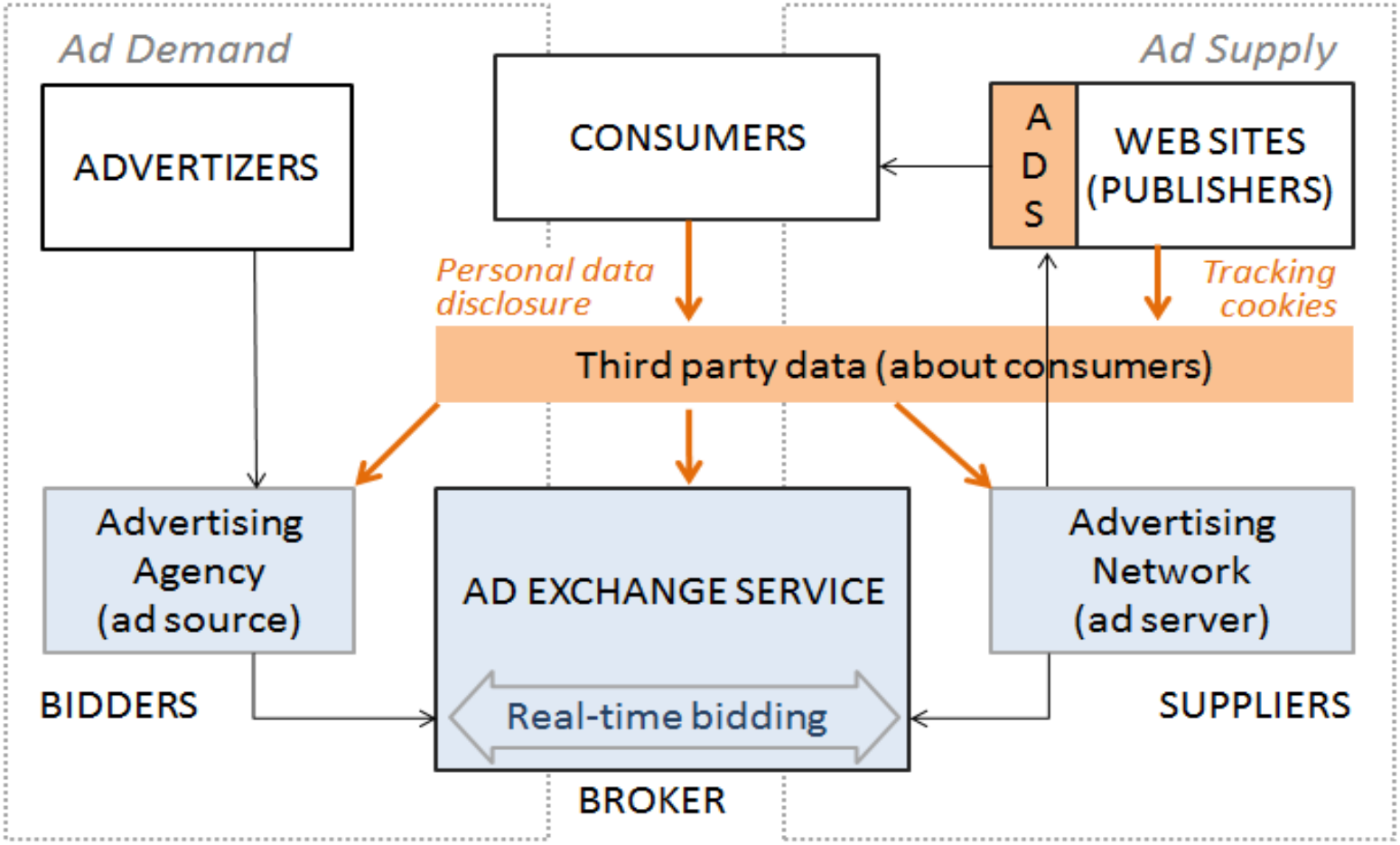
Network	Dataset	# Nodes	# Edges
SOCIAL	TOPICS	151,624	214,327
	TOP USERS	286,389	300,697
TRACKING	TOPICS	25,044	174,840
	TOP USERS	10,474	66,609


