Virtual silos on the web

Exploring website clusters using hyperlinks and tracking code

Jelmer Neeven

Martin Lopatka (Mozilla)

Maarten Marx (UvA)

Hosein Azarbonyad (UvA PhD)

A collaboration with Mozilla

June 25, 2017

Introduction

There is an increasing trend of recording, analyzing, and even selling users information and browsing behavior by third-party websites to companies.

Hypothesis: relationship between structural organisation of the Web and advertisement revenue model.

Goals

1. Detect trackers using the embedded scripts and placed cookies for roughly 5 million webpages.
2. Build a communication graph of webpages, trackers, and stakeholders.
3. Use community detection algorithms to find virtual silos on the web.
4. Analyse and compare the silos to gain insight into hypothesised relationship.

Seed URLs of pre-existing communities

- Need a starting point to get a list of websites.
- Using different communities, maximum diversity.
- Do the silos correspond to the communities?
- URLs gathered from `reddit.com`.

Crawling the pages

- Get domain of URL.
- Get all third-party scripts and cookies.
- Count internal/external links.
- *robots.txt* permissions.

Crawler results

- Data from roughly 355,000 URLs.
- 83,657 domains.
- jQuery not in top 50.

Crawler results

A few of the most frequently used script sources:

- `google-analytics.com/analytics.js`
- `platform.twitter.com/widgets.js`
- `securepubads.g.doubleclick.net/gampad/ads`
- `b.scorecardresearch.com/beacon.js`

Communication graphs

Three graphs:

- Hyperlink
- Tracking
- Total

Detect communities with Louvain Modularity Optimisation algorithm.

Communication graphs

Graph:	Communities	Modularity
Hyperlink	671	0.7
Tracking	4169	0.62
Total	435	0.64

