ANÁLISIS DE MONITORIZACIÓN DE PÁGINAS WEB

GRUPO 5

Laura Sanz García

Guillermo Monserrate Sánchez

Blanca de la Torre Fuertes

Pablo Verdugo Garrido

ÍNDICE

1. Obtención de datos

1.1. Fuente de obtención

Lightbeam. Creación y uso

Para la obtención de datos de este proyecto se ha utilizado Lightbeam, un complemento para Firefox **que brinda un control visual de nuestro historial de navegación** y analiza quién está recopilando nuestros datos y cómo se produce el seguimiento de nuestra actividad en Internet.

La versión inicial de esta extensión fue presentada por el CEO de Mozilla, Gary Kovacs, a inicios de 2012 bajo el nombre de Collusion en una charla TED. En ella explicaba cómo, gracias a este complemento, la ciudadanía podría correr la cortina del desconocimiento hacia los roles de las third parties, cómo los datos controlan la mayoría de las experiencias Web y el poco control que tenemos sobre dicha experiencia y la pérdida de nuestros datos.

Charla TED : https://www.ted.com/talks/gary_kovacs_tracking_the_trackers/transcript

Con este trabajo, nuestro equipo pretende continuar de manera humilde la labor definida por todo el equipo de Mozilla y poder concienciar a la gente que nos rodea sobre el fenómeno del rastreo de nuestros datos en la sombra.

El funcionamiento de esta herramienta se centra en la creación de un registro de eventos para los sitios visitados y los de terceros que estén activos en esas páginas y almacenados localmente en nuestro ordenador. Con ello, Lightbeam muestra gráficos visuales a tiempo real de estos eventos para resaltar las interacciones; además de continuar añadiendo información más detallada sobre dichas relaciones mientras se navega por la Web.

Adicionalmente, se puede encontrar la documentación y código de Lightbeam en Github, donde cualquiera puede informarse de los formatos de archivos que utilizan y trastear por el código y su estructura: https://github.com/mozilla/lightbeam

Fuentes y obtención de datos

Para poder realizar un análisis más amplio del campo a estudiar (monitorización de páginas web, dirigido a búsquedas realizadas por estudiantes) se han definido 2 fuentes primordiales de información, desde las que se extraen datos de sus diferentes navegaciones:

- **Navegación realizada por Asociaciones de Estudiantes**

Estos datos se han recopilado desde la extensión de Lightbeam instalada en los ordenadores del despacho de la asociación LibreLab (con permiso de los responsables de Junta). Con esta fuente se ha querido extender el abanico y variedad de las búsquedas realizadas por estudiantes, ya que entendemos que podría haber quedado limitado al haber sido solo de los miembros del equipo.

- **Navegación dirigida a páginas populares**

La recopilación de estos datos se ha realizado desde el ordenador de uno de los miembros del equipo siguiendo una lista de páginas a visitar. Estas páginas se han elegido por ser más visitadas, conflictivas, dudosas, de carácter censurable o dirigidas a minorías específicas; y creímos conveniente que se realizara un análisis específico y exhaustivo sobre ellas.

Dichas páginas son:

- The Pirate Bay: https://thepiratebay-proxylist.org/
- Wikipedia: https://es.wikipedia.org/wiki/Wikipedia
- Github: https://github.com/mozilla/lightbeam
- UCM: https://www.ucm.es/
- Comillas: http://www.comillas.edu/es/
- Facebook: https://es-es.facebook.com/
- Twitter: https://twitter.com/?lang=es
- Minijuegos: http://www.minijuegos.com/
- Seriesblanco: https://seriesblanco.com/listado/
- Pornhub: https://es.pornhub.com/

Tipos de extracciones

Las extracciones de los datos **difieren según la versión de Firefox del equipo y las funcionalidades que Lightbeam ofrece para cada uno de ellos**. Por ello, las recopilaciones de datos con las que vamos a trabajar se separan en dos tipos:

- ❖ Simples: son aquellas extracciones de datos realizadas desde Lightbeam en un ordenador con la versión de Firefox 57.0.4. Estos datos se descargan en un archivo .json con los siguientes campos:


```
"d2k1ftgv7pobq7.cloudfront.net": {
  "hostname": "d2k1ftgv7pobq7.cloudfront.net",
  "favicon": "",
  "firstPartyHostnames": [
    "trello.com"
  ],
  "firstParty": false,
  "thirdParties": []
},
"d3ahinqqx1dy5v.cloudfront.net": {
  "hostname": "d3ahinqqx1dy5v.cloudfront.net",
  "favicon": "https://d3ahinqqx1dy5v.cloudfront.net/favicon.ico",
  "firstPartyHostnames": [
    "jkanime.net"
  ],
  "firstParty": true,
  "thirdParties": [
    "www.loadmill.com",
    "fonts.googleapis.com",
    "eu.uspostly.info",
    "uspostly.info",
    "gk.loadmill.com"
  ]
},
```

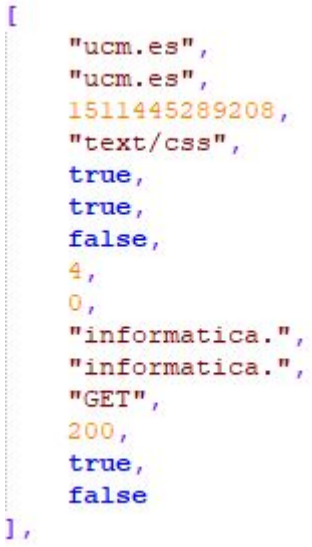

De estos campos se extrae la siguiente información:

- **hostname**: el nombre de la página web.
- **favicon**: si tiene una imagen definida como icono de la página.
- **firstPartyHostnames**: si la página es un tercero, contendrá una lista de sitios de origen.
- **firstParty**: es un booleano que determina si la página es de origen o no (tercero).

- **thirdParties**: si la página es de origen, contendrá una lista de los sitios de terceros que tiene activos.

Este tipo de extracción se corresponde con los datos obtenidos de la navegación de la **asociación de estudiantes**.

❖ Complejas: son aquellas extracciones de datos realizadas desde Lightbeam en un ordenador con la versión de Firefox 45.9.0. Estos datos se descargan en un archivo .json con los siguientes campos:



De estos campos se extrae la siguiente información:

- **source**: URL de la página solicitada.
- **target**: URL de la página cargada por una third-party site.
- **timestamp**: número entero en milisegundos.
- **contentType**: value of the Content-Type header.
- **cookie**: booleano que marca la existencia de una o más Set-Cookie en los headers del target.
- **sourceVisited**: indica si la source se ha cargado por el usuario en una pa ǵ ina o tab.
- **secure**: indica si la página se ha cargado por protocolo HTTPS(true) o por HTTP(false).
- **sourcePathDepth**: métrica de cuántos elementos de path había en la URL de origen.
- **sourceQueryDepth**: métrica de cuántos artículos hay en el query string.
- **sourceSub**: resto del dominio después de eliminar el dominio de primer nivel.
- **targetsub**: igual que sourceSub pero para la target.
- **method**: si la conexión ha sido via GET, POST, PUT, etc...
- **status**: status numérico(entero) de la respuesta.
- **cacheable**: atributo que marca si el servidor ha respondido a un mecanismo de cacheado ("Cache-control: no-cache", "Pragma: no-cache", "Expires: 0", or "Expires") con una fecha del “Date” header.

Este tipo de extracción se corresponde con los datos obtenidos de la **navegación dirigida a páginas populares**.

1.2. Asignación de nodos y aristas

Nuestro proyecto tiene en total dos grandes bloques. Por un lado la navegación llevada a cabo por Libre Lab y por otro, la navegación dirigida a páginas populares. Los scripts que detallaremos en el siguiente punto (1.3) generan de cada navegación múltiples .csv con diferentes asignaciones de nodos y aristas.

Navegación de Libre Lab

- Analizaremos un grafo dirigido con:

  Nodos: Ambas, tanto first-parties como third-parties serán nodos.

  Aristas: Las aristas salen de las first-parties y se dirigen a las third-parties. También se da el caso en el que una arista salga de una third party hasta otra third party. Además, puede ocurrir que una first party salga hasta otra first party. Por lo tanto se tendrá en cuenta el grado de salida de las first-parties y el grado de entrada de las third-parties.

- Analizaremos un grafo dirigido con:

  Nodos: Third-parties

  Aristas: Una arista une dos third-parties entre sí.

  Con este grafo se pretende estudiar la relación entre las third parties.

- Analizaremos un grafo dirigido con:

  Nodos: First-parties

  Aristas: Una arista une dos first parties entre sí.

  Con este grafo se pretende estudiar la relación entre las first-parties.

Navegación dirigida a páginas populares

- Analizaremos un grafo dirigido con:

  Nodos: Ambas, tanto first-parties como third-parties serán nodos.

  Aristas: Las aristas salen de las first-parties y se dirigen a las third-parties. También se da el caso en el que una arista salga de una third party hasta otra third party. Además, puede ocurrir que una first party salga hasta otra first party.

  Por lo tanto se tendrá en cuenta el grado de salida de las first-parties y el grado de entrada de las third-parties.

- Analizaremos un grafo dirigido con:

  Nodos: Third-parties

  Aristas: Una arista une dos third-parties entre si.

  Con este grafo se pretende estudiar la relación entre las third parties.

- Analizaremos un grafo dirigido con:

  Nodos: First-parties

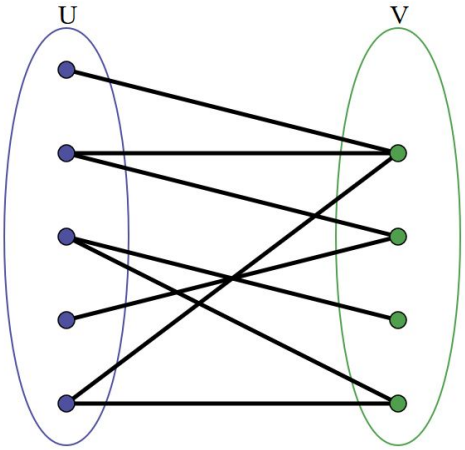
  Aristas: Una arista une dos first parties entre si.

  Con este grafo se pretende estudiar la relación entre las first-parties.

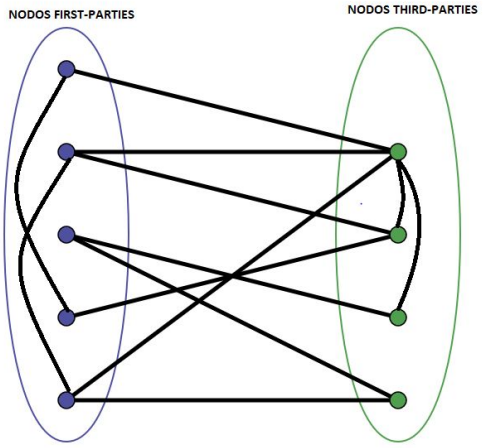
Además, las aristas de esta red pueden mostrar si la third parties **guardan cookies** o si **usan protocolo https**.

No generamos grafos bipartitos en nuestro proyecto. Un grafo bipartito es un grafo cuyos vértices se pueden separar en dos conjuntos disjuntos U y V de manera que las aristas sólo pueden conectar vértices de un conjunto con vértices del otro.

Nuestro grafo no sólo conecta first parties con third parties (lo que encajaría en la definición de grafo bipartito) sino que además conecta third parties entre ellas e incluso algunas first parties, por lo que no se podría dividir en dos conjuntos disjuntos ya que hay aristas que también conectan nodos del mismo tipo.



GRAFO BIPARTITO



NUESTROS GRAFOS CON DOS TIPOS DE NODOS

1.3. Scripts utilizados

Teniendo los diferentes archivos .json extraídos y explicados en el punto (1.1) se han programado dos scripts de Python que filtran la información: `cookies.py` y `lightbeam2gephi.py`.

Estos scripts, junto con el resto de documentación del proyecto, pueden ser consultados en el siguiente proyecto de Github: https://github.com/vdeverdu/SOC

Cookies.py

Este script procesa el archivo .json de la navegación realizada por la asociación de estudiantes Libre Lab. El archivo librelab.json es de tipo **simple**, por lo que solo contiene información de la página web de origen y de las third-parties que tiene asociadas.

Se han programado diferentes funciones que generan archivos csv para hacer análisis específicos:

- createNodes() y createEdges() generan las listas de nodos y aristas que contienen el grafo entero.
- createSources() genera las listas de nodos y aristas de una first party concreta.
- createTargets() hace lo mismo que createSources pero con third parties concretas.
- thirdPartyRelations() genera una lista de aristas que relacionan solo third parties.
- firstPartyRelations() hace lo mismo pero con first parties.
- findNodesWithTarget() busca las first parties que tienen una arista con alguna de las third parties targets y las guarda, además guarda también las first parties comunes que tienen aristas con todos los targets que se le pasen, cuantas más third parties targets menos probabilidades de coincidir las first parties comunes.

Este script no sólo genera **librelab_edges.csv** y **librelab_nodes.csv**, también genera otros .csv que se han utilizado para facilitar la elaboración de otros grafos. Por ejemplo genera **top3TPsInDegree.csv** y **top3TPsOutDegree.csv** utilizados para mostrar las 3 third parties con mayor grado de salida y entrada. Además genera **fps_edges.csv** para crear el grafo que tiene únicamente como nodos first parties y **tps_edges.csv** para crear el grafo que tiene únicamente como nodos third parties.

librelab_edges.csv

Source	Target	Type
mundokodi.(	0.gravatar.cc	Directed
lamiradadeli	0.gravatar.cc	Directed
www.minec	0914.global.s	Directed
lamiradadeli	1.gravatar.cc	Directed
aidancbrady.	1.gravatar.cc	Directed
www.mint.c	4354787.fls.c	Directed
www.reddit.	a.thumbs.re	Directed
www.blizzar	a8270235338	Directed
worldofwarc	a8270235338	Directed
www.reddit.	aax-eu.amaz	Directed
www.google	aax-eu.amaz	Directed
www.offens	aax-eu.amaz	Directed

Librelab_nodes.csv

Id	FirstParty
accounts.goc	True
add0n.com	True
addons.mozi	True
aidancbrady.	True
apartamentc	True
archlinuxarm	True
aur.archlinux	True
bolotweet.d	True

Lightbeam2gephi.py

Este script procesa el archivo .json de la navegación “popular”, es decir, la navegación dirigida a las páginas populares. El archivo popular.json es de tipo **complejo**, por lo que no sólo contiene información de la página web de origen y de las third-parties que tiene asociadas,

sino que también tiene varios booleanos que aportan más datos, nosotros utilizaremos dos concretamente, si la página maneja cookies o no y si la página es segura o no.

- createAll() generan las listas de nodos y aristas que contienen el grafo entero.
- createSource() genera la lista aristas de una first party concreta.
- createTarget() hace lo mismo que createSource pero con una third parties concreta.
- thirdPartyRelations() genera una lista de aristas que relacionan solo third parties.
- firstPartyRelations() hace lo mismo pero con first parties.
- findNodesWithTarget() busca las first parties que tienen una arista con alguna de las third parties targets y las imprime en la terminal.

Estos archivos de aristas contienen un listado completo de todas las páginas de origen (**Source**) y las páginas de terceros a los que mandan información (**Target**), junto a los valores que señalan si se va a utilizar para generar aristas dirigidas (**Type**), si la conexión almacena cookies (**Cookie**) y si la página de destino o target funciona con un protocolo http o https (**Secure**).

Este script genera **popular_edges.csv** y **popular_nodes.csv** que corresponden a los nodos y aristas que utilizaremos para generar el grafo popular, y también **fps_edges.csv** para generar el grafo que tiene únicamente como nodos first parties y **tps_edges.csv** para generar el grafo que tiene únicamente como nodos third parties.

Aquí mostramos una extracción en base a Google.com :

Source	Target	Type	Cookie	Secure
google.com	gstatic.com	Directed	True	True
google.com	gstatic.com	Directed	True	True
google.com	gstatic.com	Directed	True	True
google.com	gstatic.com	Directed	True	True
google.com	gstatic.com	Directed	True	True
google.com	gstatic.com	Directed	True	True
google.com	fonts.googleapis.com	Directed	False	True
google.com	ggpht.com	Directed	False	True

Mostramos ahora un ejemplo de una extracción que tiene como Target definido Facebook:

Source	Target	Type	Cookie	Secure
elpais.com	facebook.com	Directed	True	True
elpais.com	facebook.com	Directed	True	True
elpais.com	facebook.com	Directed	True	True
tviso.com	facebook.com	Directed	True	True
tviso.com	facebook.com	Directed	True	True
tviso.com	facebook.com	Directed	True	True
seriesblanco.com	facebook.com	Directed	True	True

1.4. Contenido del proyecto

Por lo tanto, nuestro proyecto contará con los siguientes casos:

Obtención de los datos	Tipo de extracción	Asignación de nodos y aristas	Grafos generados	Script correspondiente
Navegación realizada por asociaciones de estudiantes	Simple	1. Grafo dirigido **"LibreLab"** **NODOS:** Tanto third-parties como first-parties **ARISTAS:** Salen desde las first-parties a las third-parties. También salen desde una third party a otra third party, y desde una first party a otra first party 2. Grafo dirigido cuyos nodos son las third parties **NODOS:** Third parties **ARISTAS:** Una arista une dos third parties entre sí 3. Grafo dirigido cuyos nodos son las first parties **NODOS**: First-parties **ARISTAS**: Una arista une dos first parties entre sí	3 grafos dirigidos	.json : librelab.json .py : cookies.py
Navegación dirigida a páginas populares	Compleja	1. Grafo dirigido **"popular"** **NODOS:** Tanto third-parties como first-parties **ARISTAS:** Salen desde las first-parties a las third-parties. También salen desde una third party a otra third party, y desde una first party a otra first party 2. Grafo dirigido cuyos nodos son las third parties **NODOS:** Third parties **ARISTAS:** Una arista une dos third parties entre sí 3. Grafo dirigido cuyos nodos son las first parties **NODOS:** First-parties **ARISTAS:** Una arista une dos first parties entre sí	3 grafos dirigidos	.json : popular.json .py : lightbeam2gephi.py

2. Análisis de datos

2.1. Objetivos del análisis

El objetivo de este proyecto se centra en la investigación y análisis en detalle de los datos extraídos. La visualización de los datos toma un segundo plano. Por tanto nuestra investigación se centra en detectar nodos de interés para su posterior análisis: por ejemplo, si vemos que smartadserver sale como inseguro porque ni él ni prácticamente ninguna de sus conexiones usa protocolo https, entonces sabemos que es una página de la que puede interesarnos su análisis.

Por ello podemos decir que el objetivo de este análisis es encontrar cuales son las Third-Parties a las que están más expuestas los usuarios (grado de exposición de las Third-Parties que utilizan los nodos principales de los grafos que hemos analizado).