WEB SITE
THIRD PARTY ONLY
TRACKER/AD EXCHANGE
DUAL ROLE
medicinenet.com
2mdn.net
lastminute.com
indeed.com
dexknows.com
adserverplus.com
meebo.com
flashtalking.com
hostgator.com
supercheats.com
pubgears.com
googleadservices.com
bluelithium.com
csmonitor.com
jangonetwork.com
turn.com
audience2media.com
openx.org
adsrvr.org
guardian.co.uk
tripadvisor.com
betrad.com
myspace.com
thesaurus.com
apmebf.com
visualdna.com
yellowpages.com
scorecardresea
yieldoptimizer.com
yahoo.com
mookie1.com
msn.com
esm1.net
bannerconnect.net
tfd.com
google-analytics.com
atdmt.com
wisegeek.com
quantserve.com
nba.com
linkedin.com
247realmedia.com
wordpress.com
doubleclick.net
invitemedia.com
microsoft.com
brightroll.com
revsci.net
imrworldwide.com
toysrus.com
brilig.com
mediaplex.com
bluekai.com
criteo.com
reference.com
buzzle.com
admeld.com
about.com
chango.com
wikia.com
contextweb.com
abmr.net
rtbidder.net
google.com
exponential.com
gwallet.com
answers.com
exelator.com
legolas-media.com
doubleverify.com
experian.com
serving-sys.com
adnxs.com
amgdgt.com
gstatic.com
adadvisor.net
amazonaws.com
thefreedictionary.com
yieldmanager.com
lucidmedia.com
nextag.com
youtube.com
media6degrees.com
maps.com
veruta.com
displaymarketplace.com
pubmatic.com
imdb.com
yamaha.com
sitescout.com
adjug.com
rubiconproject.com
owneriq.net
bizrate.com
starwoodhotels.com
ctasnet.com
facebook.net
adform.net
interclick.com
crwdcntrl.net
dc-storm.com
advertising.com
mathtag.com
facebook.com
simpli.fi
ehow.com
truste.com
twitter.com
traveladvertising.com
nexac.com
adviva.net
fastclick.net
acxiom-online.com
triggit.com
dotomi.com
collective-media.net
33across.com
anneklein.com
amazon.com
specificmedia.com
ebay.com
specificclick.net
photobucket.com
fetchback.com
projectsunblock.com
icefive.com
levexis.com
adtechus.com
bit.ly
adbrite.com
calibex.com
chitika.net
ml-implode.com
addthis.com
ajax.googleapis.com
bizographics.com
hiro.tv
lijit.com
agkn.com
adtegrity.net
adsonar.com
wowcher.co.uk
cloudfront.net
googleusercontent.com
hopzilla.com
stripersonline.com
atwola.com
sharethis.com