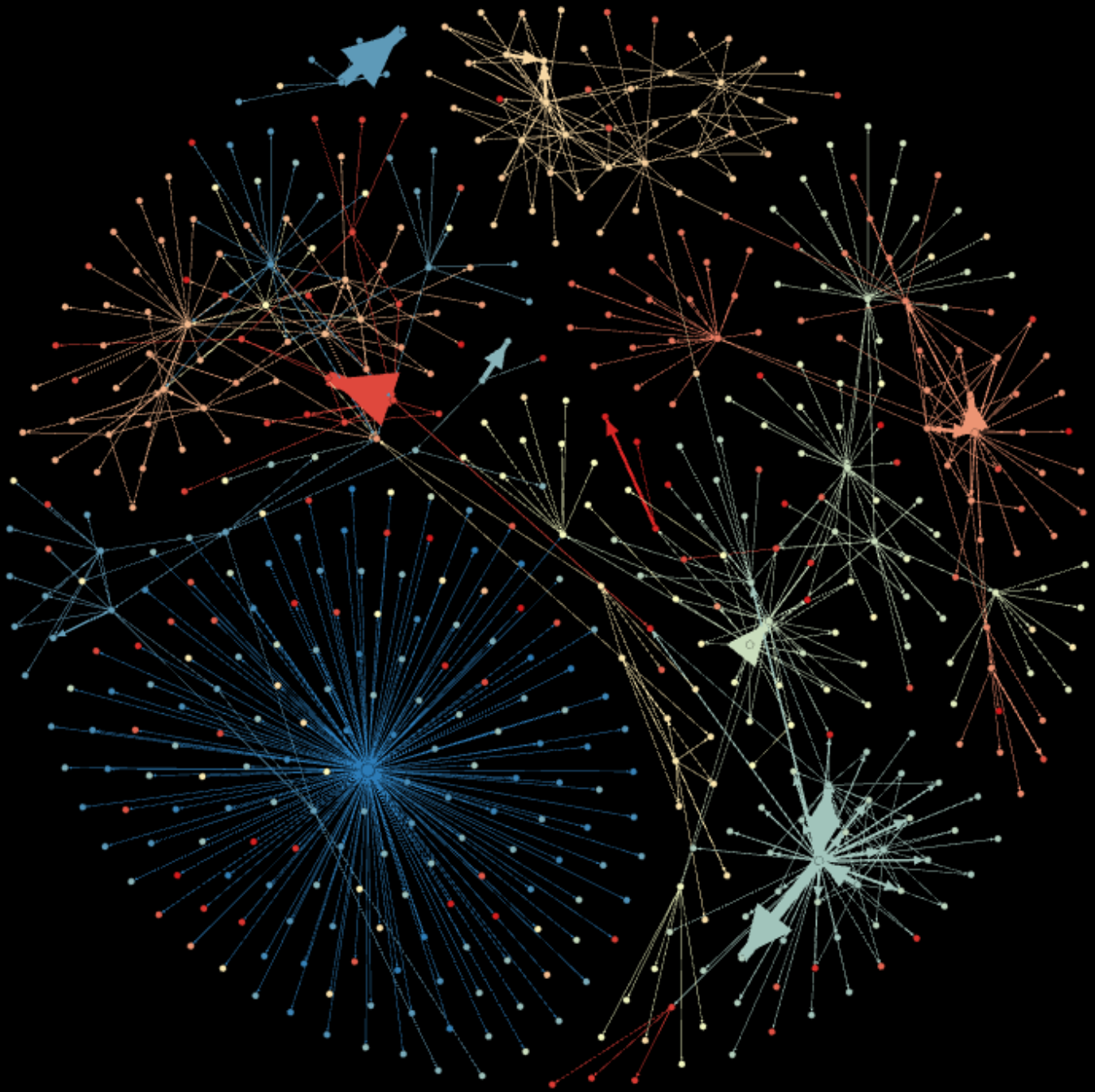
Communication graphs

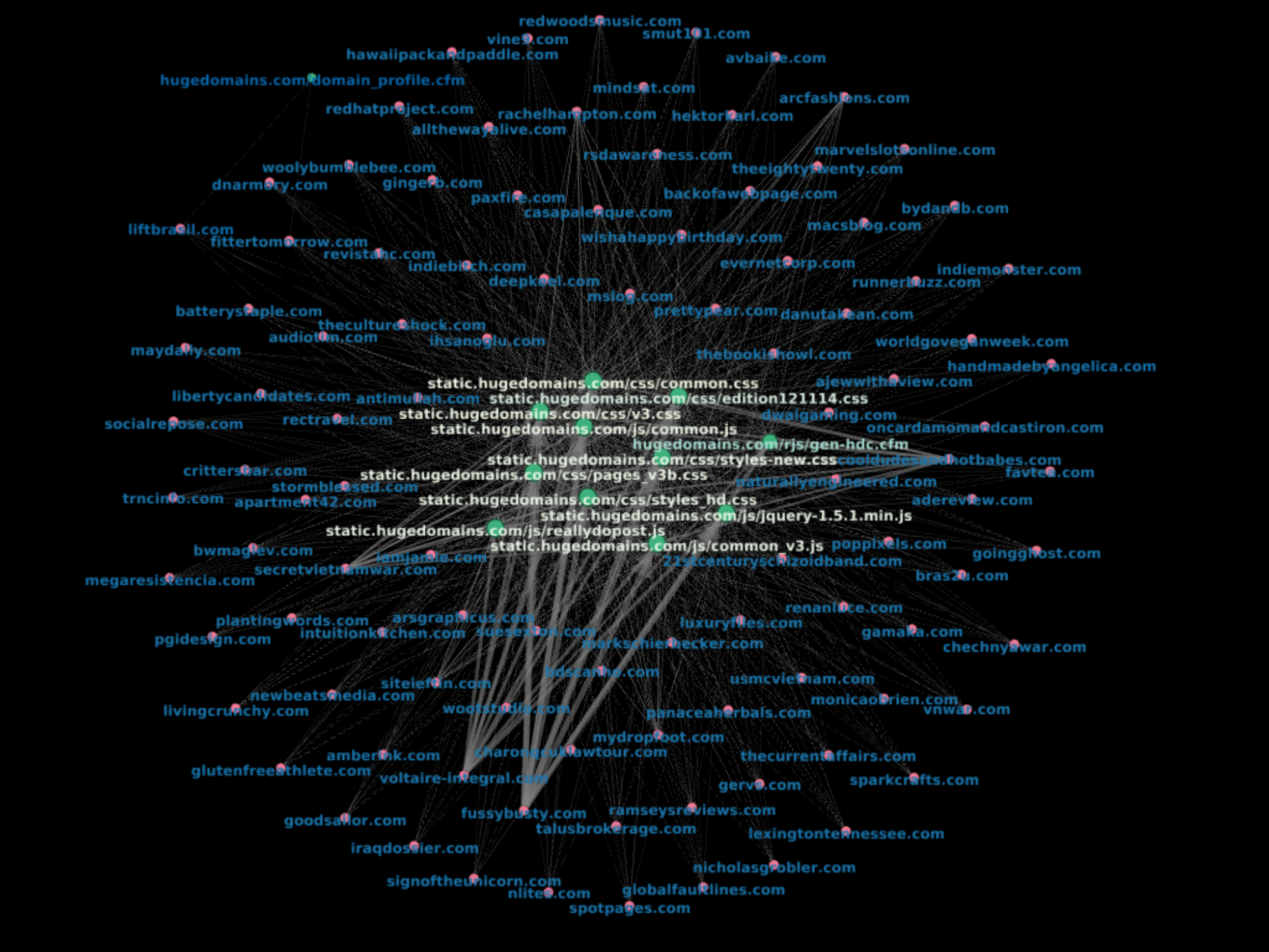
Calculate metrics of communities:

- size
- density
- degree
- connectivity

Calculated on induced graph so they can be compared meaningfully.