Calculando el grado de salida, buscamos hallar cuántos de los nodos principales tiene una Third-Party (es decir, cuantos nodos están expuestos a esa third-party).

Analizando las third parties con más grado de entrada buscamos encontrar aquellos nodos que saben mucho sobre empresas, o cuales saben sobre una empresa concreta (ej: si de facebook te vas a doubleclick.net, y de doubleclick va a muchos más, sabes que está expuesto, y que doubleclick.net sabe mucho de otras empresas).

Además, hemos visto que The Pirate Bay y SeriesBlanco comparten un CryptoMiner (SeriesBlanco no dice en ninguna parte que mine en su página web).

2.2. Métricas/algoritmos aplicados al análisis de la red

Hemos analizado **dos redes**: **una en la que se han visitado páginas concretas**, y **otra sacada de la navegación de la asociación Libre Lab**.

En la primera nos hemos centrado en analizar ciertas métricas vistas en clase: grado medio, centralidad del vector propio, PageRank, Betweenness centrality, Modularidad (comunidades). Y además hemos exportado los datos del grafo a un mapa interactivo desde el que hemos podido analizar, y visualizar de una forma más clara, las Third-Parties de cada uno de los nodos que estábamos analizando.

En la segunda red nos hemos centrado en visualizar y estudiar desde Gephi los tres nodos con mayor grado de salida y de entrada.
La extensión de Lightbeam de Libre Lab no distinguía si tenían cookies o no, o si usaban protocolo https o no. Por tanto, en este grafo nos hemos centrado más en un análisis más exhaustivo entre los respectivos nodos y sus Third-Parties.

Además, en ambas redes hemos modificado el código para mostrar únicamente las relaciones entre First-Parties y entre Third-Parties, de forma independiente y sin la influencia de sus respectivos.

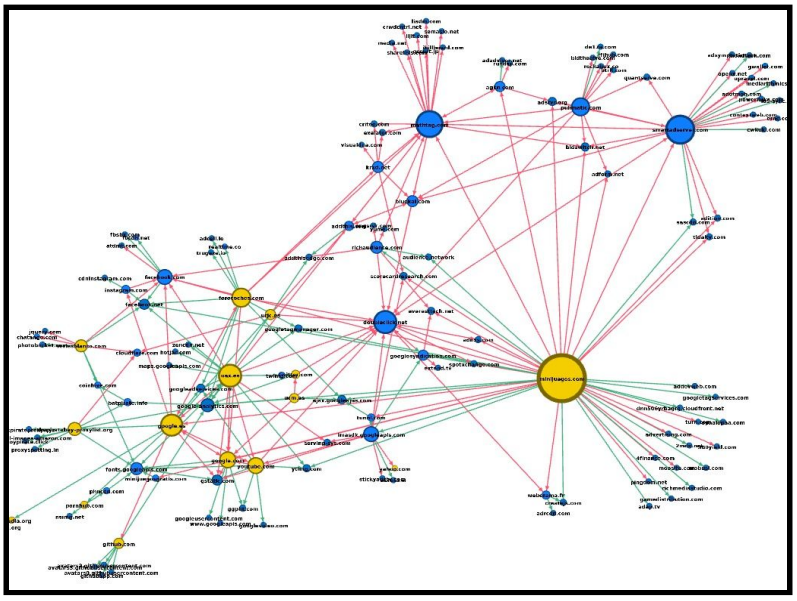
2.3. Visualización de los datos

Visualización del grafo de la navegación Popular

1. Importar archivo (primero nodos y luego edges).
2. Calcular grado medio.
3. Nodos: paleta de colores basada en Third-Parties (amarillas) y First-Parties (azules).
4. Nodos: cambiar el tamaño de 10 a 100 por el grado (tamaño → ranking → 10 a 100).
5. Distribución: force atlas 2 (escalado 200). Marcamos disuadir hubs.

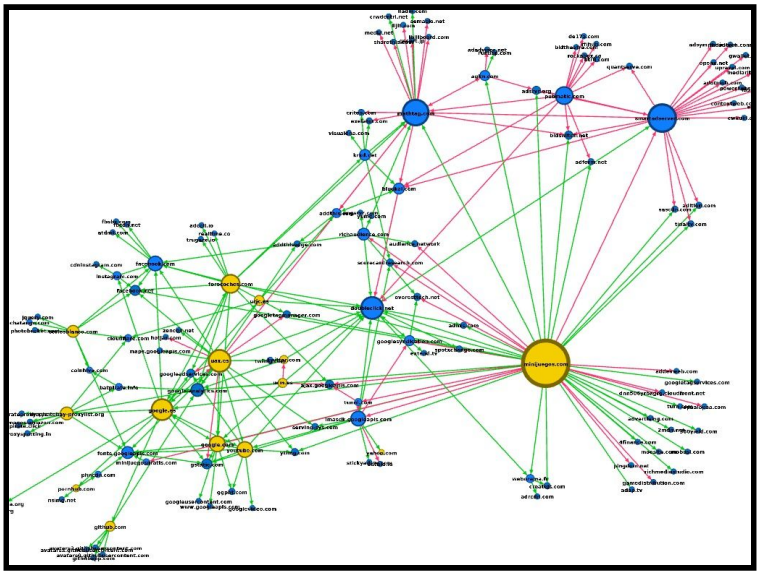
Grafo de aristas representando si tienen cookies

Al grafo se le aplica un filtro de color en las aristas: asignar partición por atributo cookie (rojo si guarda cookies, verde si no).



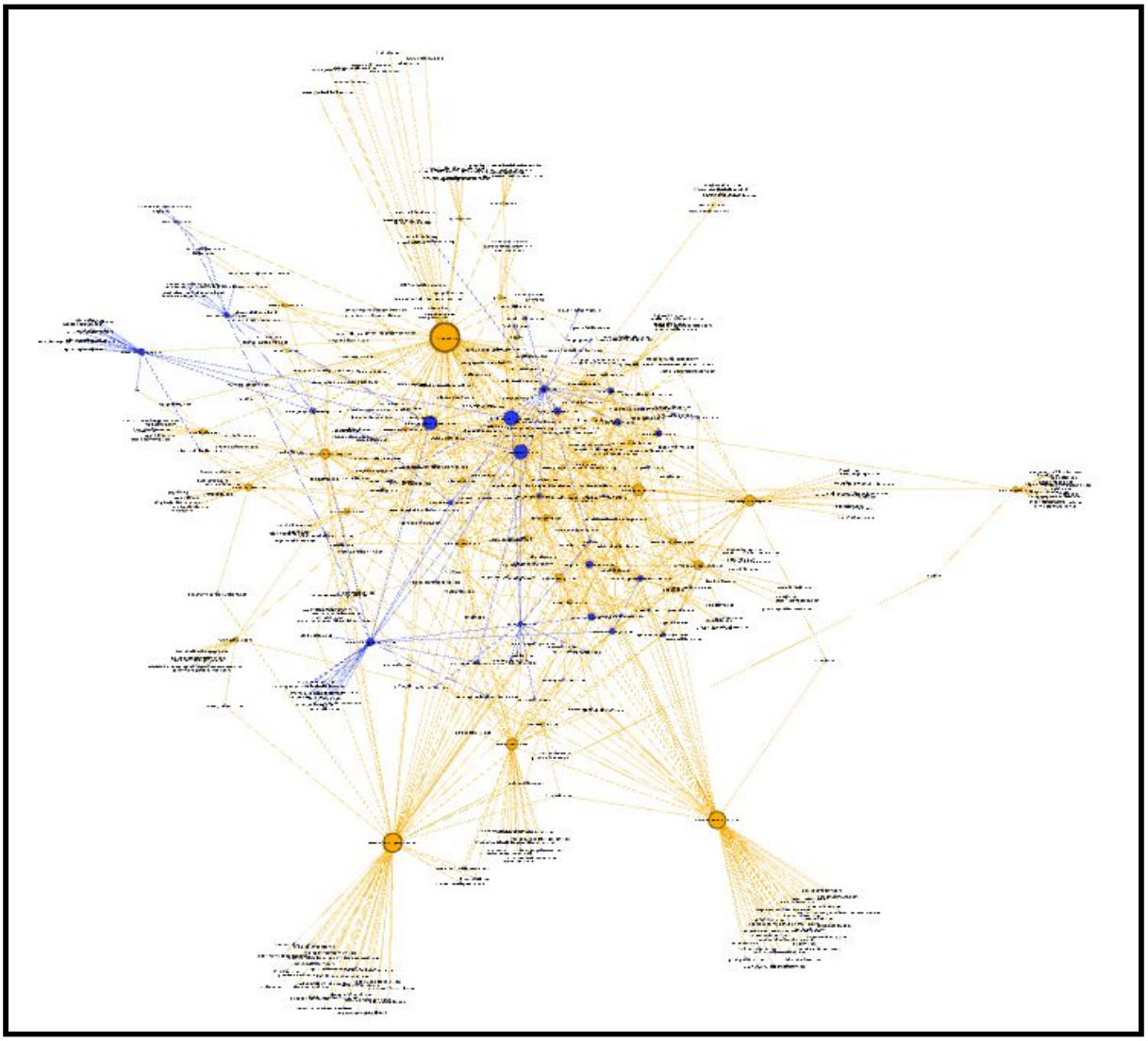
Grafo con aristas representando el tipo de protocolo

Al grafo se le aplica un filtro de color en las aristas: asignar partición por atributo secure (rosa si utiliza HTTP, verde si utiliza HTTPS).



Visualización del grafo de la navegación de LibreLab

1. Importar archivos (primero nodos y luego edges).
2. Calcular grado medio.
3. Nodos: paleta de colores basada en Third-Parties (amarillas) y First-Parties (azules).
4. Nodos: cambiar el tamaño de 10 a 100 por el grado (tamaño -> ranking -> 10 a 100).
5. Distribución: force atlas 2 (escalado 200). Marcamos disuadir hubs.
6. Las aristas por defecto se ponen del color del nodo dependiendo de si se relacionan como First-Parties o Third-Parties.



Visualización centrada en una página web concreta

Se han programado funciones específicas para poder extraer información de las páginas web sobre las que queríamos hacer un análisis en profundidad.

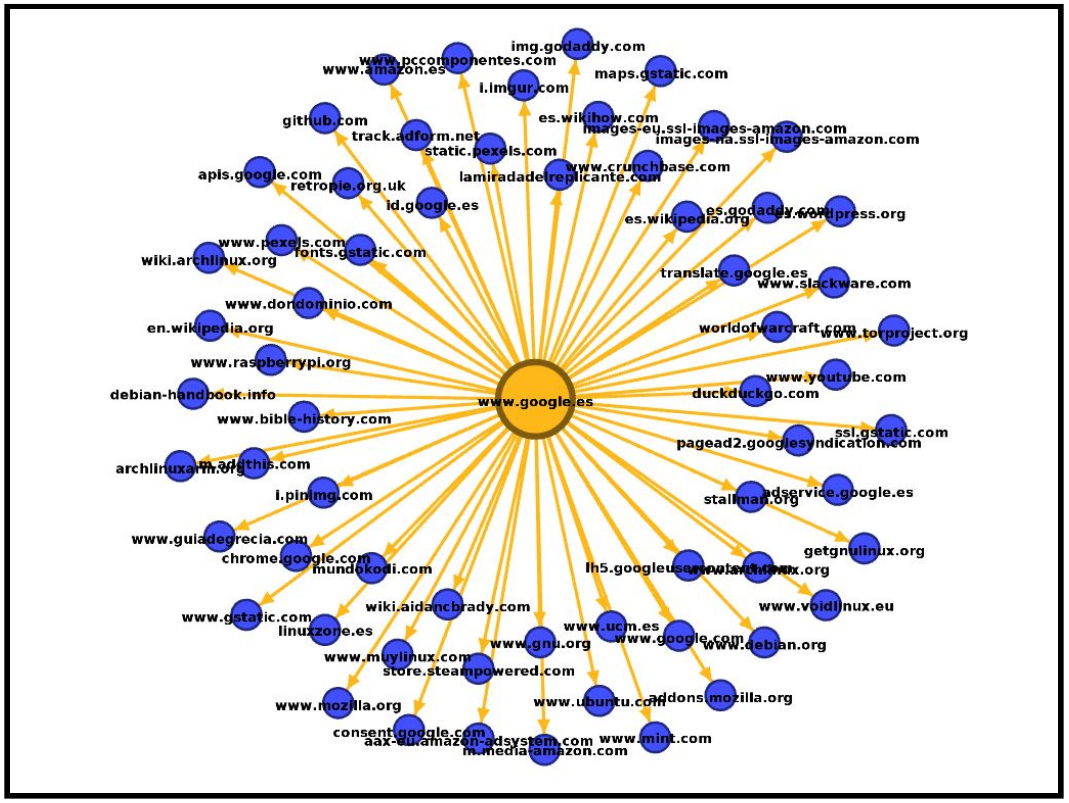
Dichas funciones permiten mostrar únicamente el nodo específico que queremos analizar y los otros nodos con los que está relacionado. Al elegirse los tres nodos que queremos, el programa busca dichos nodos y, al encontrarlos, crea aristas con todas sus Third-Parties.

Por tanto, el código crea dos archivos: sourcenodes.csv y sourceedges.csv (con el nombre de dicho nodo).

- Ej: para el grado de salida de **www.google.es** hemos mostrado un grafo individual de dicho nodo con todas sus Third-Parties.

Primero se importa el archivo de nodos, y posteriormente el de aristas. Después se cambian los atributos:

1. Nodos: paleta de colores basada Third-Parties (amarillas) y First-Parties (azules)
2. Nodos: cambiar el tamaño de 30 a 100 por el grado (tamaño -> ranking -> 10 a 100)
3. Distribución: force atlas 2 (escalado 200). Marcamos disuadir hubs.

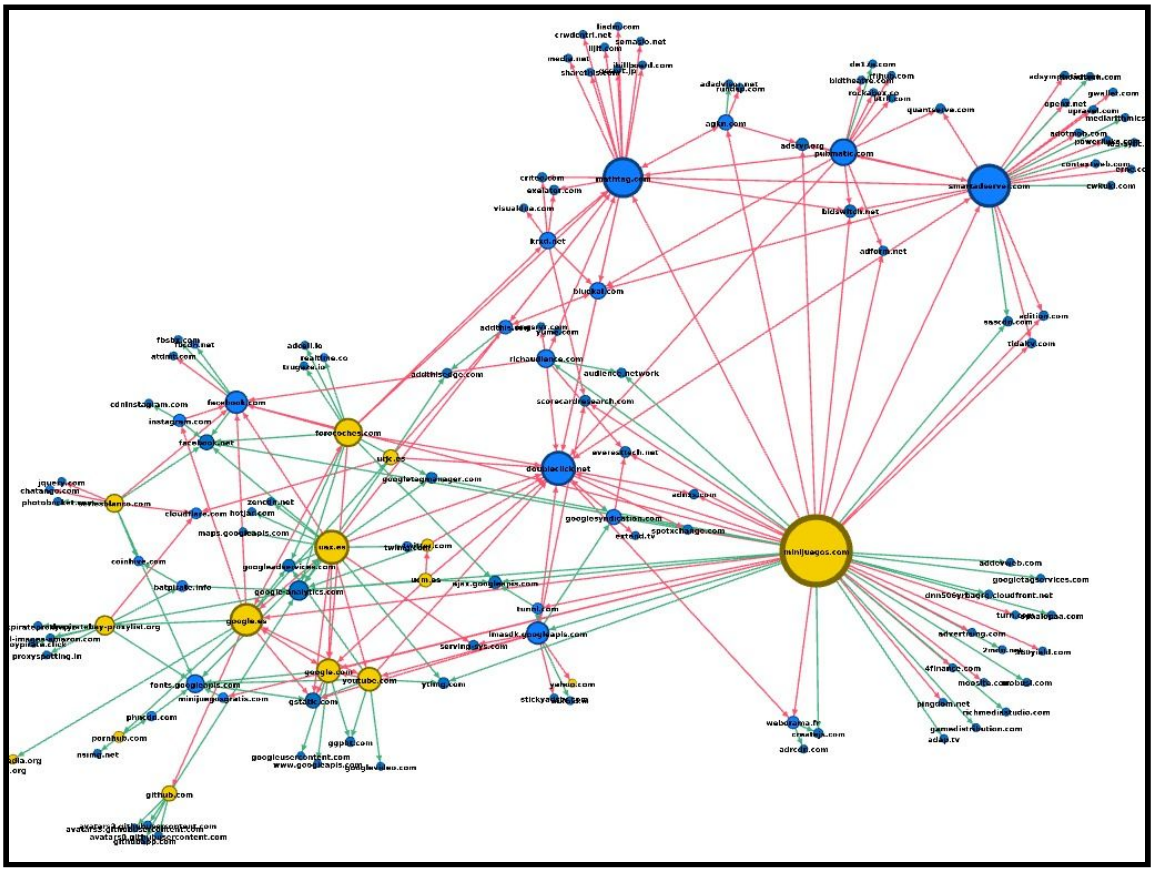


Ejemplo de www.google.es grado de salida

2.4. Análisis de los datos

Estudio Navegación popular

- Total Nodos: 139
- Total Aristas: 246



Medidas locales de centralidad

Grado de entrada (third parties más populares)

Nodos con mayor grado de entrada:

- doubleclick.net - 14
- googleanalitics.com - 9
- mathtag.com - 8

Mirando estos nodos, al tener un **alto grado de entrada**, podemos afirmar que son **Third-Parties muy usadas por otras empresas**. En general, las empresas tienden a usar mismas Third-Parties a la hora de sacar información, por eso estos nodos tienen un grado de entrada más alto de lo habitual.

(Se ve que sobretodo predomina el marketing digital).