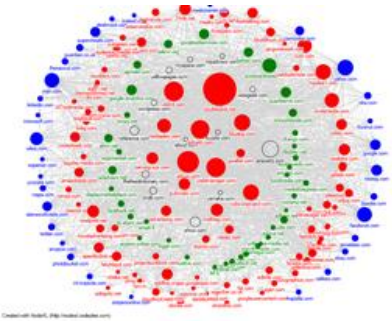

WEB SITE
THIRD PARTY ONLY
TRACKER/AD EXCHANGE
DUAL ROLE
medicinenet.com
2mdn.net
lastminute.com
adserverplus.com
dexknows.com
meebo.com
flashtalking.com
hostgator.com
googleadservices.com
supercheats.com
pubgears.com
bluelithium.com
csmonitor.com
jangonetwork.com
guardian.co.uk
audience2media.com
openx.org
turn.com
myspace.com
tripadvisor.com
thesaurus.com
apmebf.com
yellowpages.com
yieldoptimizer.com
visualdna.com
mookie1.com
yahoo.com
msn.com
esm1.net
bannerconnect.net
tfd.com
google-analytics.com
wisegeek.com
quantserve.com
linkedin.com
247realmedia.com
wordpress.com
atdmt.com
doubleclick.net
invitemedia.com
nba.com
microsoft.com
brightroll.com
revsci.net
imrworldwide.com
toysrus.com
brilig.com
reference.com
mediaplex.com
bluekai.com
criteo.com
wikia.com
contextweb.com
abmr.net
about.com
admeld.com
buzzle.com
chango.com
rtbidder.net
google.com
experian.com
exponential.com
legolas-media.com
gwallet.com
answers.com
doubleverify.com
exelator.com
serving-sys.com
adadvisor.net
amazonaws.com
thefreedictionary.com
adnxs.com
yieldmanager.com
lucidmedia.com
gstatic.com
nextag.com
youtube.com
maps.com
veruta.com
displaymarketplace.com
imdb.com
pubmatic.com
yamaha.com
media6degrees.com
sitescout.com
adjug.com
rubiconproject.com
owneriq.net
bizrate.com
starwoodhotels.com
facebook.net
advertising.com
mathtag.com
interclick.com
facebook.com
crwdcntrl.net
dc-storm.com
twitter.com
simpli.fi
ehow.com
nexac.com
adviva.net
fastclick.net
traveladvertising.com
acxiom-online.com
triggit.com
dotomi.com
anneklein.com
amazon.com
specificmedia.com
ebay.com
photobucket.com
fetchback.com
projectsunblock.com
insightexpressai.com
levexis.com
adtechus.com
calibex.com
ml-implode.com
chitika.net
bit.ly
adbrite.com
hiro.tv
addthis.com
ajax.googleapis.com
adtegrity.net
lijit.com
agkn.com
wowcher.co.uk
adsonar.com
cloudfront.net
googleusercontent.com
stripersonline.com
atwola.com
sharethis.com


WEB SITE
THIRD PARTY ONLY
TRACKER/AD EXCHANGE
DUAL ROLE
medicinenet.com
2mdn.net
lastminute.com
indeed.com
dexknows.com
adserverplus.com
meebo.com
flashtalking.com
hostgator.com
supercheats.com
pubgears.com
googleadservices.com
bluelithium.com
csmonitor.com
jangonetwork.com
turn.com
guardian.co.uk
audience2media.com
openx.org
adsrvr.org
tripadvisor.com
myspace.com
thesaurus.com
visualdna.com
yellowpages.com
scorecardresea
yieldoptimizer.com
yahoo.com
mookie1.com
msn.com
esm1.net
bannerconnect.net
google-analytics.com
wisegeek.com
quantserve.com
nba.com
linkedin.com
247realmedia.com
wordpress.com
invitemedia.com
microsoft.com
brightroll.com
revsci.net
imrworldwide.com
toysrus.com
bnlig.com
criteo.com
wikia.com
contextweb.com
abmr.net
exponential.com
google.com
experian.com
legolas-media.com
exelator.com
youtube.com
amazonaws.com
adadvisor.net
gstatic.com
nextag.com
maps.com
veruta.com
displaymarketplace.com
imdb.com
yieldmanager.com
bizrate.com
starwoodhotels.com
ctasnet.com
facebook.net
rubiconproject.com
adjug.com
twitter.com
crwdcntrl.net
advertising.com
mathtag.com
facebook.com
simpli.fi
ehow.com
amazon.com
dotomi.com
nexac.com
anneklein.com
photobucket.com
specificclick.net
fetchback.com
projectsunblock.com
ebay.com
ml-implode.com
chitika.net
addthis.com
adtechus.com
calibex.com
adtegrity.net
stripersonline.com
cloudfront.net
atwola.com
sharethis.com
googleusercontent.com
hopzilla.com
agkn.com
adsonar.com
wowcher.co.uk
lijit.com
adbrite.com
bizographics.com