gainsboroughstandard.co.uk
haylingtoday.co.uk
sheffieldtelegraph.co.uk
derryjournal.com
sussexexpress.co.uk
thametoday.co.uk
crawleyobserver.co.uk
batleynews.co.uk
spenboroughguardian.co.uk
todmordennews.co.uk
halifaxcourier.co.uk
midsussextimes.co.uk
lurganmail.co.uk
horncastlenews.co.uk
dewsburyreporter.co.uk
newsguardian.co.uk
lutontoday.co.uk
pendletoday.co.uk
wscountytimes.co.uk
yorkshirepost.co.uk
ballymoneytimes.co.uk
buxtonadvertiser.co.uk
retfordtoday.co.uk
bognor.co.uk
matlockmercury.co.uk
longridgenews.co.uk
littlehamptongazette.co.uk
londonderrysentinel.co.uk
westsussextoday.co.uk
fileymercury.co.uk
hebdenbridgetimes.co.uk
dromoreleader.co.uk
chorley-guardian.co.uk
sthelensreporter.co.uk
ryeandbattleobserver.co.uk
leighreporter.co.uk
shieldsgazette.com
warwickcourier.co.uk
biggleswadetoday.co.uk
carrickfergustimes.co.uk
wigantoday.net
rugbyadvertiser.co.uk
ripleyandheanornews.co.uk
jplocalbusiness.co.uk
thevisitor.co.uk
pontefractandcastlefordexpress.co.uk
inews.co.uk
brightonandhoveindependent.co.uk
bedfordtoday.co.uk
hemeltoday.co.uk
wetherbynews.co.uk
whitbygazette.co.uk
newtownabbeytoday.co.uk
johnstonpress.co.uk
chichester.co.uk
northumberlandgazette.co.uk
morleyobserver.co.uk
mirfieldreporter.co.uk
southyorkshiretimes.co.uk
larnetimes.co.uk
banbridgeleader.co.uk
colerainetimes.co.uk
eastbourneherald.co.uk
lancasterguardian.co.uk
belpernews.co.uk
wakefieldexpress.co.uk
sunderlandecho.com
ilkestonadvertiser.co.uk
tringtoday.co.uk
northyorkshirenews.com
hucknalldispatch.co.uk
burnleyexpress.net
skegnessstandard.co.uk
investorrelations.co.uk
doncasterfreepress.co.uk
pocklingtonpost.co.uk
ballymenatimes.com
hemsworthandsouthelmsallexpress.co.uk
harboroughmail.co.uk
kenilworthweeklynews.co.uk
marketrasenmail.co.uk
sleafordstandard.co.uk
derbyshiretimes.co.uk
leylandguardian.co.uk
eastwoodadvertiser.co.uk
worksopguardian.co.uk
newsletter.co.uk
daventryexpress.co.uk
hastingsobserver.co.uk
buckinghamtoday.co.uk
portadowntimes.co.uk
leightonbuzzardonline.co.uk
banburyguardian.co.uk
midulstermail.co.uk
chad.co.uk
wow247.co.uk