Grado de salida

Nodos con mayor grado de salida:

- minijuegos.com - 45
- smartadserver.com - 22
- uax.es - 18

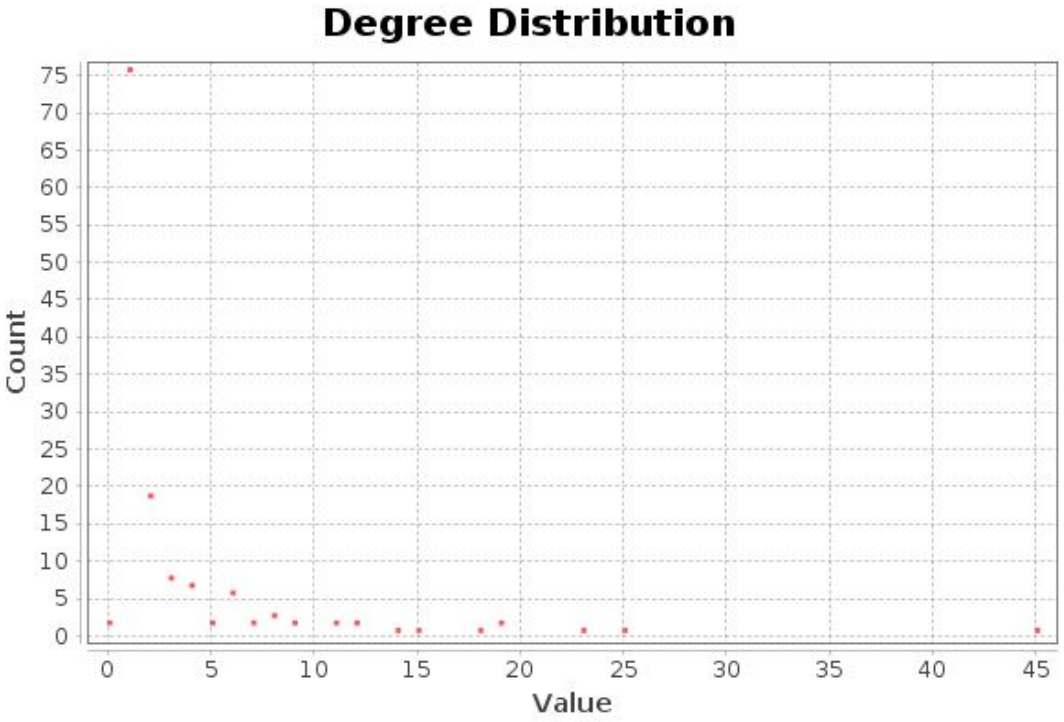
Podemos apreciar que **las tres principales son First-Parties con un alto grado de salida**. Esto se debe a que tienen asociadas muchas Third-Parties a ellas, a través de las que reciben información.

Grado medio

- Grado medio: 1,77
- Grado medio con pesos: 12,604

Nodos con mayor grado medio de entrada y de salida:

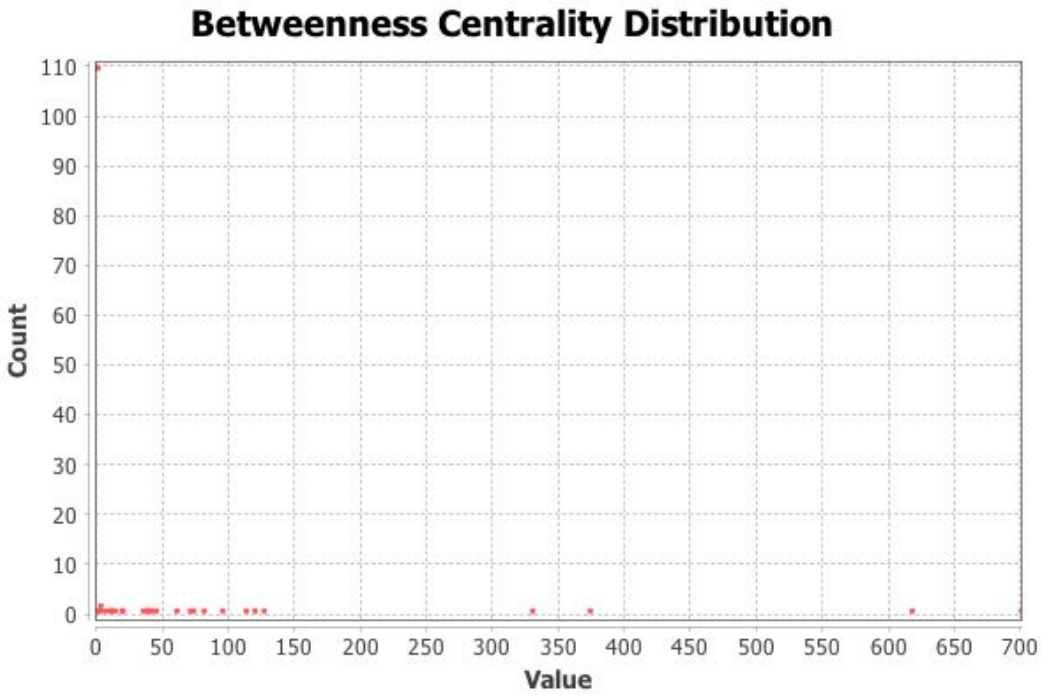
- minijuegos.com - 45
- smartadserver.com - 25
- mathtag.com - 23



Betweenness Centrality

Nodos con mayor Intermediación:

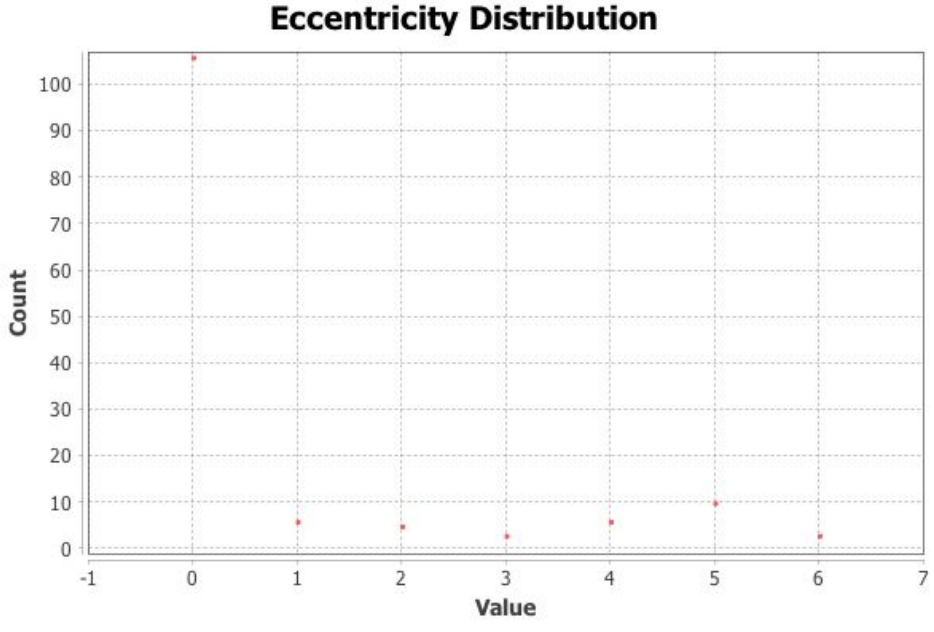
- doubleclick.net - 699,33
- smartadserver.com - 616,83
- google.es - 372,87



Como podemos ver, **los nodos más periféricos son Third-Parties**, mientras que **los más centrados son First-Parties**. Gracias a esto, podemos ver el **impacto que tiene la página en la red social**. Las tres que mostramos son las que más impacto tienen y las que más interesantes serían de estudiar.

<u>Excentricidad</u>

- Google.com - 6.0
- Addthis.com - 6.0
- Tunnl.com - 6.0



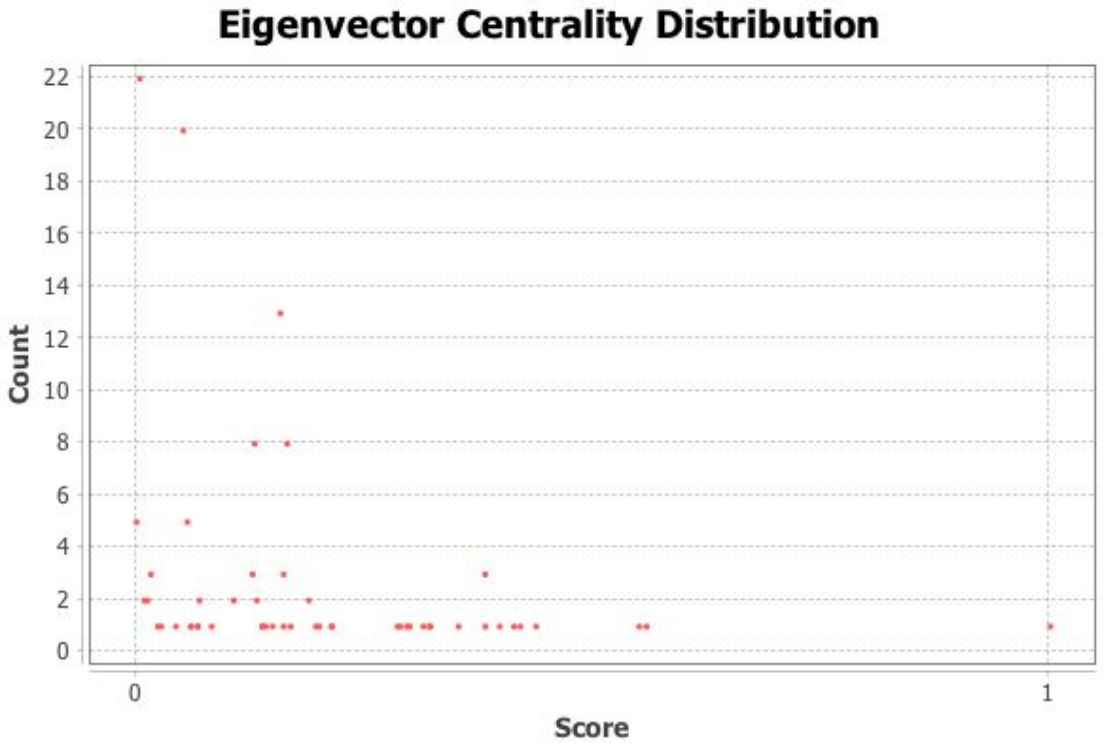
Centralidad de vector propio

Parametros calculados:

- Network Interpretation: directed
- Number of iterations: 100
- Sum change: 0.009539292864450832

Nodos con mayor centralidad de vector propio:

- doubleclick.net - 1,0
- gstatic.com - 0,56
- bluekai.com - 0,55



Podemos observar que **doubleclick.net es el nodo con más prominencia en el grafo**. Esto se debe a que es el nodo más conectado y que más Third-Parties tiene para poder sacar información.

PageRank

- **Probability(p):** 0.85
- **Epsilon:** 0.01

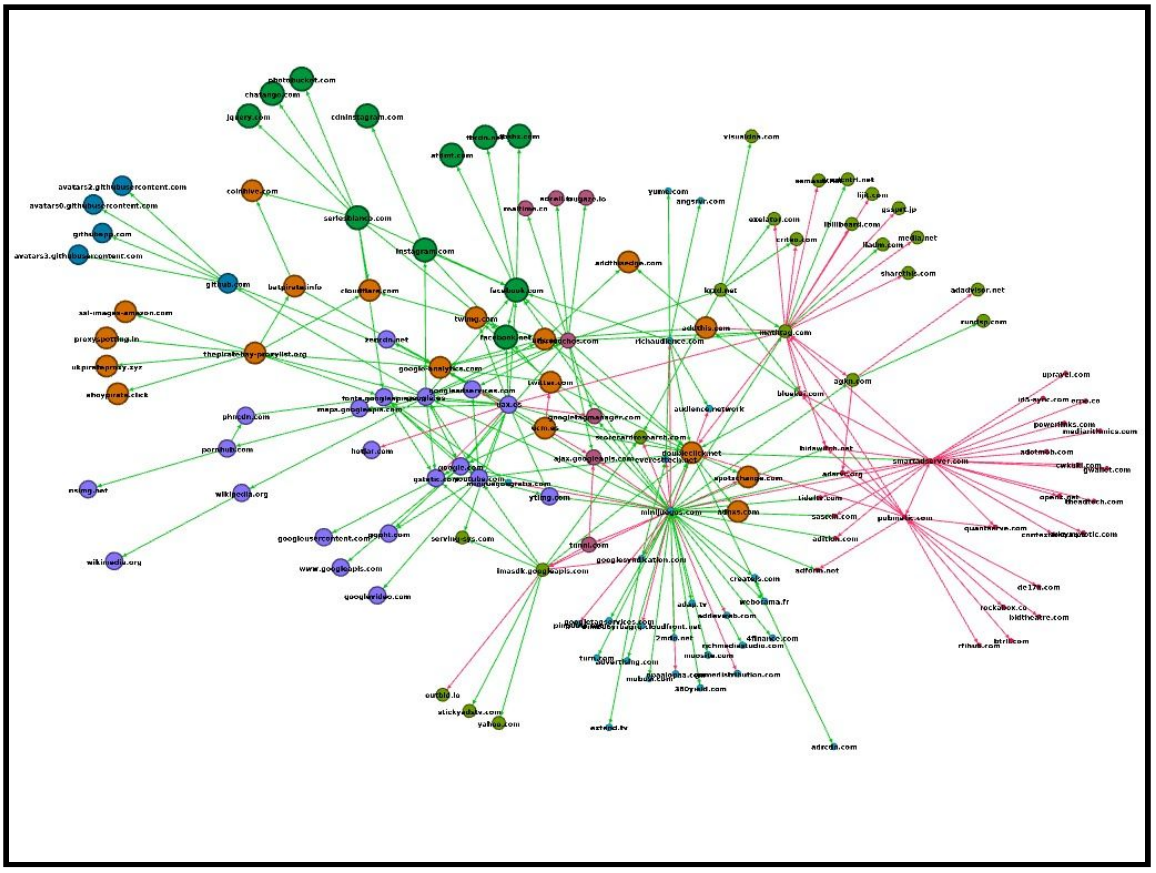
Nodos con mayor PageRank:

- gstactic.com - 0,026
- doubleclick.net - 0,022
- forms.googleapis.com - 0,020

Al fijarnos en el PageRank podemos ver que, efectivamente, **las páginas que más prominencia tenían están conectadas a su vez con otras muchas páginas**. Por tanto, la red de información a la que están conectadas es aún mayor de lo que podemos observar únicamente con la centralidad del vector propio.

Modularidad

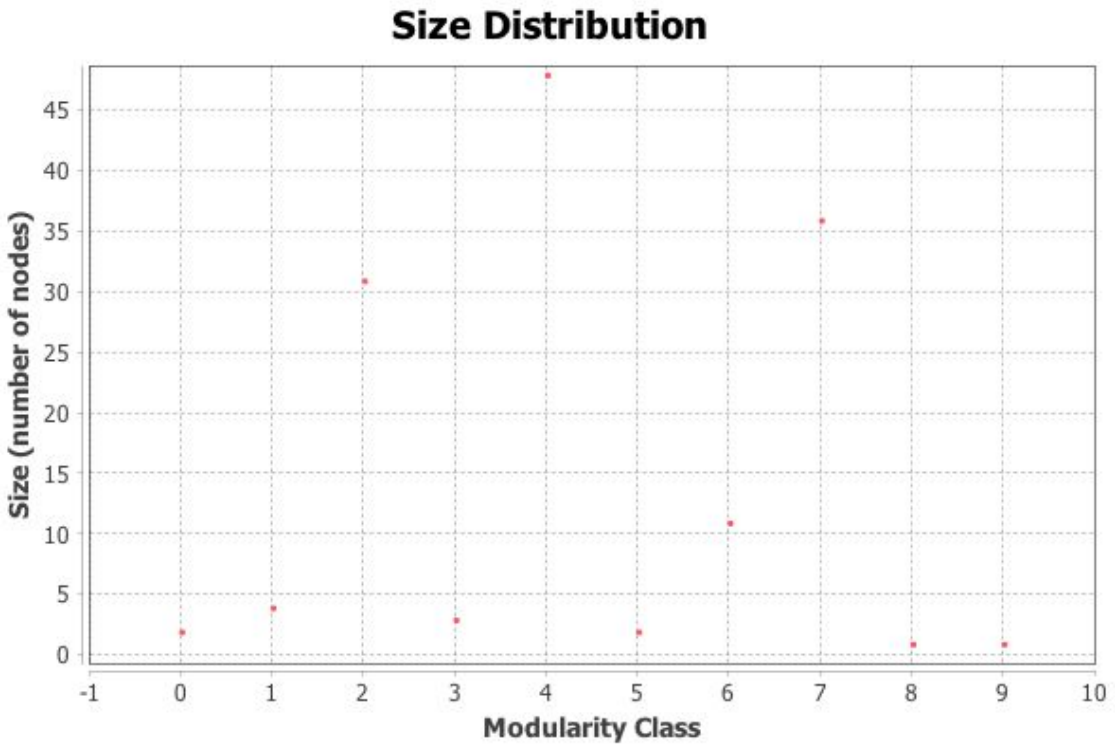
Cambiando el tamaño por **centralidad** y usando la **modularity class**, logramos una visualización más apropiada para mostrar las **comunidades**, que, como podemos ver, se distribuyen por un nodo principal junto a sus Third-Parties.



Hemos puesto una distribución **Yifan Hu** con **distancia óptima "300"** y, si cambiamos el color de las aristas a seguro (verde) o no seguro (rojo), se ve claramente cómo hay **una zona segura y una menos segura** (predominando smartadserver).

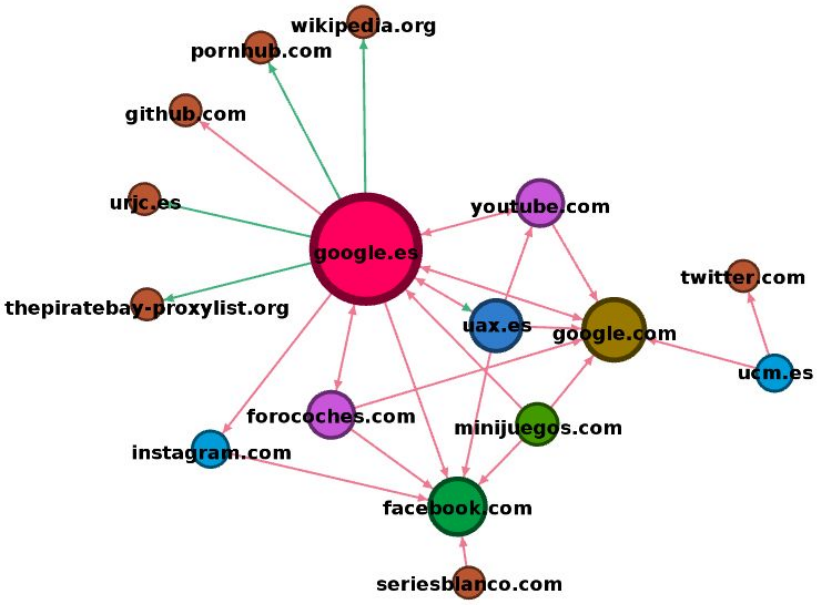
Modularidad - 0,725

Número de comunidades: 10



Por tanto, podemos afirmar que en general las Third-Partie**s asociadas a páginas seguras se conectan entre ellas**, mientras que **las páginas que no tienen dicho protocolo activado**, y que por ende son más inseguras, **se relacionan entre ellas también**. Esto es un claro ejemplo de cómo en internet predominan los espacios inseguros o seguros como tal, en vez de estar interconectados de formas aleatorias. Por ejemplo, si entras en **google maps, facebook, ucm.es**, lo normal es que estas estén conectadas con páginas con protocolo https. Mientras que si entras en una página como **smartadserver**, las conexiones serán en su mayoría inseguras.