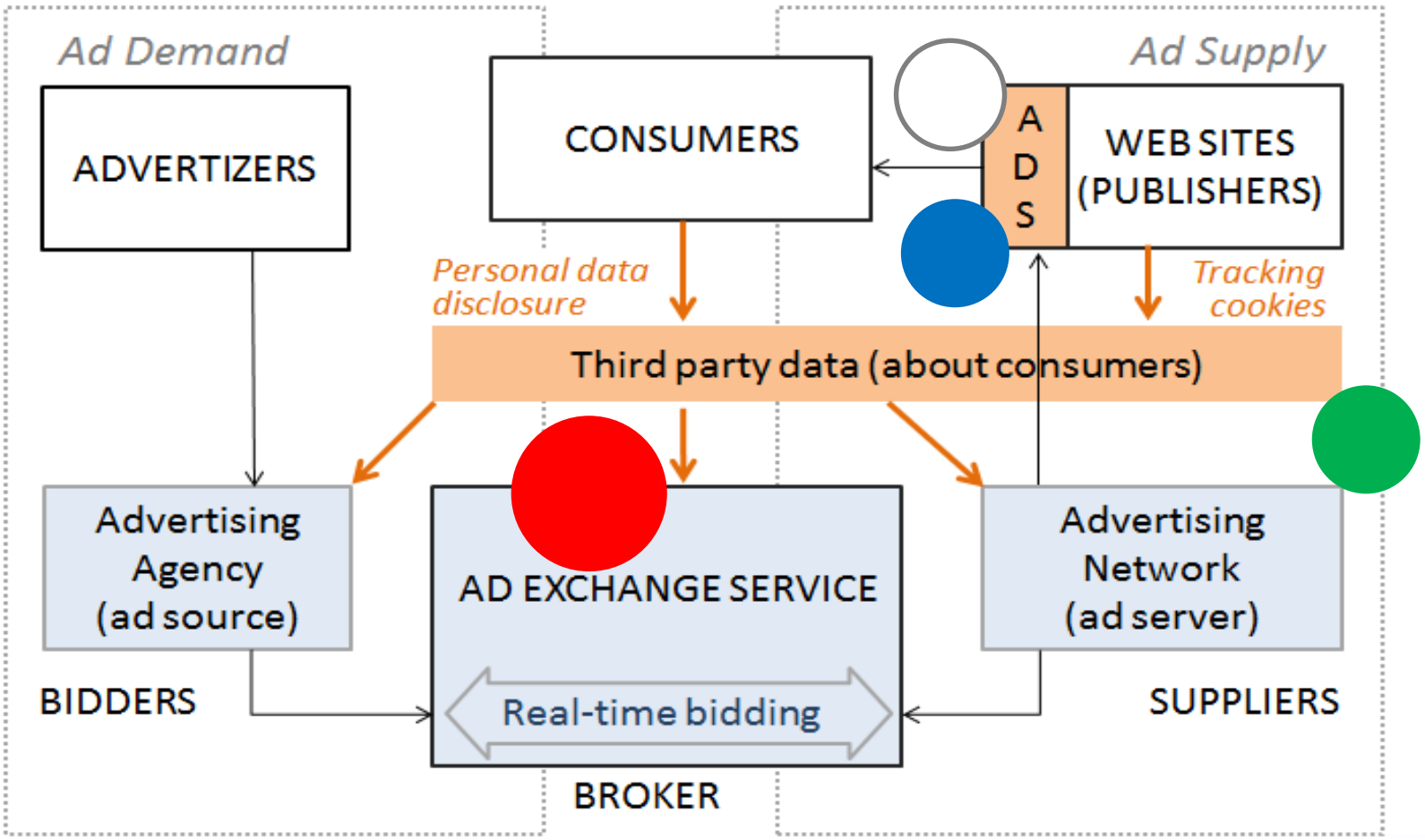
Business Ecosystem



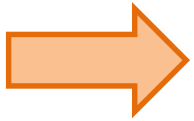
Business Ecosystem



WEB SITE
THIRD PARTY ONLY
TRACKER/AD EXCHANGE
DUAL ROLE



new slogan

WYSIWYG  WY’^{n}SI **WYP**

***W**hat **Y**ou **D**on’t **S**ee is **W**hat **Y**ou **P**ay*

- Interaction with applications and services through UI is captured and forms a digital footprint that is used for (behavioural) profiling
- Protocols used to enable communication between the PC and Web services enable device fingerprinting and user tracking.

research investigation