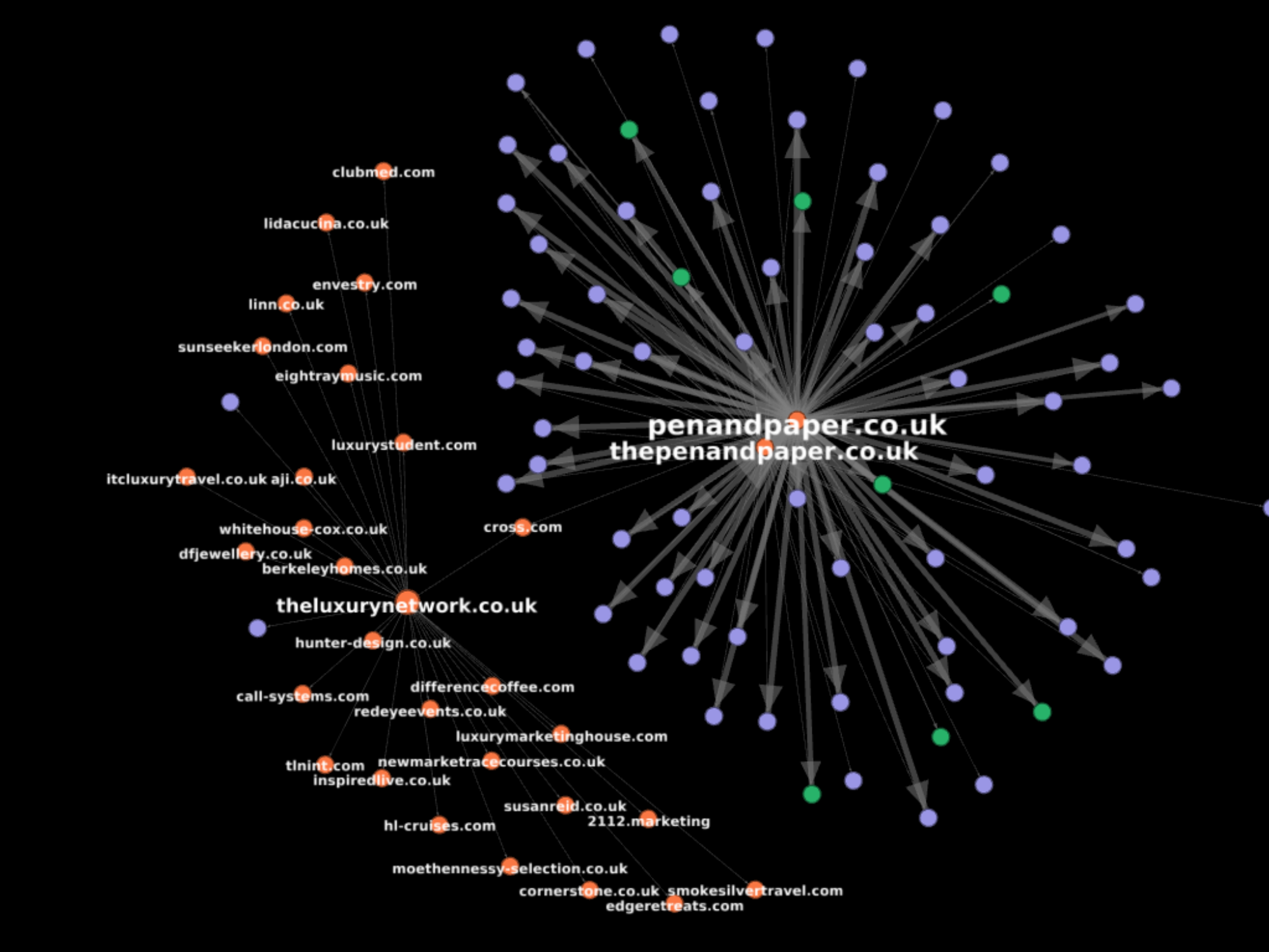



redwoodsmusic.com
smut101.com
vine9.com
hawaiipackandpaddle.com
avbaise.com
hugedomains.com/domain_profile.cfm
mindset.com
arcfashions.com
redhatproject.com
rachelhampton.com
hektorkarl.com
allthewayalive.com
rsdawareness.com
marvelslotsonline.com
woolybumblebee.com
theeightytwenty.com
dnarmory.com
gingerb.com
paxfire.com
backofawebpage.com
casapalenque.com
bydandb.com
liftbrasil.com
macsblog.com
fittertomorrow.com
wishahappybirthday.com
revistahc.com
indiebitch.com
evernetcorp.com
indiemonster.com
deepkeel.com
runnerbuzz.com
mslog.com
batterystaple.com
prettybear.com
danutakean.com
theculturehack.com
audiotom.com
ihsanoglu.com
worldgoveganweek.com
maydaily.com
thebookishowl.com
handmadebyangelica.com
static.hugedomains.com/css/common.css
ajewwithaview.com
libertycandidates.com
antimuslah.com
static.hugedomains.com/css/edition121114.css
static.hugedomains.com/css/v3.css
dwaigaming.com
socialrepose.com
rectravel.com
static.hugedomains.com/js/common.js
oncardamomandcastiron.com
hugedomains.com/rjs/gen-hdc.cfm
static.hugedomains.com/css/styles-new.css
cooldudesandhotbabes.com
critterspar.com
static.hugedomains.com/css/pages_v3b.css
favtee.com
stormblessed.com
naturallyengineered.com
trncindo.com
apartment42.com
static.hugedomains.com/css/styles_hd.css
aderevtew.com
static.hugedomains.com/js/jquery-1.5.1.min.js
static.hugedomains.com/js/reallydopost.js
static.hugedomains.com/js/common_v3.js
poppixels.com
bwmaglev.com
iamjamie.com
goingghost.com
21stcenturyschizoidband.com
secretvietnamwar.com
megaresistencia.com
bras2u.com
renanlace.com
plantingwords.com
arsgraphicus.com
luxuryfiles.com
intuitionkitchen.com
suesexton.com
gamaka.com
pgidesign.com
markschieldecker.com
chechnyawar.com
bdscan.com
siteiefan.com
usmcvietnam.com
newbeatsmedia.com
monicaobrien.com
livingcrunchy.com
woofstudio.com
panaceaherbals.com
vnwar.com
mydropfoot.com
amberink.com
charonnculawtour.com
thecurrentaffairs.com
glutenfreeathlete.com
voltaire-integral.com
sparkcrafts.com
gerve.com
goodsailor.com
fussybusty.com
ramseysreviews.com
talusbrokerage.com
lexingtontennessee.com
iraqdossier.com
nicholasgrobler.com
signoftheunicorn.com
nlites.com
globalfaultlines.com
spotpages.com

Total silos

Manually analysed twenty silos:

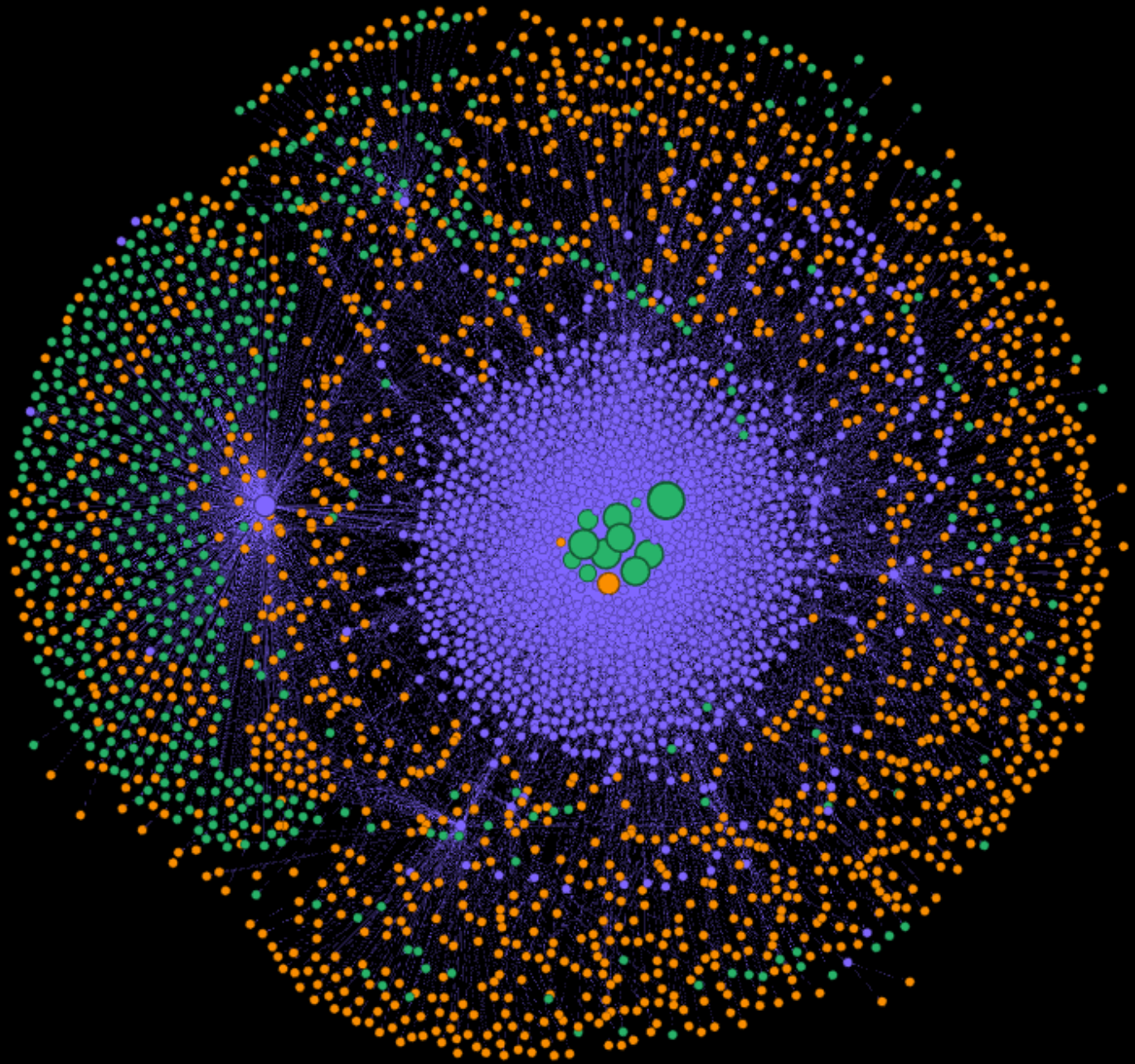
- Twelve contain exactly the same domains.
- Six silos were (part of) a bigger silo, now separate because of cookies/scripts. However, can also be identified by Louvain.
- One lost a domain because it uses popular cookies/scripts.
- The following silo only exists within the total communities, which is remarkable:

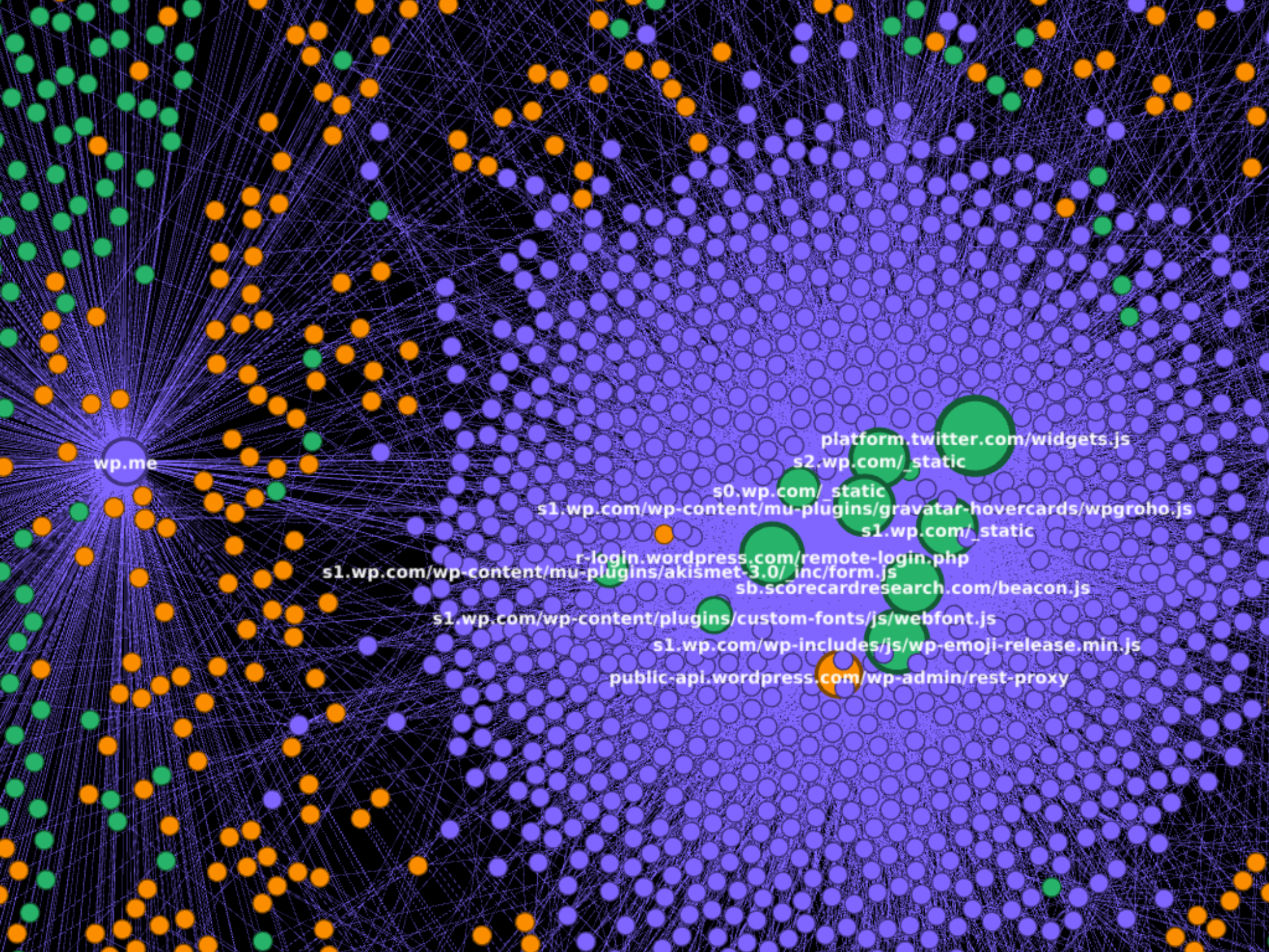

clubmed.com
lidacucina.co.uk
envestry.com
linn.co.uk
sunseekerlondon.com
eightraymusic.com
luxurystudent.com
itcluxurytravel.co.uk aji.co.uk
whitehouse-cox.co.uk
dfjewellery.co.uk
berkeleyhomes.co.uk
cross.com
theluxurynetwork.co.uk
hunter-design.co.uk
call-systems.com
differencecoffee.com
redeyeevents.co.uk
luxurymarketinghouse.com
tlnint.com
newmarketracecourses.co.uk
inspiredlive.co.uk
susanreid.co.uk
2112.marketing
hl-cruises.com
moethennessy-selection.co.uk
cornerstone.co.uk
smokesilvertravel.com
edgeretreats.com
penandpaper.co.uk
thepenandpaper.co.uk