Relaciones entre First-Parties del grafo popular

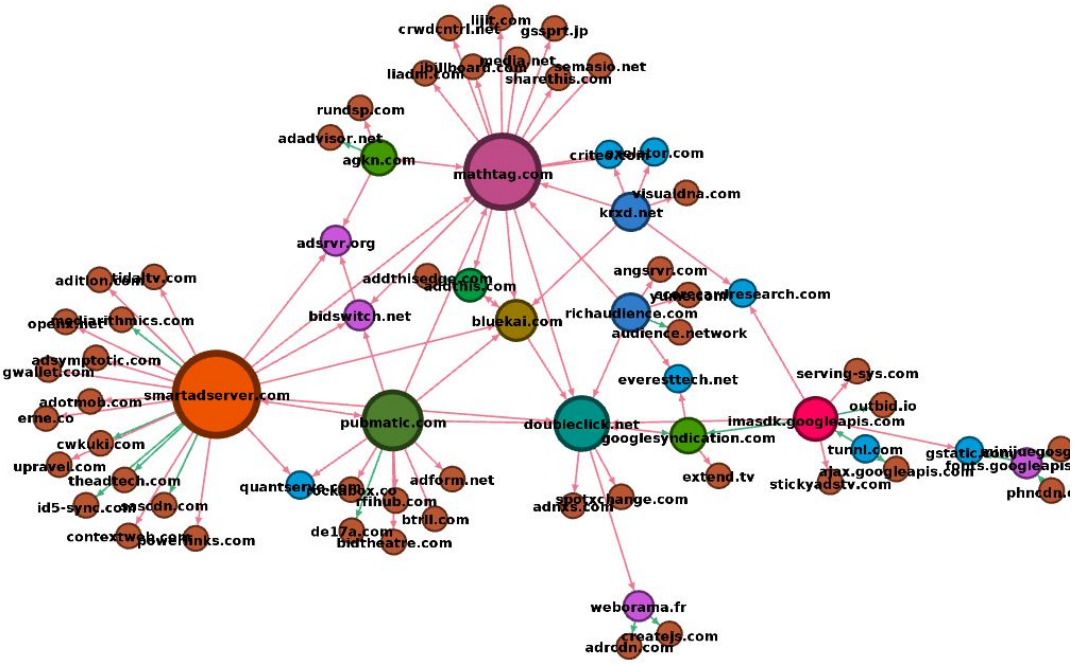


Podemos diferenciar dos hubs principales: uno de **smartadserver.com** y **mathtag.com**.
Este grafo nos muestra que las First-Parties se conectan básicamente mediante **google.es**.

Como ejemplo, podemos ver que **doubleclick.net** es una First-Party que actúa a su vez como Third-Party para otras páginas. Por un lado las cookies de DoubleClick están asociadas con doubleclick.net, el dominio de DoubleClick (si un navegador visita un sitio que muestra anuncios de DoubleClick, el navegador no está en un sitio del dominio de DoubleClick).
Por tanto, el servidor de DoubleClick se convierte en un servidor de terceros, por lo que las cookies que envía el servidor en este contexto se denominan cookies de terceros.

-(Información de explicación de uso de cookies por doubleclick.net)-

Relaciones entre Third-Parties del grafo popular



Al igual que en las First-Parties, en este grafo podemos ver que las Third-Parties también se relacionan entre ellas. Es decir, tenemos Third-Parties que a su vez tienen Third-Parties.

Podemos ver que si por ejemplo google usa google apis para registrar a sus visitantes, google apis puede tener varias Third-Parties que distribuyan sus funciones a varios servidores (por tanto tendremos Third-Parties conectadas con Third-Parties).

Mapa interactivo

Después de darle forma con Gephi hemos exportado los datos creando un mapa interactivo y guardándolo en un servidor. Puedes entrar a él y ver las Third-Parties que acceden a los nodos de nuestra investigación.

Nota: google aparece relacionado con varios nodos ya que se accede a los nodos desde su search engine, además almacena cookies. Esto ocurre principalmente en las páginas que tienen un inicio de sesión de google, también pasa a veces con facebook. Ignorad estos nodos excepto en su propio análisis.

Los nodos se evaluarán según la cantidad de Third-Parties que tenga un nodo y si guardan cookies o no, de tres formas diferentes:

1. Seguro: Sin Third-Parties o Third-Parties propias.
2. Dudoso: Tiene Third-Parties extrañas que no almacenan cookies o Third-Parties conocidas que almacenan cookies.
3. Expuesto: Tiene Third-Parties extrañas que almacenan cookies.

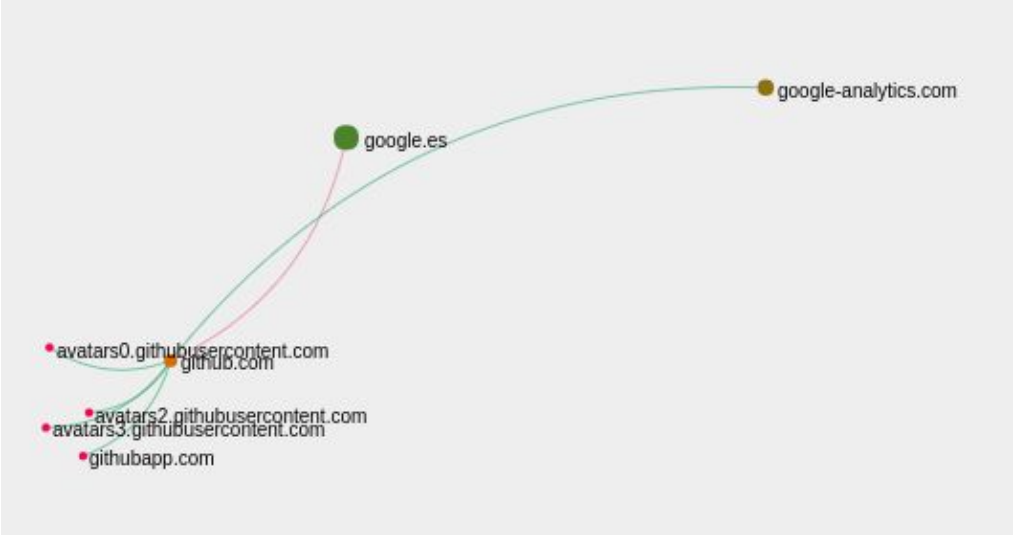
Vamos a analizar las 10 First-Parties que hemos considerado más interesantes teniendo en cuenta la visualización que nos proporciona el mapa.

- Wikipedia ---> Seguro

Tiene una sola arista con **wikimedia.org**. Wikimedia es un movimiento global cuyo objetivo es proporcionar contenido educacional gratuito,.Wikipedia es parte de este movimiento, por lo que tiene sentido que tenga acceso. Además no almacena cookies de ningún tipo, así que navegar por Wikipedia es totalmente seguro.

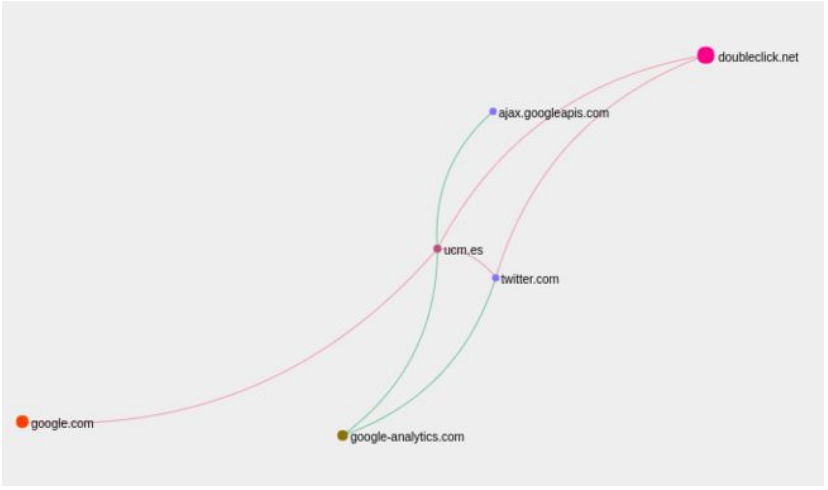
- Github ---> Seguro

Tiene 4 aristas con Third-Parties propias que utiliza para recibir el contenido de los usuarios desde su servidor. Además tiene una arista con **google-analytics.com**, que es un servicio de google muy popular utilizado para monitorear webs. Por ejemplo, ver la cantidad de visitas o tiempo de navegación media del usuario, no guarda datos personales y además no almacena cookies, así que es seguro.



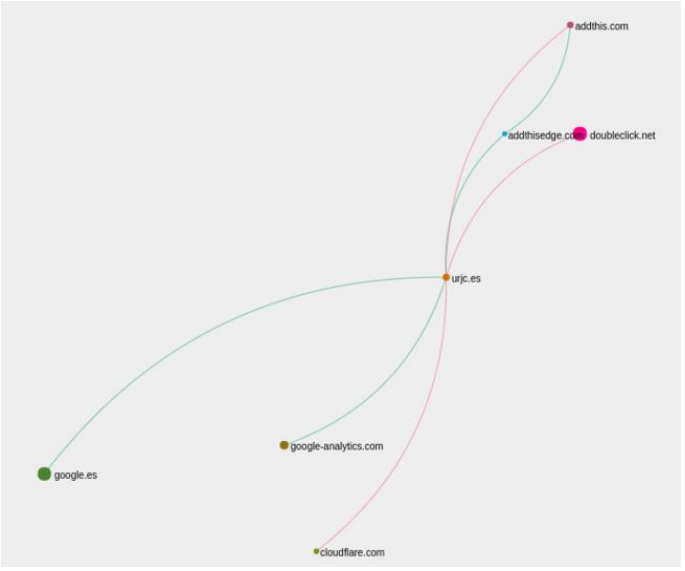
- UCM ---> Dudoso

La web de la Complutense utiliza varios servicios externos de google, ajax.google.apis y google-analytics.com, aunque estos dos no almacenan cookies. Sin embargo el tercero, **doubleclick.net**, sí que las almacena. Este último es un servicio dedicado a proveer soluciones digitales y de marketing. Aunque sea un servicio de google y por ello podamos dar por hecho que es seguro, está almacenando cookies de los usuarios desde la web de la Complutense. También tiene una arista con twitter que almacena cookies, esto se debe a que en la web de la UCM se pueden implantar feeds de twitter para proveer información de algún tema específico, como el que podemos encontrar en la sección de la **biblioteca**.



- URJC ---> Dudoso

La web de la Universidad Rey Juan Carlos tiene similitudes con la de la Complutense pero está más expuesta. Tiene una arista que no guarda cookies con **google-analytics** y otra que sí con **doubleclick.net**, pero además utiliza **cloudfare.com** y **addthis.com**, para optimizar su web. Estos también guardan cookies de los usuarios desde la URJC.



- Facebook ---> Dudoso (Pero está expuesto a otras webs que pueden no ser confiables)

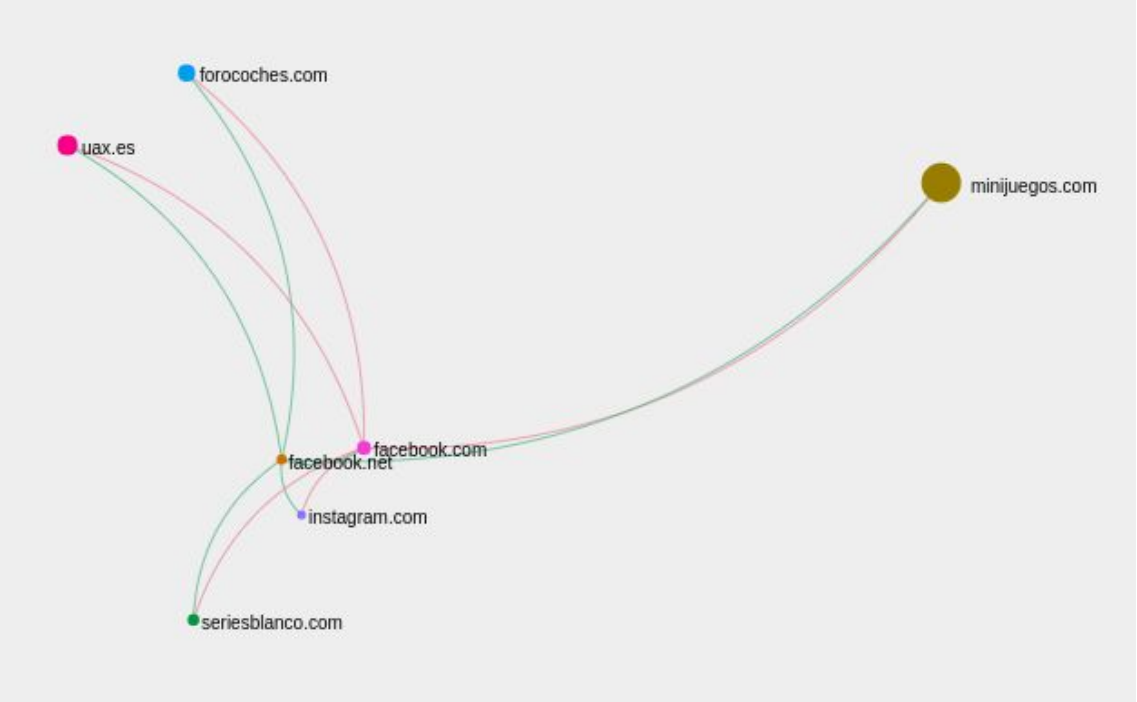
Facebook tiene una arista con una API propia, facebook.net, que no almacena cookies.

Hasta ahí bien, sin embargo podemos ver 5 nodos conectándose a ambas, estos nodos son **minijuegos.com**, **instagram.com** , **uax.es** (Universidad Alfonso X), **seriesblanco.com** y **forocoches.com**.

Curiosamente estos nodos no almacenan cookies cuando acceden a facebook.net pero sí lo hacen con facebook.com, como hemos dicho antes esto se debe a que estas páginas tienen una opción de registrarse con facebook, todas menos forocoches. Según indican los términos de uso en la web de forocoches, es para mostrar publicidad más eficientemente, pero no podemos estar seguros de ello.

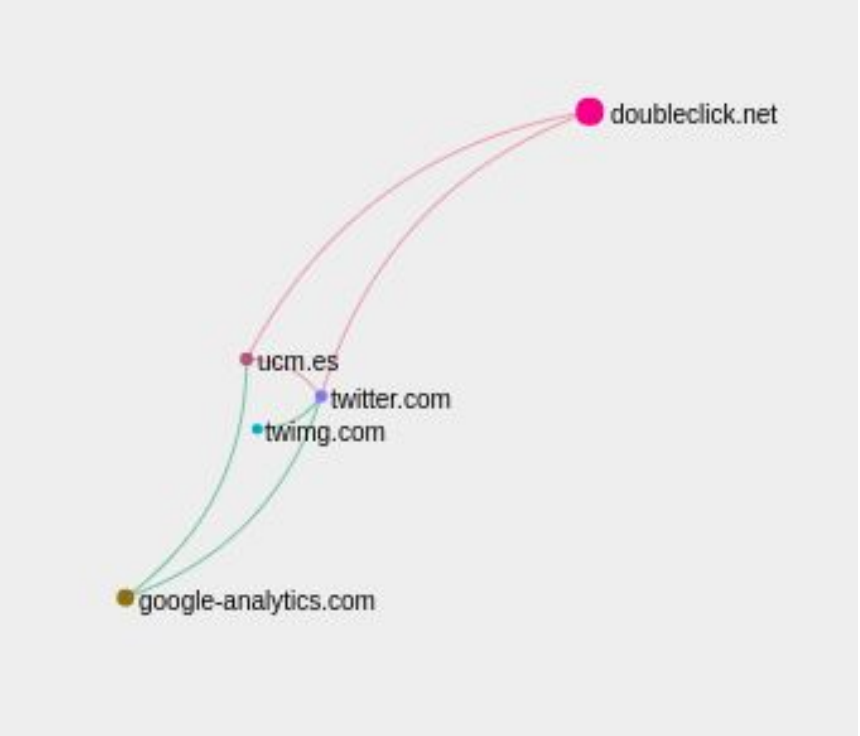
Esto no perjudica directamente a facebook, porque facebook no se conecta con forocoches sino al revés, pero si utilizas forocoches pueden obtener información de tu facebook.

Forocoches también se conecta a varias APIs que almacenan cookies y de las que no hemos encontrado datos, por lo que no sabemos qué hacen con ellos.



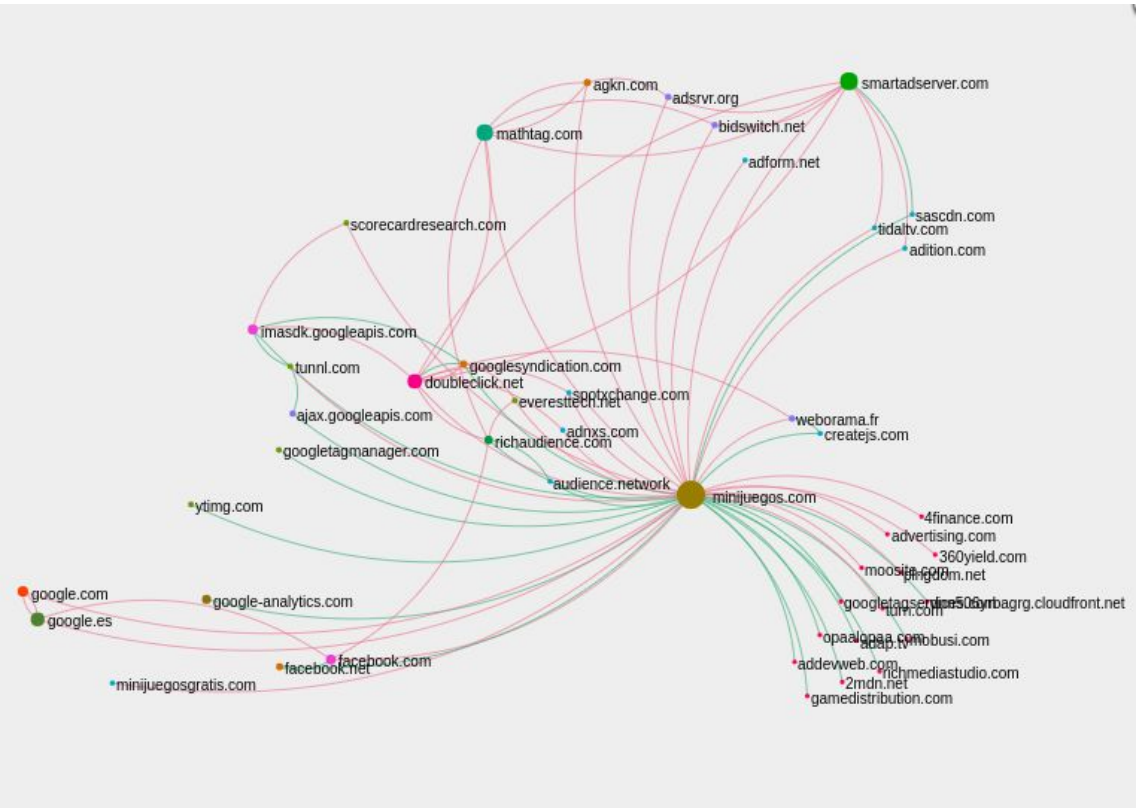
- Twitter ---> Dudoso

Tiene dos aristas que no almacenan cookies, **google-analytics.com** y **twimg.com**, esta última es una API propia de Twitter para enviar y recibir contenido multimedia desde su servidor. También aparece aquí la UCM por los feeds que hemos mencionado antes. Lo único que hace que no sea completamente seguro es que también utiliza **doubleclick.net**.