Designs of computing systems lack transparency about the personal data capture and data flow.

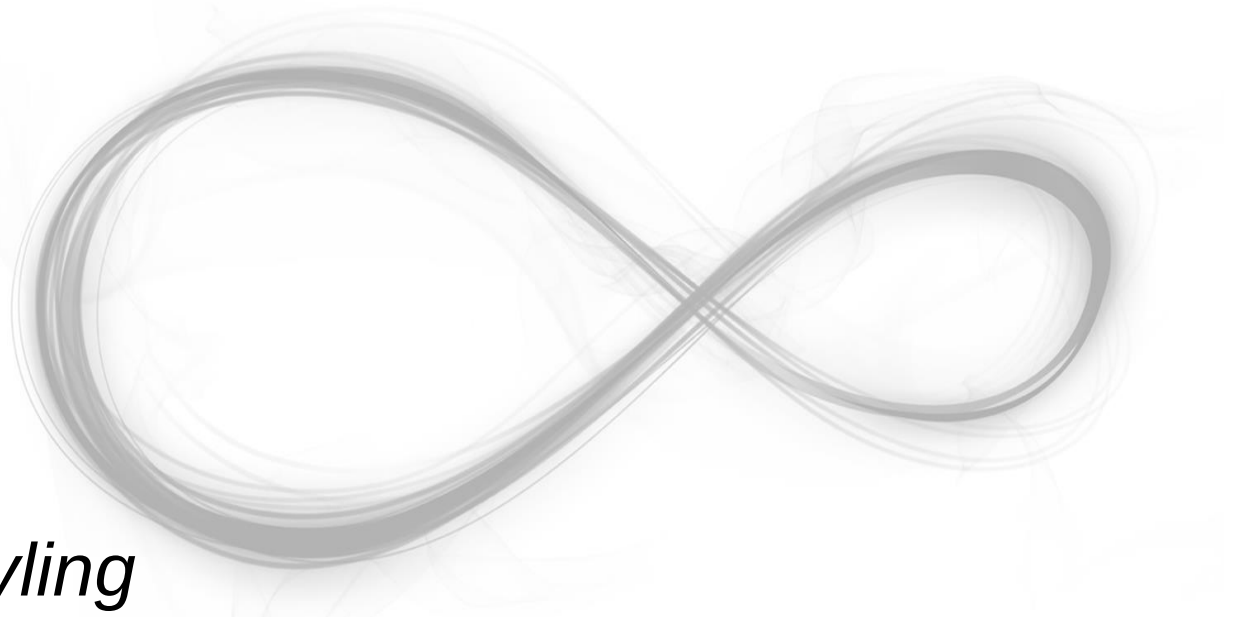
Implications:

We are unable to make informed decisions and assume responsibility for our own actions, and ensure we do not harm others.

We have been stripped of the ability to determine our own self within the (digital) society.

Information is collected by first and third parties. We have no say and no control over what is collected, to whom and how information is presented, and how it is used.

Thank you



Natasa Milic-Frayling

natasamf@microsoft.com