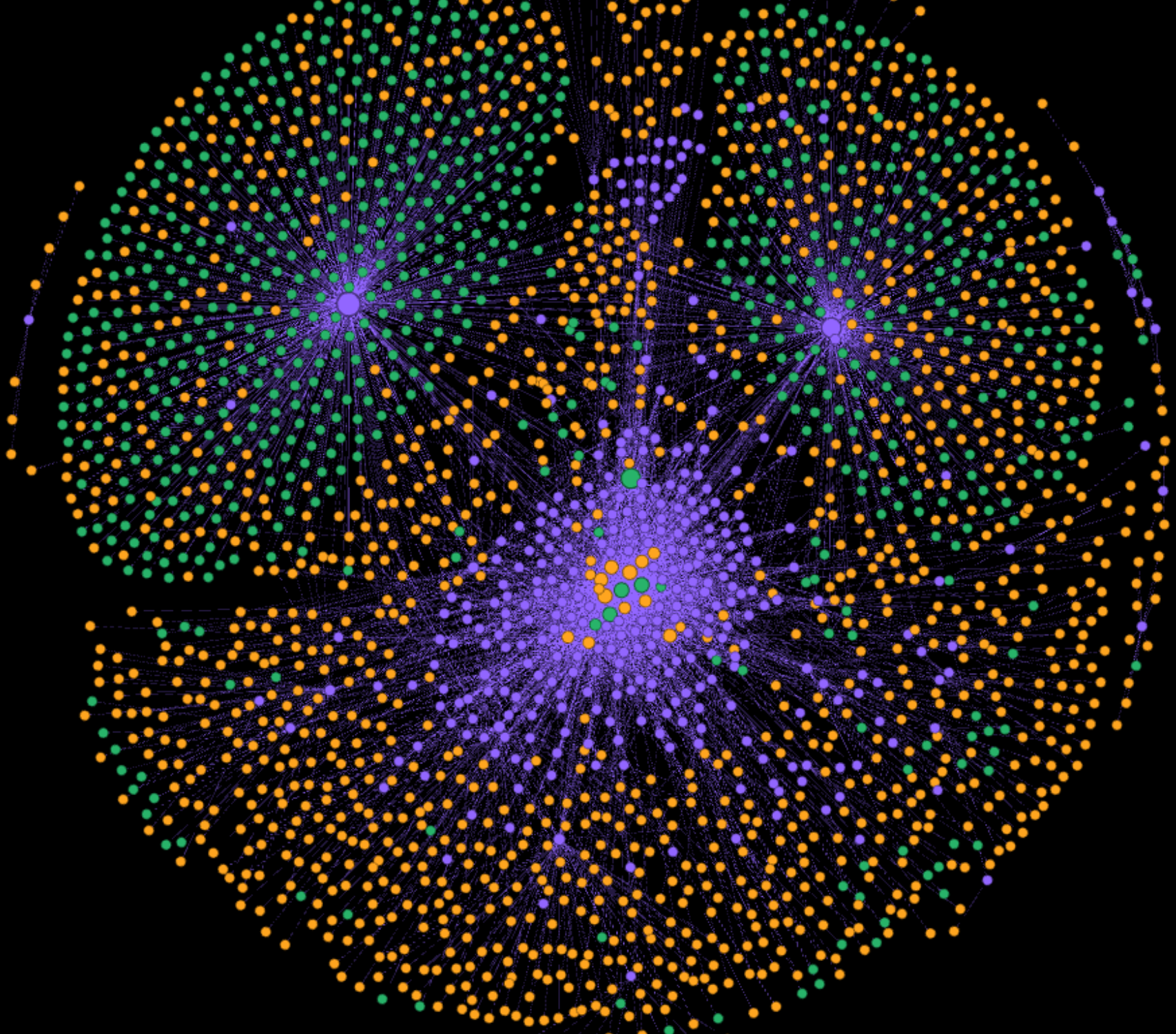
Preliminary conclusion

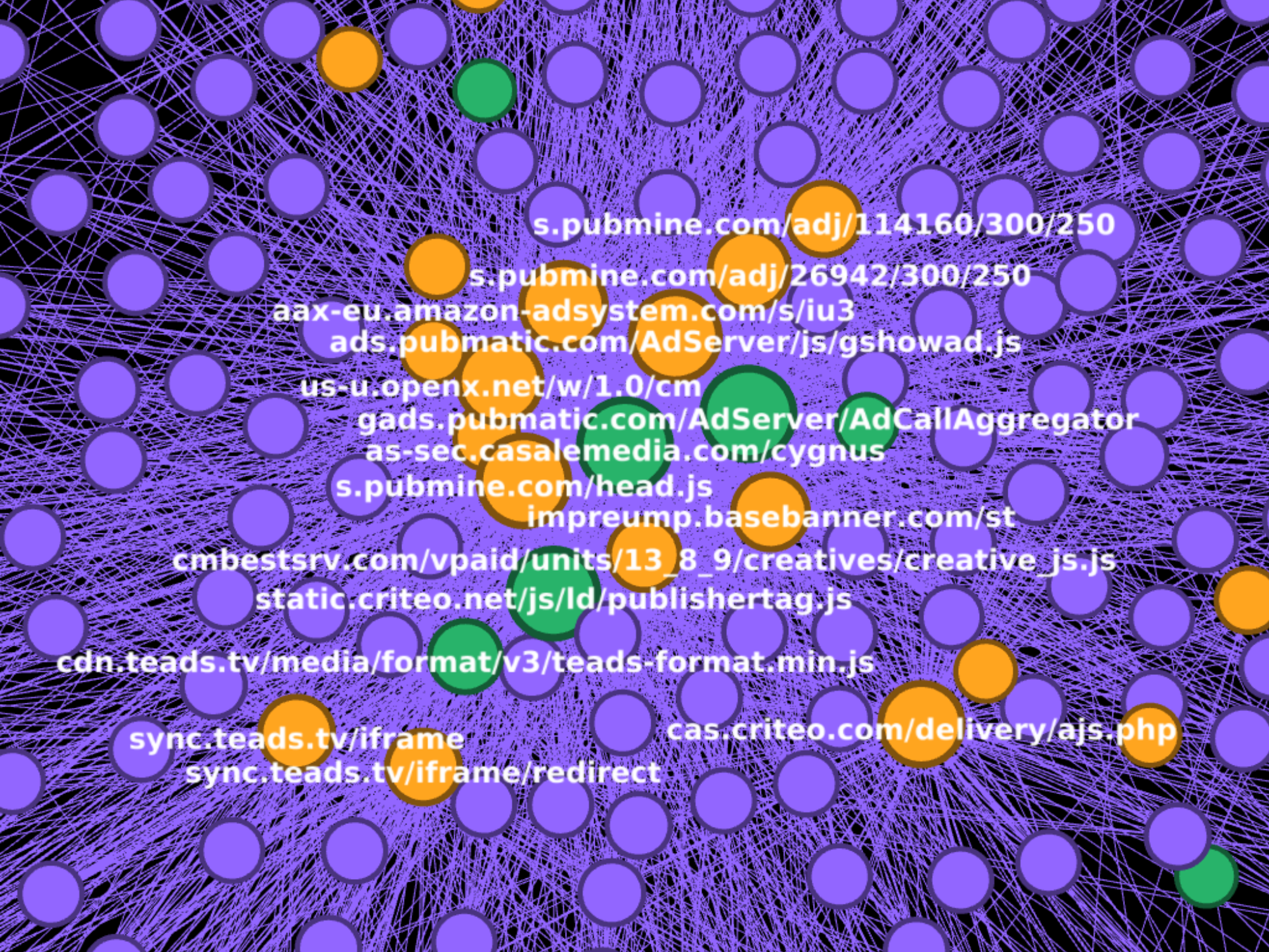
- Incorporating tracking data in communication graph does not yield many interesting results.
- In fact, sometimes sites are clustered together that are not linked in any way.
- Reduced connectivity, density, and average degree.