- Minijuegos ---> Expuesto

Minijuegos es el nodo que más nos ha sorprendido, tiene conexiones con varias Third-Parties de marketing digital mencionadas anteriormente. Y sobre todo al poner anuncios en su página web permiten almacenar cookies a muchas webs de terceros, la mayoría de ellas son de publicidad.

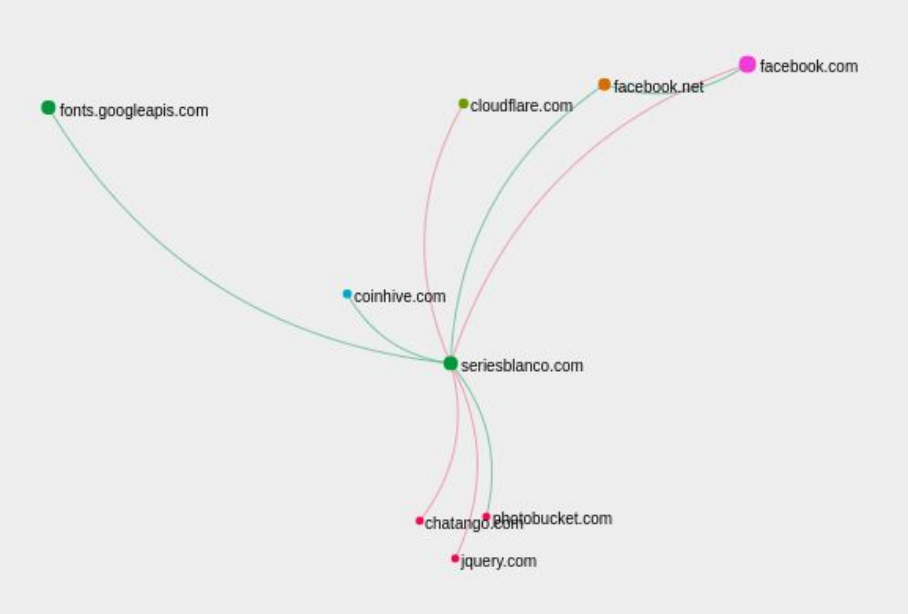


- Seriesblanco ---> Expuesto

Se conecta con 3 Third-Parties que almacenan cookies, Cloudfare que ya la hemos visto varias veces, **jquery.com** que es una librería de frontend y **chatango.com** para moderar chats.

Además se conecta con dos TPs que no almacenan cookies, **photobucket.com** para almacenar contenido multimedia y **coinhive.com**. Esta última es muy interesante puesto que a pesar de no almacenar cookies es un cryptominer que utiliza la potencia computacional de los visitantes de Seriesblanco para minar criptomonedas y así conseguir una remuneración.

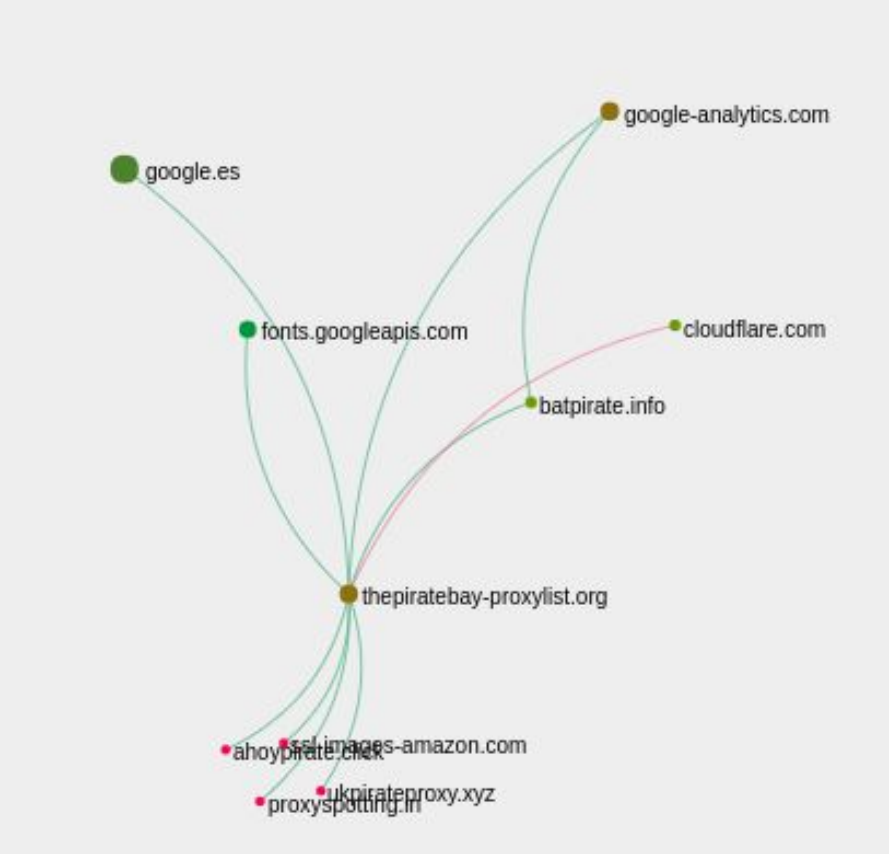
Esto no tendría porque ser un problema si indicarán en algún lado de su web que utilizan un cryptominer y que los usuarios siguieran visitando su web siendo conscientes de esto, sin embargo no se indica en ningún lado.



- ThePirateBay ---> Expuesto

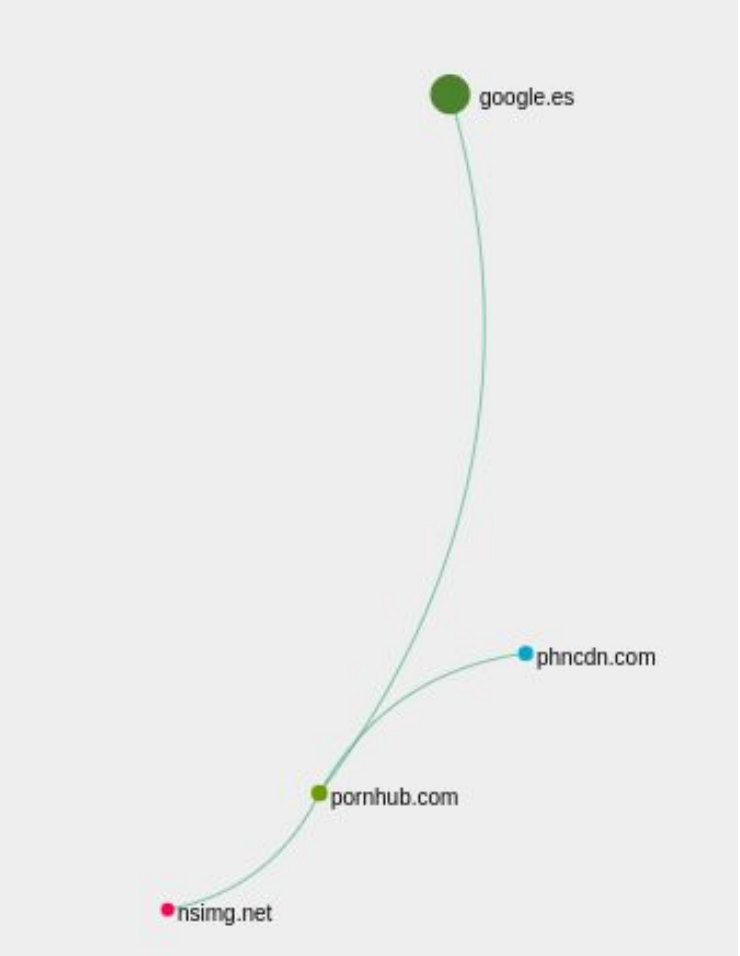
Ocurre algo muy similar a Seriesblanco, utiliza Cloudfare y APIs propias para gestionar el contenido. Pero la página que hemos analizado es la lista de proxies de TPB, **thepiratebay-proxylist.org** y al entrar en uno de estos proxies, **batpirate.info** hemos visto que también enlaza con coinhive.

Al igual que Seriesblanco, TPB usa un cryptominer y en su momento cuando se descubrió se montó un escándalo.



- Pornhub ---> Limpio

Sorprendentemente Pornhub solo accede a dos APIs propias para gestionar sus contenidos y estas no almacenan cookies, es segura.



Conclusión de análisis popular (mapa interactivo)

Tenemos páginas que creíamos seguras y lo son como **Wikipedia** y **Github** y que no lo son del todo como las de la **UCM** y **URJC**.

Las redes sociales sabíamos que no sería del todo seguras, sobre todo según las Third-Parties que tienen acceso a ellas como pasa con **Facebook** y **Forococches**, páginas que a primera vista parecen inofensivas ya que solo ponen un poco de publicidad como **Minijuegos** en realidad envían datos de los usuarios a muchas Third-Parties.

Seriesblanco que pone poca publicidad, sin embargo mina criptomonedas y no se lo indica a sus usuarios. Les acabará pasando como a **TPB** y se descubrirá (no parece que se haya anunciado en ningún sitio y en google no sale nada). Y también tenemos páginas como **Pornhub** de las que podríamos esperar algo extraño pero sin embargo está completamente limpia.

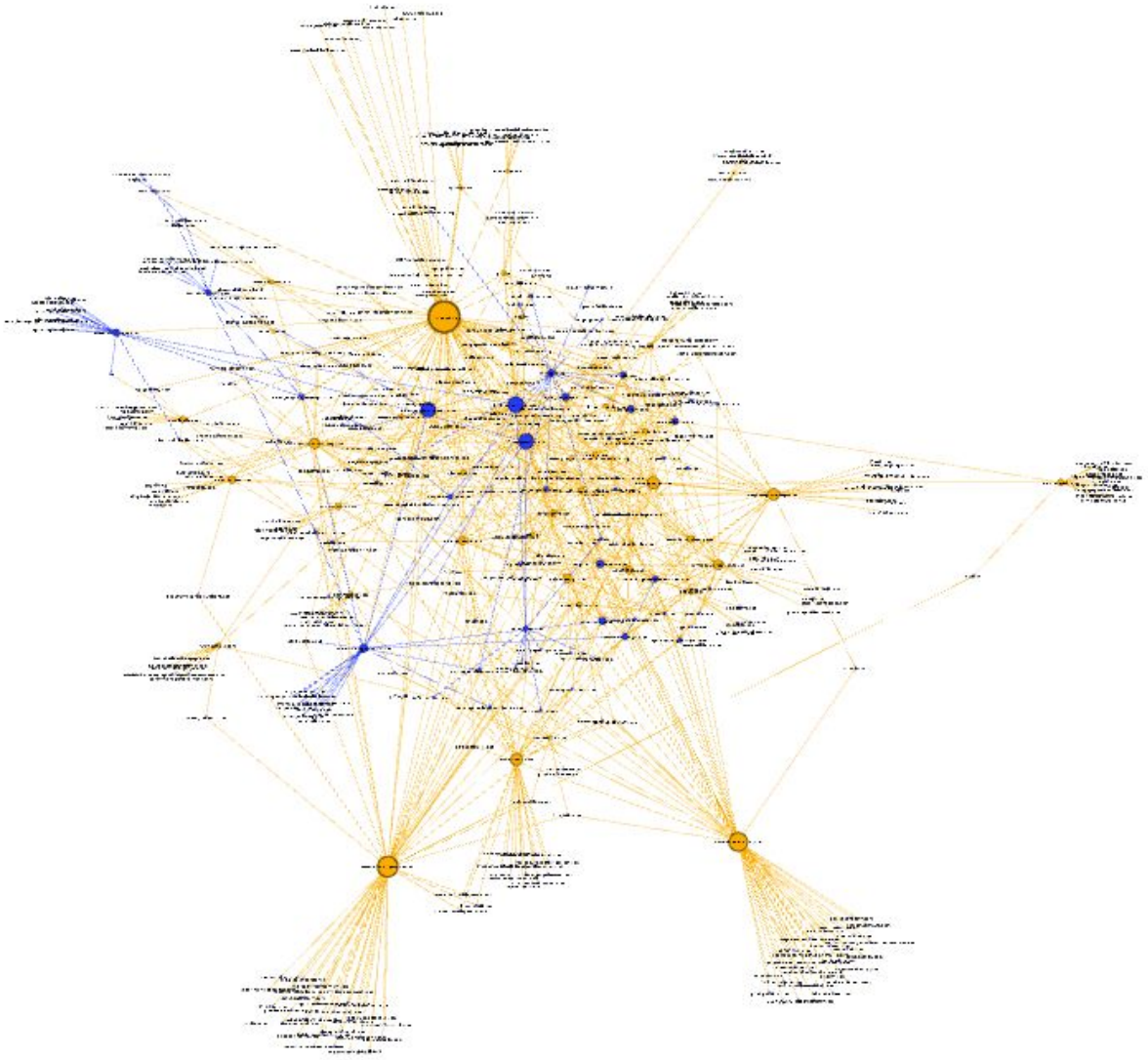
Estudio Navegación de Libre Lab

- Número total de nodos: 482.
- Número total de aristas: 981.

Medidas locales de centralidad

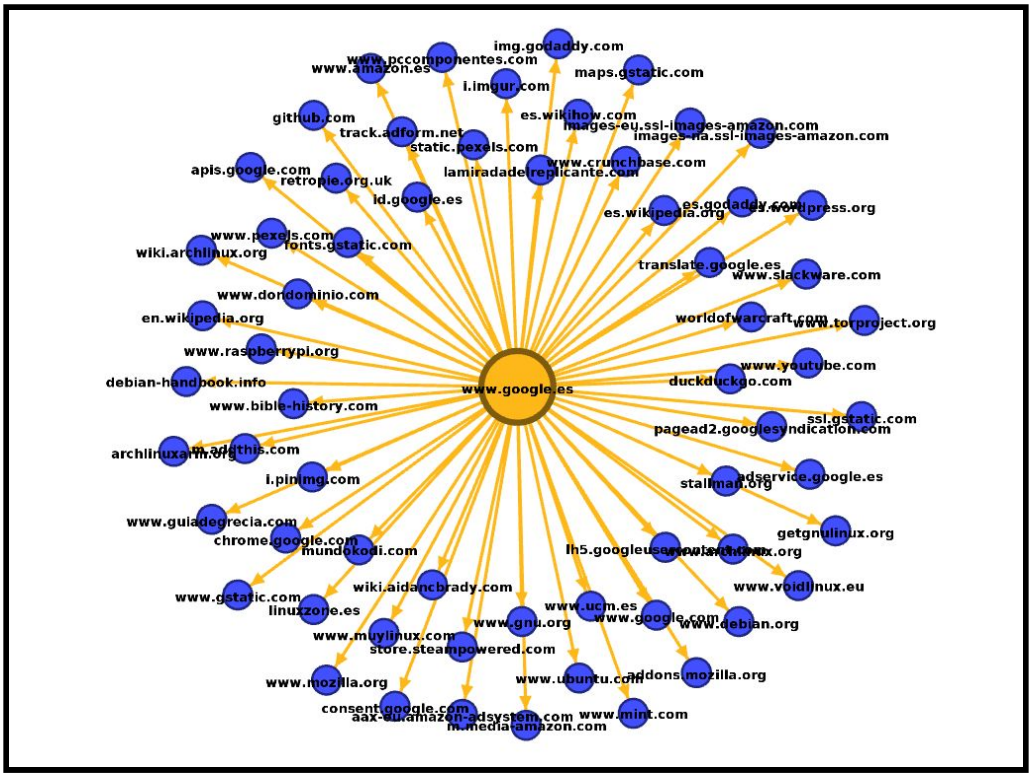
- Grado medio: 2,035
- Diámetro de la red: 6
- Densidad del grafo: 0,004
- Modularidad: 0,539
- Componentes conexos: 22
- Coeficiente medio de Clustering: 0,023
- Longitud media de camino: 2,644

Como podemos observar, este grafo contiene un número mayor de nodos comparado con el generado por la navegación popular, analizado anteriormente.

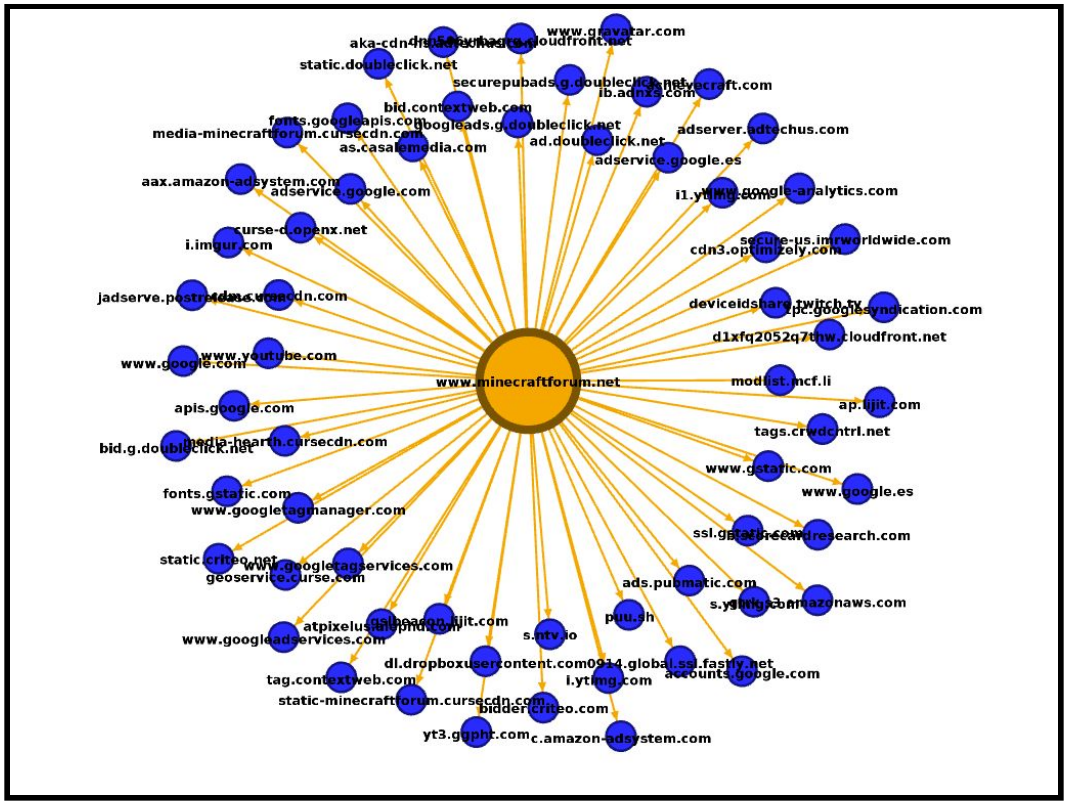


Top 3 de nodos con mayor grado de salida (todos First Parties)

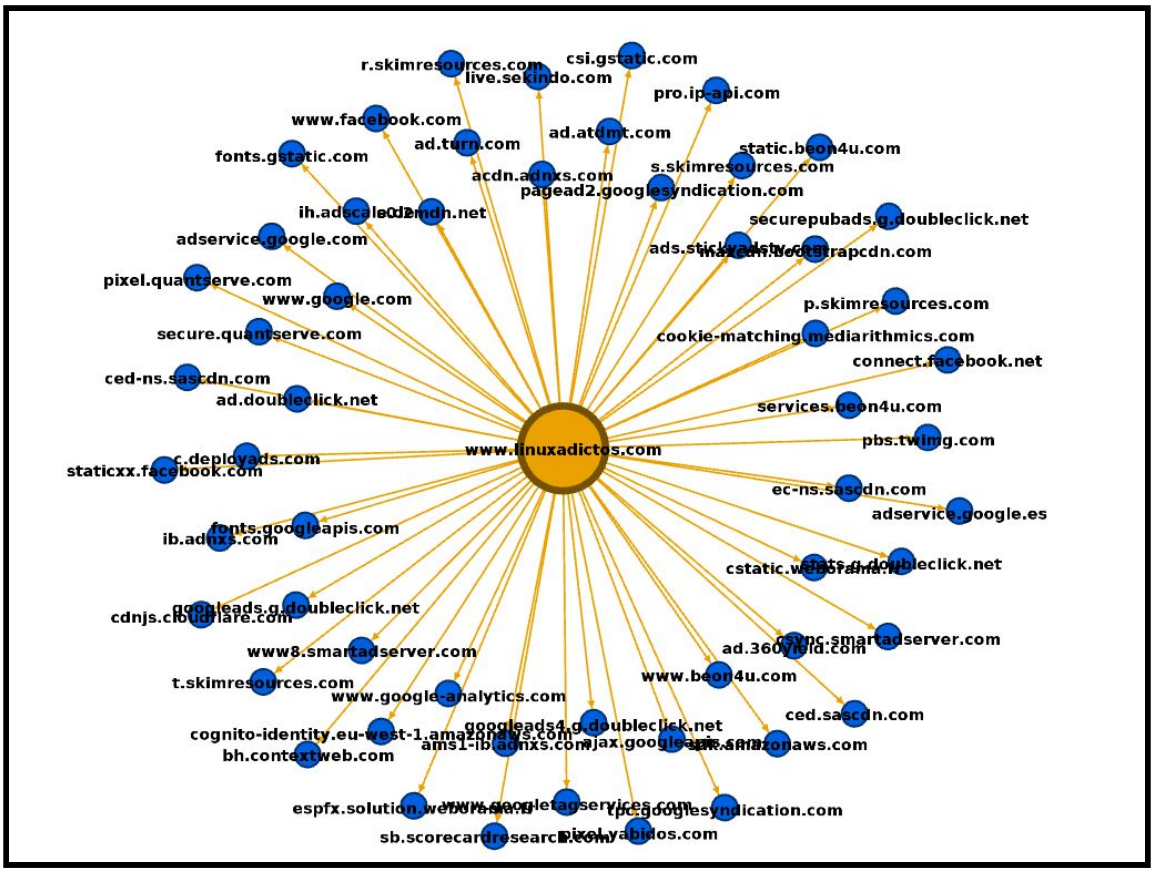
- www.google.es → 63



- www.minecraftforum.net → 62



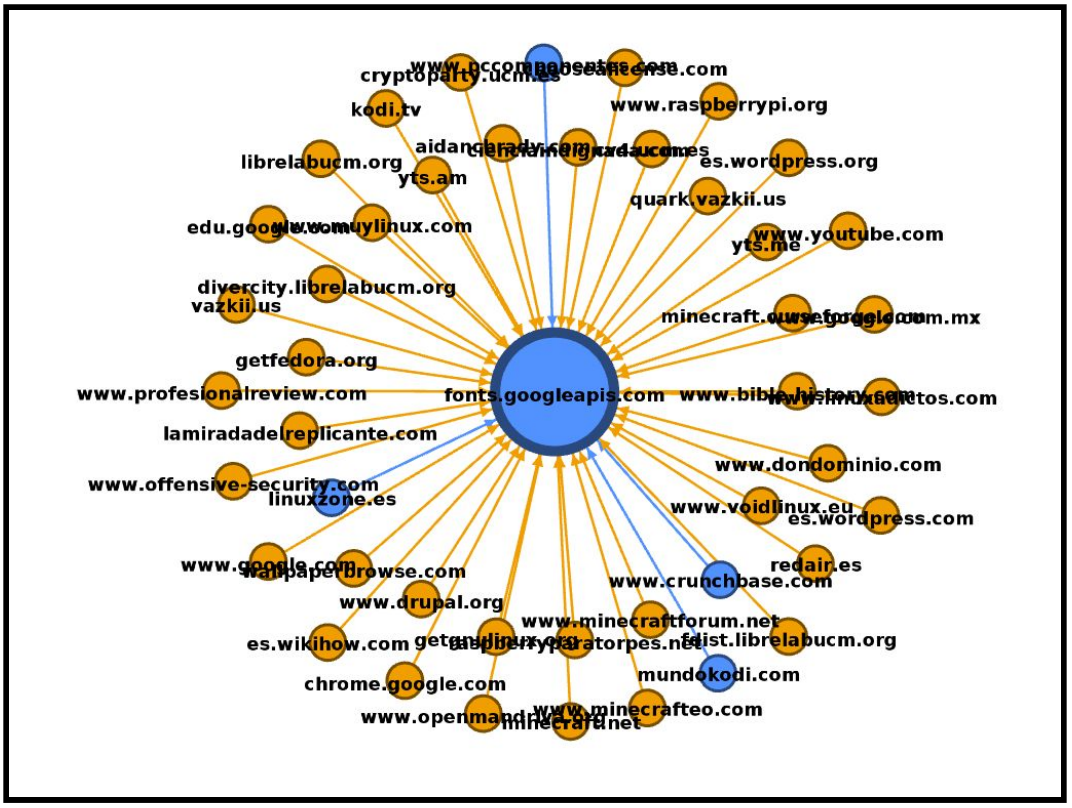
- www.linuxadictos.com → 56



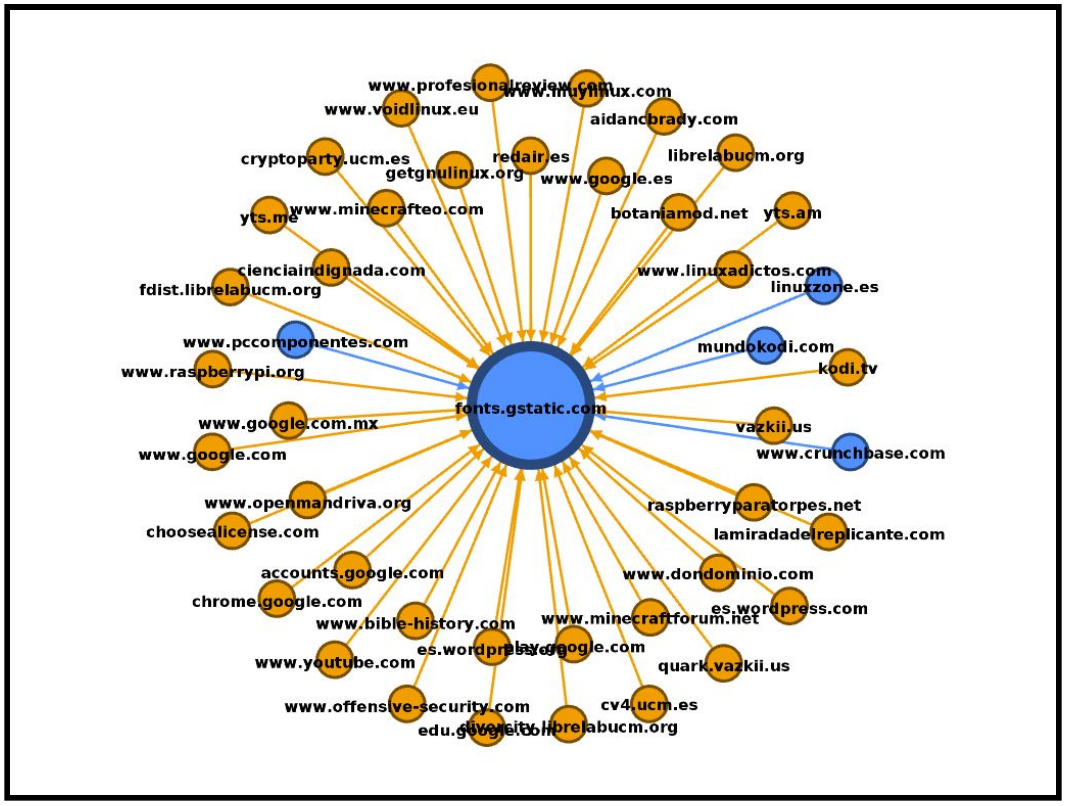
Como podemos apreciar, estos tres nodos son empresas que tienen asociadas muchas Third-Parties a ellas. Por tanto, estos tres hubs son nodos que están mucho más conectados que el resto, y a los que por tanto les llega mucha más información de otras Third-Parties que a las demás empresas que aparecen en el grafo.

Top 3 de nodos con mayor grado de entrada (todos Third Parties)

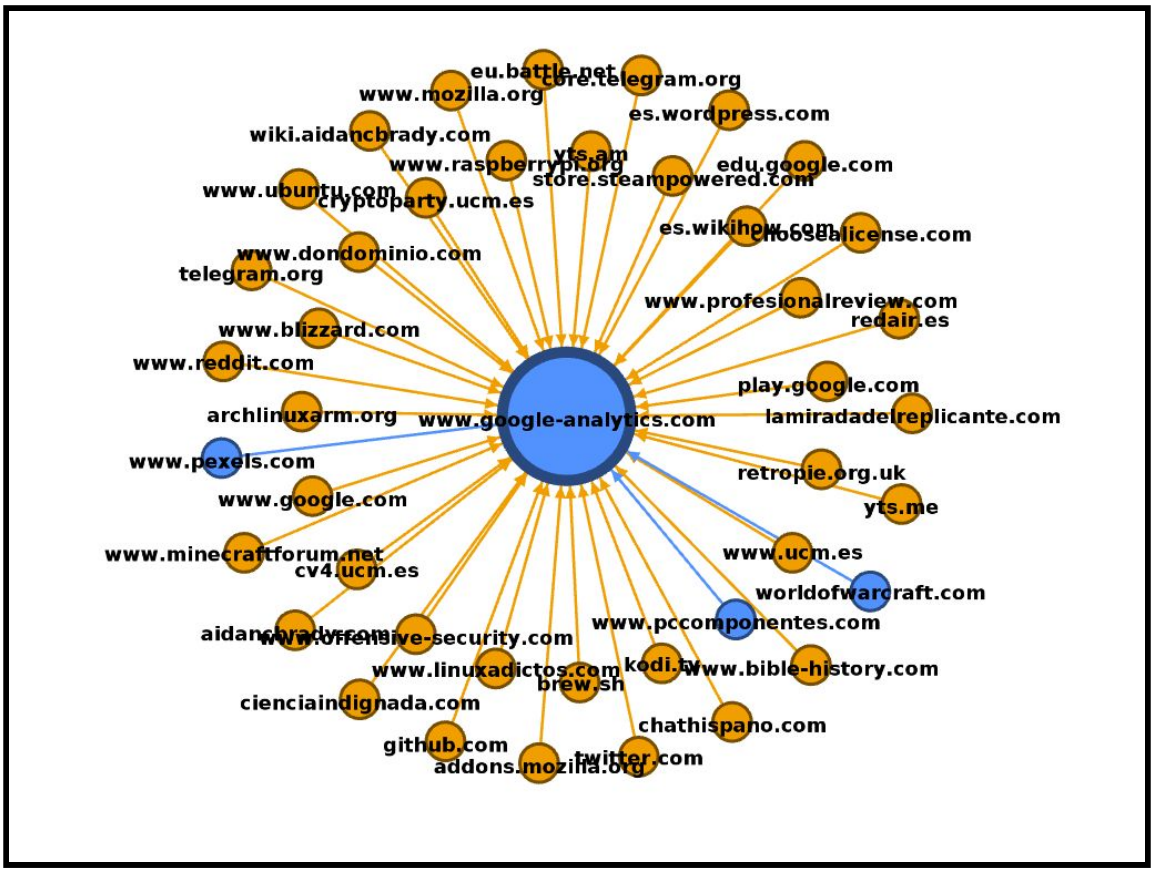
- fonts.googleapis.com → 45



- fonts.gstatic.com → 43



- www.google-analytics.com → 42

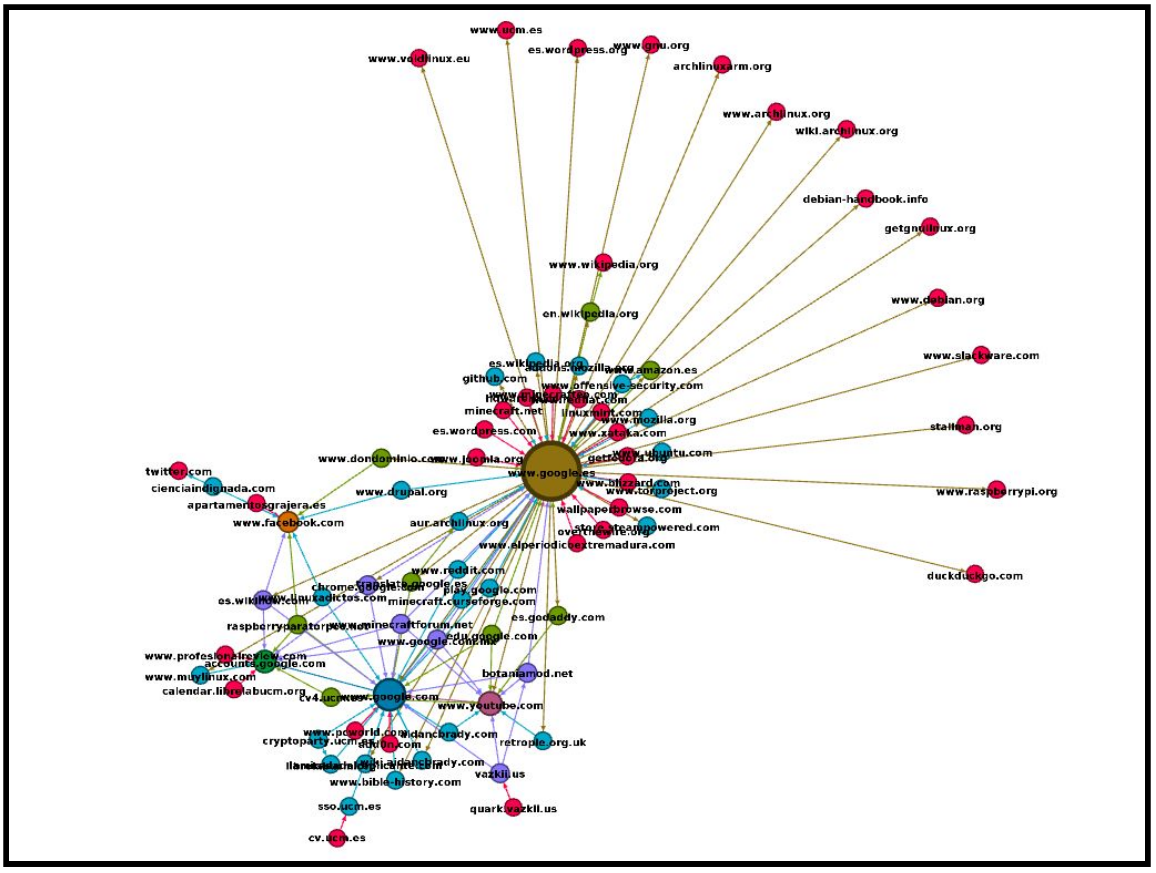


En este caso, podemos apreciar que **las tres que más grado de entrada tienen son Third-Parties**, las cuales proveen mucha información diferentes páginas.

Es normal que tengan un grado de entrada tan alto porque todas son Third-Parties, las cuales van dirigidas desde First-Parties en su mayoría.

Por ejemplo, **si nos fijamos en fonts.gstatic.com**, podemos ver que los otros nodos que están conectados son páginas que seguramente estén usando las fonts de google, y por tanto tengan como un tercero esta página.