However, the bigger silos with a higher degree have not yet been considered.




wp.me
platform.twitter.com/widgets.js
s2.wp.com/_static
s0.wp.com/_static
s1.wp.com/wp-content/mu-plugins/gravatar-hovercards/wpgroho.js
s1.wp.com/_static
r-login.wordpress.com/remote-login.php
s1.wp.com/wp-content/mu-plugins/akismet-3.0/_inc/form.js
sb.scorecardresearch.com/beacon.js
s1.wp.com/wp-content/plugins/custom-fonts/js/webfont.js
s1.wp.com/wp-includes/js/wp-emoji-release.min.js
public-api.wordpress.com/wp-admin/rest-proxy




s.pubmine.com/adj/114160/300/250
s.pubmine.com/adj/26942/300/250
aax-eu.amazon-adsystem.com/s/iu3
ads.pubmatic.com/AdServer/js/gshowad.js
us-u.openx.net/w/1.0/cm
gads.pubmatic.com/AdServer/AdCallAggregator
as-sec.casalemedia.com/cygnus
s.pubmine.com/head.js
impreump.basebanner.com/st
cmbestsrv.com/vpaid/units/13_8_9/creatives/creative_js.js
static.criteo.net/js/ld/publishertag.js
cdn.teads.tv/media/format/v3/teads-format.min.js
sync.teads.tv/iframe
cas.criteo.com/delivery/ajs.php
sync.teads.tv/iframe/redirect

Conclusion

- Although very labour-intensive, it is indeed possible to detect tracking networks using this approach.
- The tracking networks seem to be 'hidden' within larger networks.
- In general, it appears that the detected communities can be further split up.

Future research

- Automated approach: recursive Louvain
- Tracking probabilities
- Different community detection algorithms

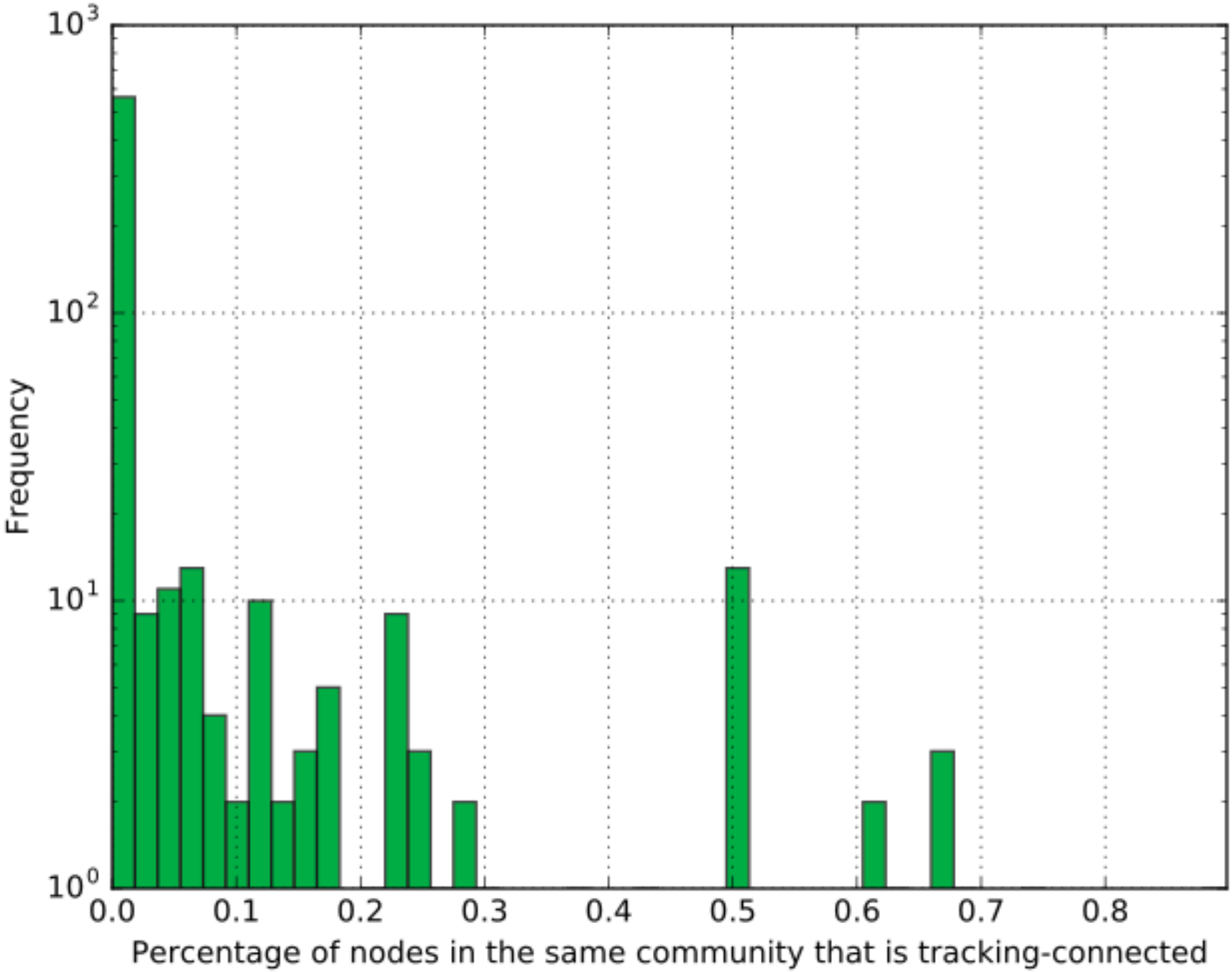
End of presentation

Thank you for your attention!

Appendix

Advertisement network:
- Inner edges: 66972
- Outer edges: 112843

Hyperlink connected:
- Total hyperlinks found: 827724
- Amount of connected domains sharing at least one piece of tracking code: 84234


10^3
10^2
10^1
10^0
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Percentage of nodes in the same community that is tracking-connected