Relaciones entre las First-Parties

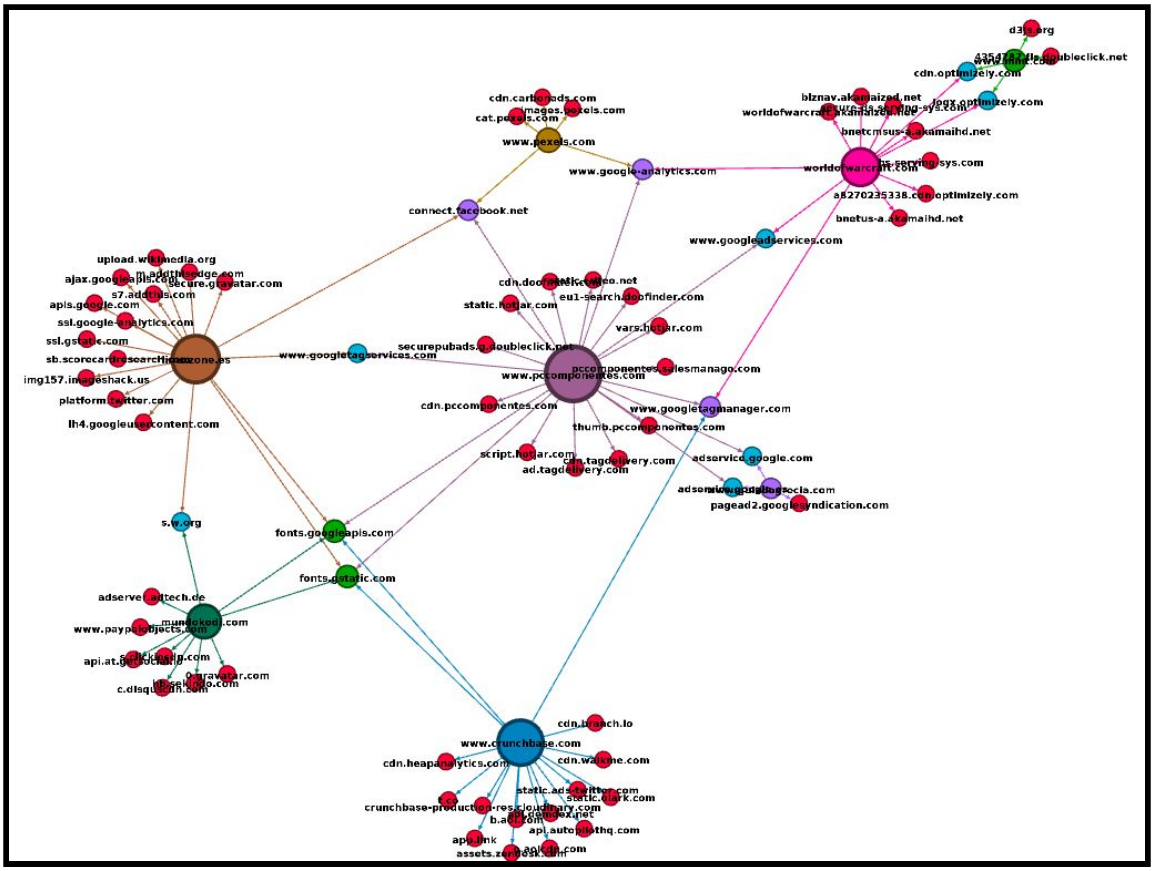


Podemos diferenciar dos hubs principales. Ambos son de google.
Esto se debe a que desde Google se accede a todas las otras páginas (que luego tendrán sus correspondientes Third-Parties).

Por otra parte, si ya estás en una página y buscas otra en el navegador, se te redirige a los resultados de Google y desde ahí eliges la página a la que quieres acceder. Por eso podemos apreciar que muchas páginas vienen y van de Google.

En el caso de aquellas páginas que no van a Google se debe a que ya se había accedido antes desde el navegador y este autocompleta la url.

Relaciones entre Third-Parties



Con este grafo obtenemos las relaciones entre las third parties sin la influencia de las first parties, asi podemos ver más fácilmente cuáles de estas son las que más exponen al usuario.

Podemos visualizar algunos hubs en este grafo, los tres más grandes en cuanto a grado de salida, y por tanto los que más exponen al usuario, son:

www.pccomponentes.com → 21

linuxzone.es → 17

www.crunchbase.com → 16

Top 3 third parties con más grado de entrada	First parties que usan las 3
• fonts.googleapis.com • fonts.gstatic.com • www.google-analytics.com	• www.linuxadictos.com • www.offensive-security.com • minecraft.net • aidancbrady.com • getgnulinux.org • yts.am • edu.google.com • raspberryparatorpes.net • www.youtube.com • www.dondominio.com • es.wordpress.com • www.muylinux.com • www.bible-history.com • lamiradadelreplicante.com • cv4.ucm.es • minecraft.curseforge.com • www.google.com • www.raspberrypi.org • librelabucm.org • www.openmandriva.org • getfedora.org • cryptoparty.ucm.es • chrome.google.com • www.minecraftforum.net • www.profesionalreview.com • cienciaindignada.com • fdist.librelabucm.org • www.google.com.mx • wallpaperbrowse.com • quark.vazkii.us • www.minecrafteo.com • kodi.tv • divercity.librelabucm.org • www.drupal.org • es.wordpress.org • choosealicense.com • es.wikihow.com • www.voidlinux.eu • vazkii.us • yts.me • redair.es

Top 3 third parties con más grado de salida	First parties que usan las 3
• www.pccomponentes.com • linuxzone.es • www.crunchbase.com	• www.google.es

3. Interpretación de los datos

3.1. Interpretación de resultados y conclusiones relevantes

Navegación popular

Para iniciar el estudio de esta red, se han aplicado ciertas métricas explicadas anteriormente. En este apartado pasaremos a explicar qué implican los resultados obtenidos.

Estudio del grado de entrada → Soporte

Realizando un estudio del grado o prestigio de entrada de los nodos del grafo general de la red (los nodos representan tanto a First Parties como a Third Parties, siendo la página de origen la que activa a la página de destino); se ha obtenido un listado de nodos con el valor del prestigio de entrada calculado.

Teniendo en cuenta que los nodos a estudiar cumplen que si un nodo tiene un grado de entrada es una Third Party (ha sido activada por una página de origen, ya fuera una First Party u otra Third Party), nos ha parecido interesante recalcar cuales son los nodos con un mayor valor,es decir, los **más prominentes**. Aquí mostramos el top 3 de las páginas con un mayor de entrada:

1. doubleclick.net: 14
2. google-analytics.com: 9
3. mathtag.com: 8

Analizando estas páginas podemos ver que se cumple un patrón: estas tres páginas son pertenecientes a empresas de publicidad que ofrecen tanto anuncios como gestión de datos adquirida de información del tráfico que llega a los sitios web de origen. También es destacable (aunque no sorprendente) que las dos páginas que reciben más información de otras páginas pertenecen a Google. Si hacemos una vista más amplia y analizamos las 10 páginas más llamadas, sigue estando a la cabeza Google junto a Facebook y Oracle a través de BlueKai.

Estudio del grado de salida → Influencia

En cuanto al grado o prestigio de salida, este parámetro marca la cantidad de páginas de terceros que tiene activas ese nodo (página de origen). Las tres páginas con un mayor valor de prestigio de salida, es decir, las **más influyentes** son:

1. minijuegos.com: 45
2. smartadserver.com: 22
3. uax.es: 18

Estas páginas difieren mucho en cuanto al servicio que ofrecen, pero todas ellas utilizan una gran cantidad de páginas de terceros para análisis y almacenamiento de datos.

Hemos investigado un poco en la información que proporciona la página Minijuegos al ser la maś polémica en cuanto al uso masivo de este tipo de políticas de cookies. En su página informan a los usuarios del uso de cookies, de los tipos que hay y enumeran a los terceros que usan para el almacenamiento. Pero la lista se les queda corta, pues hemos detectado otras muchas Third-Parties (sobre todo relacionadas con publicidad) que no mencionan.

A su vez, Smartadserver es una página de una compañía de publicidad. En nuestro análisis está clasificada como una Third Party, pues es usada por otras páginas para políticas de marketing; pero a su vez mejora sus servicios mandando información a otras páginas.

Dentro de este estudio nos ha sorprendido la página de la Universidad Alfonso X el Sabio. Tienen un apartado de aviso legal en el que, entre otras cosas, hacen una breve explicación del uso de cookies. Cuando explican las cookies de terceros que utilizan, solo mencionan a Google Analytics; pero nosotros hemos encontrado también que tienen activas tanto a Facebook como a Youtube y a Mathtag (una compañía que hemos analizado en el apartado del grado de entrada).

Estudio del grado medio

El cálculo del grado medio para el grafo general de la red da un valor de 1,77.

Los nodos con un mayor grado en total (suma del grado de entrada y de salida) son:

1. Minijuegos.com: 45
2. Smartadserver.com: 25
3. Mathtag.com: 23

Esto nos da una idea global de los sitios más activos en cuanto a relaciones entre páginas web.

Además, de un estudio visual del gráfico podemos concluir que la mayoría de nodos tienen un grado 1, siendo solo una minoría de ellos los que tienen un grado superior.

Estudio de Intermediación

Tras realizar el cálculo de la intermediación de los nodos del grafo general de la red, las páginas web más centrales o con un mayor grado de intermediación son:

1. Doubleclick.net: 699,33
2. Smartadserver.com: 616,83
3. Google.es: 372,83

Este es un parámetro muy interesante de analizar en nuestra red, pues nos muestra los nodos más **centrales** y que tienen un rol más predominante en la red. Traducido al tema sobre el que gira nuestro estudio, nos lleva a la conclusión de que estas tres páginas son las que más información manejan dentro del rango de navegación de la asociación de LibreLabUcm. Estas páginas comparten además que las tres tienen un rol de third party (son activadas por la página de origen). Aunque Google.es en nuestro estudio aparece como first Party porque ha sido visitada por los usuarios, también se comporta como una página de tercero para muchas de las páginas visitadas (Youtube, Facebook, Minijuegos...).

Estudio de Excentricidad

El cálculo de excentricidad de la red general nos devuelve la máxima distancia geodésica entre cada uno de los nodos y el resto de los nodos de la red. Las tres páginas con un mayor valor son:

- Google.com: 6.0
- Addthis.com. 6.0
- Tunnl.com: 6.0

Estas páginas son las más **periféricas** de la red.

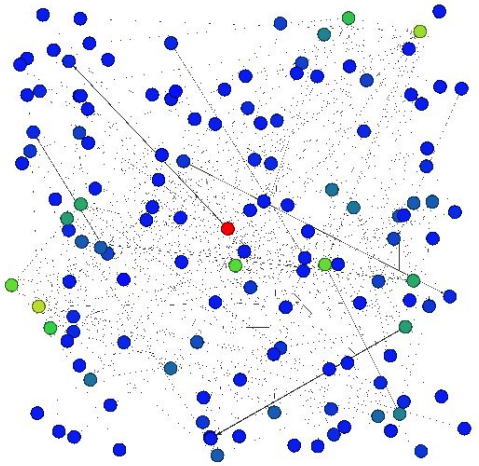
Estudio de la Centralidad de vector propio

Basándonos en el resultado de la medida de Centralidad de vector propio (CPV), observamos que los nodos más centrales son:

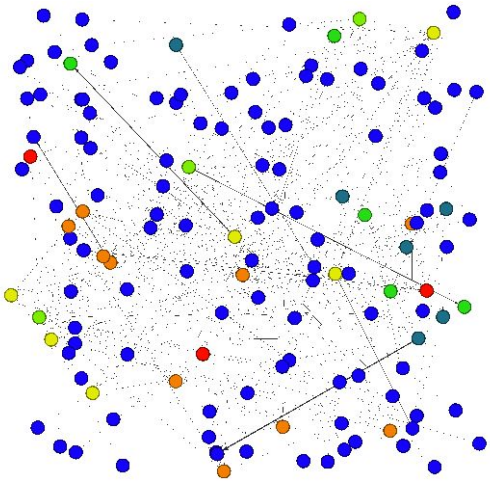
1. Doubeclick.net: 1.0
2. Gstatic.com: 0.56
3. Bluekai.com: 0.55

Esta centralidad toma como base la centralidad de sus vecinos, y por ello estos nodos representan a las páginas con más poder o **prominencia**.

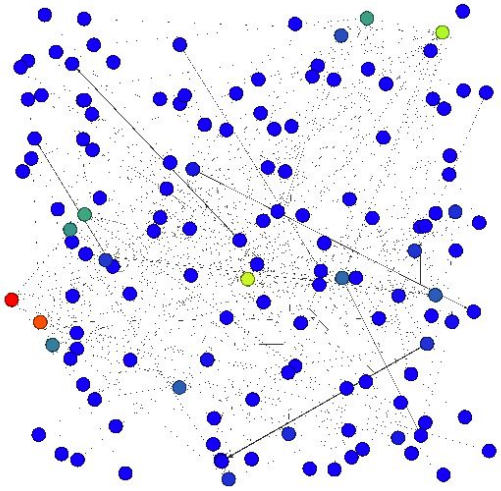
Comparativa de las medidas locales de centralidad



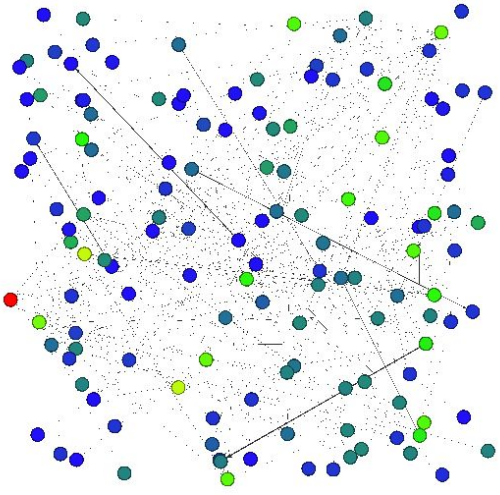
Centralidad de grado



Excentricidad



Intermediación



Centralidad de vector propio

Estudio del PageRank

El cálculo del PageRank nos mide la **relevancia** de las diferentes páginas web teniendo en cuenta tanto las páginas web que apuntan a una en concreto, como las páginas que apuntan a las anteriores. Es una variante del cálculo de centralidad propia, y los nodos con un mayor valor son:

1. Gstatic.com: 0.026
2. Doubleclick.net: 0. 22
3. Forms.googleapis.com: 0.020

Remarcando el tipo de páginas que son, observamos que se tratan de 3 páginas que actúan como Third Parties. El tener un alto valor de PageRank aporta importancia a estas páginas, que a su vez se lo transmiten a sus nodos vecinos. Se comprueba como algunas de las páginas son similares a las calculadas con la centralidad de vector propio:

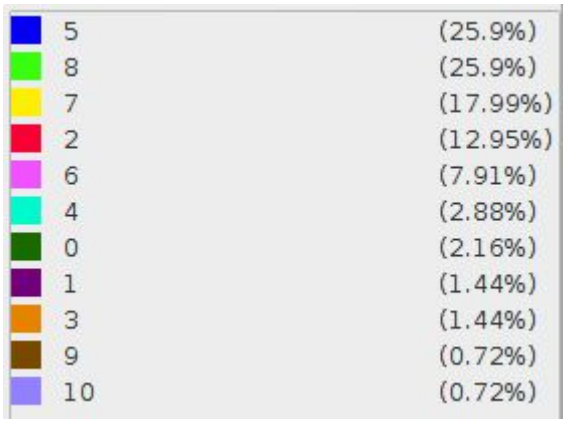
Id	Eigenvector Ce...
doubleclick.net	1.0
gstatic.com	0.55825
bluekai.com	0.550263
smartadserver.com	0.43723
mathtag.com	0.420101
addthis.com	0.413078
googlesyndication.com	0.397362
bidswitch.net	0.381752
spotxchange.com	0.381528
weborama.fr	0.381528

Id	PageRank
gstatic.com	0.026585
doubleclick.net	0.021927
fonts.googleapis.com	0.020663
google-analytics.com	0.014436
bluekai.com	0.013049
addthis.com	0.012642
addthisedge.com	0.011921
wikimedia.org	0.010716
facebook.net	0.010712
googlesyndication.com	0.010119

Estudio de la Modularidad

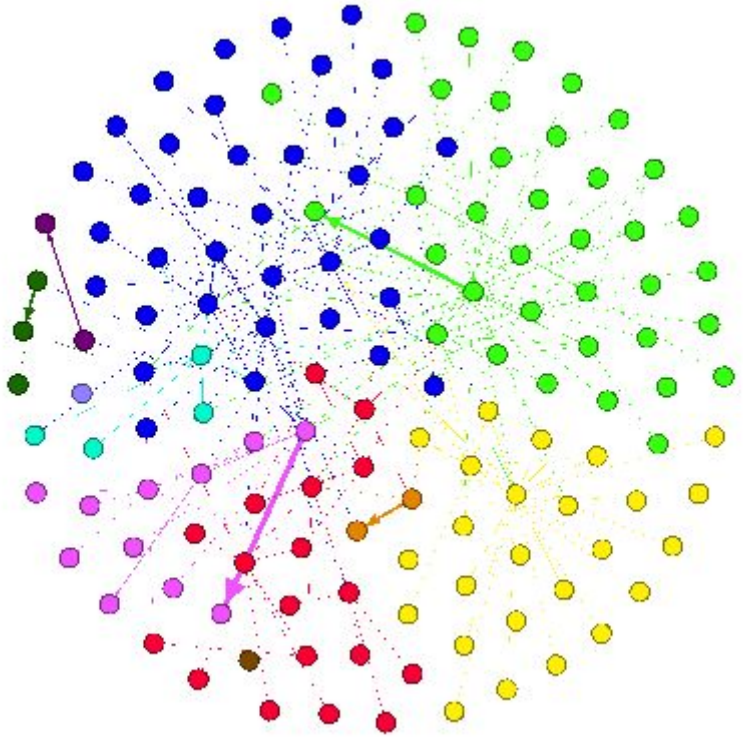
Habiendo calculado la modularidad al grafo general de la red, se ha obtenido un valor de modularidad general de 0.734. Este número, al ser positivo, refleja que el número de enlaces dentro de los grupos supera el número esperado sobre la base de pura casualidad.

Se han formado 11 comunidades, que ordenamos a continuación por población:



Visualizaciones del grafo por Comunidades

Si aplicamos al grafo un layout de tipo Fruchterman Reingold, podemos visualizar claramente el grado de población de los diferentes grupos:



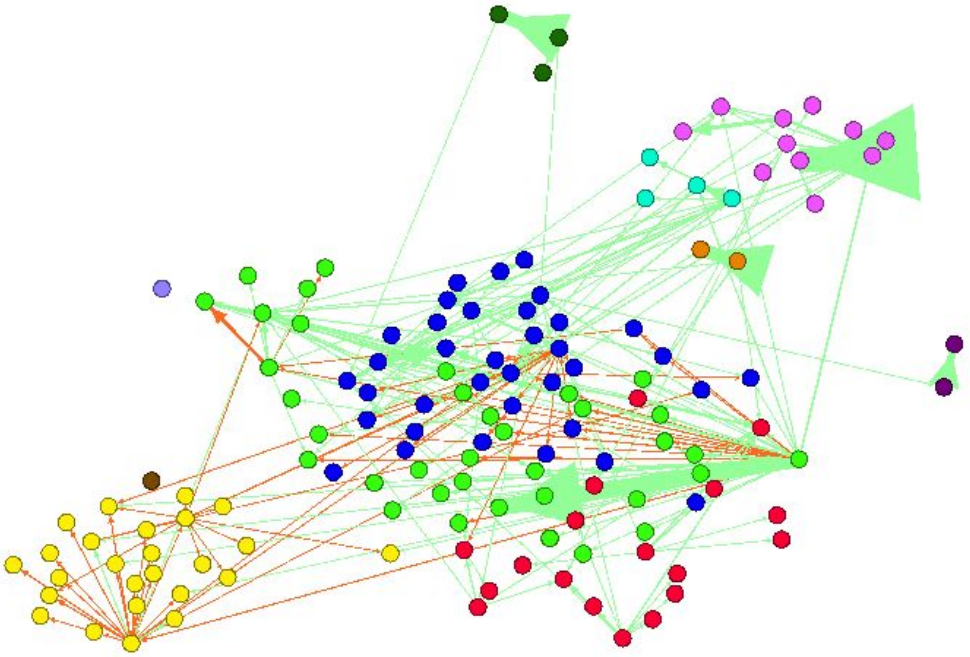
Se puede observar claramente como la Comunidad 5 (azul oscuro) es la más poblada con 36 nodos; y las Comunidades 10 y 9 las menos pobladas con solo un nodo.

Adentrándonos en los nodos que componen cada comunidad, podemos apreciar cómo cada grupo se genera entorno a una First Party visitada o una Third Party con un grado de centralidad muy alto. Aquí mostramos un listado de las agrupaciones más significativas:

- Comunidad 5 (azul): se identifica con la página web Google.es como principal, aunque también incluye a Forocoches.com.
- Comunidad 8 (verde): se identifica con la página web Minijuegos.com.
- Comunidad 7 (amarillo): se identifica con la página web Smartadserver.com.
- Comunidad 2 (rojo): se identifica con la página web Github.com y Ucm.es.

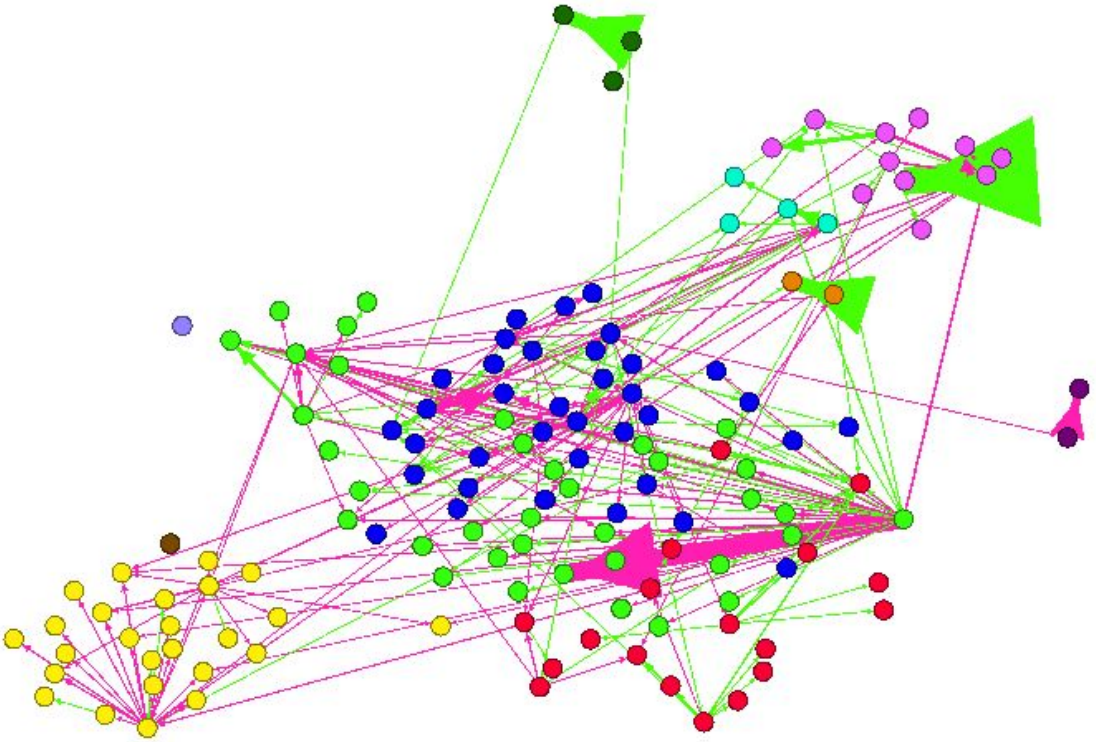
También es importante destacar como, si resaltamos el color de las aristas según si las conexiones son seguras o no (aristas verdes si se utiliza el protocolo https y aristas rojas si utiliza el protocolo http), se pueden distinguir las comunidades por su seguridad.

Para realizar esta vista se ha aplicado el layout OpenOrd, y aplicando a las aristas una partición por color según la variable secure, obtenemos:



Con esto se detecta que la mayoría de Comunidades son seguras, identificandose claramente a la comunidad 7 como la más conflictiva. Esto no nos sorprende, al haber explicado anteriormente que es una de las comunidades que , a la vez que poblada, gira entorno a una Third Party.

Si por otro lado aplicamos a las aristas una partición por color según la variable cookies, siendo aristas verdes las conexiones que no almacenan cookies y rosas las que si, obtenemos:



En este caso también destaca la comunidad 5 como una de las que mueve más cookies, aunque es un comportamiento mucho más extendido entre todas las páginas web de la red.

Además, interesa identificar al nodo de Minijuegos.com como el que más aristas de cookies genera (análisis que coincide con el realizado al estudiar el prestigio de salida).

Navegación de LibreLabUCM

Para el estudio de la segunda red nos hemos centrado en visualizar y estudiar desde Gephi los tres nodos con mayor grado de salida y de entrada y sus vecinos.

Estudio del grado de entrada → Soporte

Al calcular el grado de entrada para el grafo general de nodos, hemos obtenido esta lista con las páginas más **prominentes**:

1. fonts.googleapis.com: 45
2. fonts.gstatic.com: 43
3. www.google-analytics.com: 42

Lo más destacable de este estudio de la red es la predominancia de Google como entidad que más información recibe y que está activa en el mayor número de páginas web.

Fijándonos en el tipo de páginas que tienen activos a estos nodos, podemos concluir que son todo First Parties. Este tipo de páginas suelen buscar la gestión que ofrece Google tanto para el número de visitas como para realizar posteriormente un control de los usuarios interesados y que tener en cuenta como potenciales clientes/visitas.

Estudio del grado de salida → Influencia

También hemos creído oportuno generar grafos a partir de cada uno de los nodos dentro del top 3 de páginas con un mayor prestigio de salida::

1. www.google.es: 63
2. www.minecraftforum.net: 62
3. www.linuxadictos.com: 56

En este caso la lista sale mucho más variada en cuanto a las funciones que ejercen las páginas (google como buscador a la cabeza y dos páginas de entretenimiento).

Aún así todas ellas cumplen el patrón de buscar en páginas que actúan como Third Parties (la mayoría de Google) una gestión de sus usuarios y de las visitas que reciben, tanto para potenciar su popularidad (futuro plan de marketing) como para fomentar su financiación a través de publicidad.

3.2. Limitaciones encontradas en el análisis

Lamentablemente, una de las limitaciones que hemos tenido a la hora de extraer los datos ha sido la versión de firefox. Sólo un miembro del grupo tenía la versión antigua de firefox (45.9.0.) que permite que Lightbeam obtenga más información de las páginas. La versión que conseguimos poner en la asociación Libre Lab era la (57.0.4). Si hubiésemos conseguido más de un ordenador con la versión antigua de firefox, podríamos haber comparado ambas redes y haber obtenido más información.

3.3. Comparativa con modelos de redes estudiados

Vamos a comparar nuestras redes generales (Popular y Librelab) con los modelos vistos en clase. Estos modelos son el modelo de red aleatoria y el modelo de red de libre escala.

Navegación Popular

N	L	<k>	ln N	<c>	<d>
139	246	1,77	4,93	0,09	2,859

Comparación con el modelo de red aleatoria

Como la definición de red aleatoria establece que es una red donde cada enlace entre dos nodos se ha creado siguiendo un proceso aleatorio (con una probabilidad p), suponemos que puede hacerse una comparación tanto con grafos no dirigidos como con grafos dirigidos.

De acuerdo a los datos de nuestra red, se cumple que <k> > 1 por lo que tendría una componente gigante. No se observa una componente gigante significativa en el grafo de esta red. Sin embargo, en esta red al ser <k> menor que lnN según el modelo de red aleatoria, significaría que se encuentra en la etapa supercrítica y que entonces existen nodos y componentes aislados. Nunca llegaría a estar en la parte conectada puesto que el grado medio <k> no puede llegar a ser mayor o igual que el ln N. Esta red se encontraría siempre en la fase supercrítica.

Usando como referencia la distribución de Poisson, se tendría que observar que la mayoría de los nodos tienen un grado entorno a la media <k> y que los nodos de mayor grado tienen sólo unos pocos más que la media. Las redes aleatorias tampoco cuentan con concentradores o hubs dado que la probabilidad de que un nodo tenga el grado mucho mayor que los demás es muy baja. Esto difiere mucho de los resultados de nuestra red, ya que la media de grados sale 1,77 y hay nodos que tienen grado 45 de salida y hasta grado 14 de entrada, teniendo un máximo total de grado Gin + Gout = 45, muy alejado de la media.

En cuanto a el coeficiente de clustering, según el modelo de red aleatoria, decrece cuanto mayor es el tamaño de la red, con una razón de 1/N. Además es independiente de su grado. En nuestra red, se puede observar en el laboratorio de datos, que el coeficiente de clustering Ci si que es dependiente del grado, y que los nodos con grado más alto tienen un coeficiente de clustering mayor.

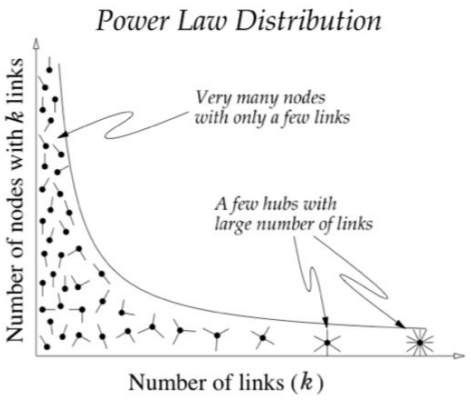
La longitud media de camino según el modelo de red aleatoria se calcula como <d> = Log N/log <k>. Resolviendo esta ecuación saldría una longitud media de camino de 8.64 mientras que nuestra red tiene una <d> = 2.859. Por lo tanto aunque muchas redes reales se aproximen a la red aleatoria en este aspecto, la nuestra no es uno de esos casos. Si consideramos la misma fórmula de la red aleatoria de <d> que según el modelo de red aleatoria también sirve para calcular el diámetro (dmax) sí que se parecerían más los datos, dado que el diámetro de la red (sin normalizar centralidades) es 6, que se acerca más a 8.64.

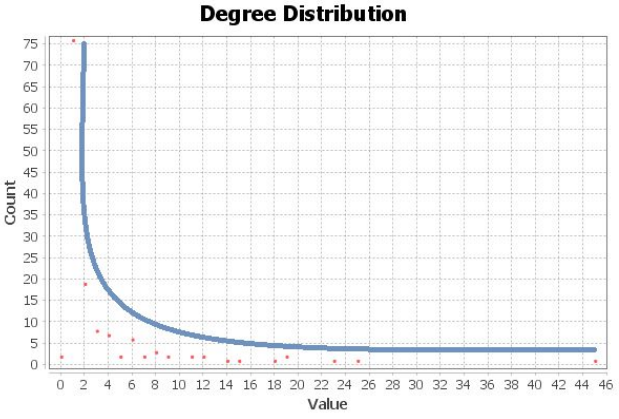
Por lo tanto podemos concluir que nuestra red no se aproxima al modelo de red aleatoria.

Comparación con el modelo de red de libre escala

El modelo de libre escala, contempla una red cuya distribución de grados sigue una ley potencial. Al contrario que el modelo de red aleatorio, este modelo contempla la existencia de hubs, que se pueden observar claramente en nuestra red. Por ejemplo en esta red minijuegos.com tiene un valor de hub de 0.75 estando mucho más conectado que el resto. Además la red de libre escala presenta una mayor probabilidad de encontrar nodos tanto de grados muy altos como de grados muy pequeños. Esto se ajusta perfectamente a nuestra red, puesto que contiene ambos casos.

Sobre la distribución de grados nuestro grafo presenta una distribución de cola ancha, es decir nodos de grado bajo conviven con nodos de grado muy alto (hubs). Si comparamos la gráfica que nos da gephi con una gráfica que representa una red de libre escala, vemos que el comportamiento es el mismo.





Además nuestra red cumple la propiedad de los mundos pequeños puesto que la distancia media de camino salía con un valor mucho menor que el propuesto por la red aleatoria. Además ésta se incrementa en un orden menor que el logaritmo cuando el tamaño de la red aumenta.

El modelo de crecimiento de Barabasi-Albert establece que los nodos nuevos prefieren conectarse a los nodos más conectados. En el caso de nuestra red, esto podría ser posible pero no se tiene porque cumplir siempre. Por ejemplo, una nueva third party que quiere extraer información, tendrá la intención de conectarse con la first party de mayor grado de salida, dado que es la que más third parties tiene conectadas y por lo tanto probablemente sea la first party más visitada. También puede ser que una third party quiera extraer información de un

sitio particular por ciertas razones, el enlace no tiene por qué ser con la first party de mayor grado.

Por otro lado, nuestra red si que crece a lo largo del tiempo, expandiéndose e incorporando nuevos nodos y enlaces cada vez que el usuario estudiado visita nuevas páginas.

Comparando las métricas de nuestra red con las métricas del modelo de Barabasi-Albert, vemos que la distancia media de camino <d> = log N / ln(lnN) es igual a 1.342. Este valor se aproxima mucho más al valor de nuestra red.

Por lo tanto, podemos concluir que nuestra red es mucho más parecida al modelo de red de libre escala. Aún así tenemos que tener en cuenta que es necesario para generar una red de libre escala que esta esté en crecimiento y haya conexión preferencial. Si en se diera el caso comentado anteriormente en el cual la red no siguiese una conexión preferencial, perdería semejanza con el modelo de libre escala.

Navegación de LibreLab

La comparación de esta navegación es mucho más breve dado que sólo comentaremos las métricas y resultados relevantes (las dos redes se parecen mucho, han sido obtenidas de las misma manera)

N	L	<k>	ln N	<c>	<d>
482	981	2,035	6,178	0,023	2,644

Comparación con el modelo de red aleatoria

Si comparamos esta red con el modelo de red aleatoria, podemos ver que se cumple también que <k> > 1 pero que a su vez es menor que el lnN por lo que según el modelo de red aleatoria se encuentra también en la etapa supercrítica, por lo que existen nodos y componentes aislados.

En esta red tampoco se cumple que los nodos tengan un grado entorno a la media <k> ni que los nodos de mayor grado tengan sólo unos pocos más que la media. En esta red hay nodos con un valor de grado de entrada de 45 y de grado de salida de 63, valores bastante alejados de la media.

En cuanto a el coeficiente de clustering, según el modelo de red aleatoria, decrece cuanto mayor es el tamaño de la red, con una razón de 1/N. Además es independiente de su grado. En el laboratorio de datos cuando evaluamos esta red, observamos que en cierto modo es independiente del grado. Esta red es muy densa, y el coeficiente de clustering más alto (valor 1) lo tiene un nodo con grado 2.

La longitud media de camino según el modelo de red aleatoria se calcula como <d> = Log N/log <k>. Resolviendo esta ecuación saldría una longitud media de camino de 8.6 mientras que

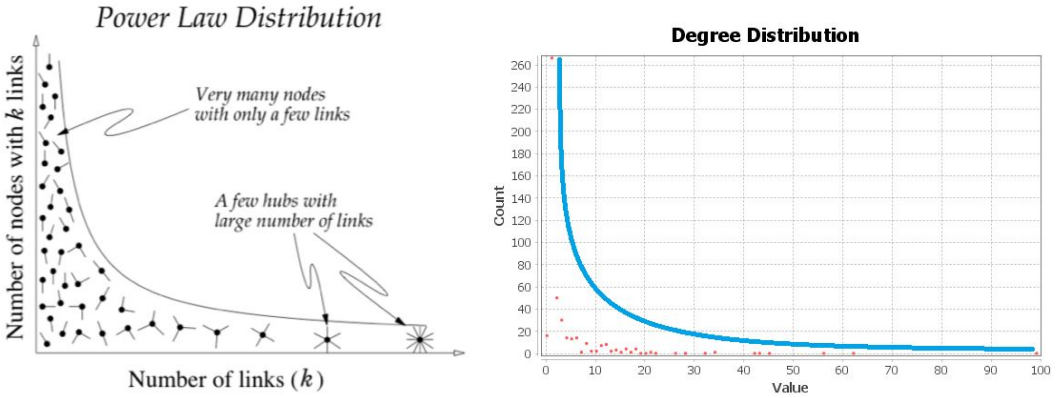
nuestra red tiene una <d> = 2.64. Esta red y la red “popular” tienen una diferencia de décimas entre estos valores.

Por lo tanto podemos concluir que esta red tampoco se aproxima al modelo de red aleatoria.

Comparación con el modelo de red de libre escala

Al comparar la red generada por la navegación de librelab con el modelo de red de libre escala podemos observar que también presenta una mayor probabilidad de encontrar nodos tanto de grados muy altos como de grados muy pequeños.

Sobre la distribución de grados nuestro grafo también presenta una distribución de cola ancha, es decir nodos de grado bajo conviven con nodos de grado muy alto (hubs). Si comparamos la gráfica que nos da gephi con una gráfica que representa una red de libre escala, vemos que el comportamiento es el mismo.



Como la red anterior, esta también cumple la propiedad de los mundos pequeños y difiere en la misma proporción con los valores de la distancia media de camino propuestos por el modelo de red aleatoria.

Comparando las métricas de esta red con las métricas del modelo de Barabasi-Albert, vemos que la distancia media de camino <d> = log N / ln(lnN) es igual a 1.47 que se aproxima mucho más al valor de nuestra red.

Esta red también crece a lo largo del tiempo e incorpora nuevos nodos y enlaces cada vez que un usuario de la asociación visita nuevas páginas. También como comentamos en el caso anterior, es muy probable que se dé una conexión preferencial.

Por lo tanto, podemos concluir que esta red también es más parecida al modelo de red de libre escala que al modelo de red aleatoria.