CLIQZ Anti-Tracking

Whitepaper, February 2016

CLIQZ IS ...

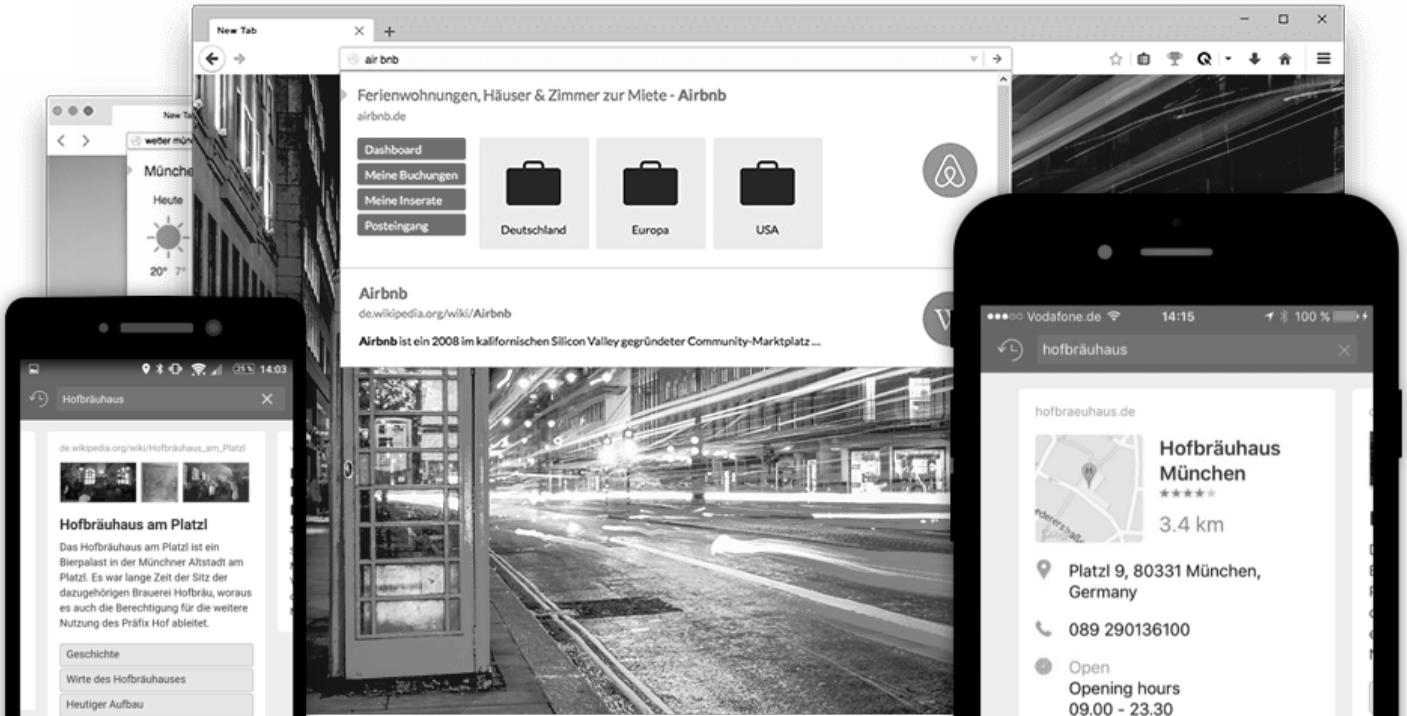
... a proprietary, radically new search engine.

... directly integrated into the browser.

… available for all major platforms (iOS, Android, Mac, Windows).

… developed by a multi-national team of 92 digital enthusiasts in Munich.

… pursuing a re-design of the internet for our users.



Tracking has various legitimate use-cases for websites, but the users' privacy should always be protected

TRACKERS CAN FULFIL VERY LEGITIMATE USE CASES FOR WEB-SITES AND PLAY AN IMPORTANT PART IN THE INTERNET ECO-SYSTEM

Exemplary use cases for trackers

- Site performance analytics
- On- and offsite user journey and conversion tracking
- Sales and basket optimization
- Advertising
- Audience measurement
- Content adaption
- Additional functionalities (e.g. sharing content)

and many more …

- Can play an important role for web-sites.
- Do not necessarily, by default invade the users' privacy.
- **CLIQZ tries to support these use-cases as long as the user's privacy is not at risk.**

HOWEVER SOMETIMES THE IMPLEMENTATION OF TRACKERS INTRODUCES UNDESIRED SIDE-EFFECTS FOR USERS AND WEB-SITES

Exemplary implementations of trackers

- Often outsourced to 3rd parties (e.g., for measurement, convenience, …)
- Often loaded as part of backfill advertisement (and often without the publisher knowing)
- JavaScript code executed on the user‘s browser
- Information in HTTP-referrer sent to site owner and tracking company

```
bklc=55f6ad4d
l=https://www.DOMAIN.de/
ua=f82610bef1d54776cde605b90b0c7949
t=1444203542439
m=020810a3483fc8307caa483fd192bc02
lang=07ef608d8a7e9677f0b83775f0b83775
sr=1440x900x24
cpu=4b4e4ecaab1f1c93ab1f1c93ab1f1c93
platform=6d44fad93929d59b3929d59b3929d59b
plugins=d4de4a68c91685d0ff4838ce3714359a
cn=df62ddfcfa96f717f2ee5a7d912e7102
```

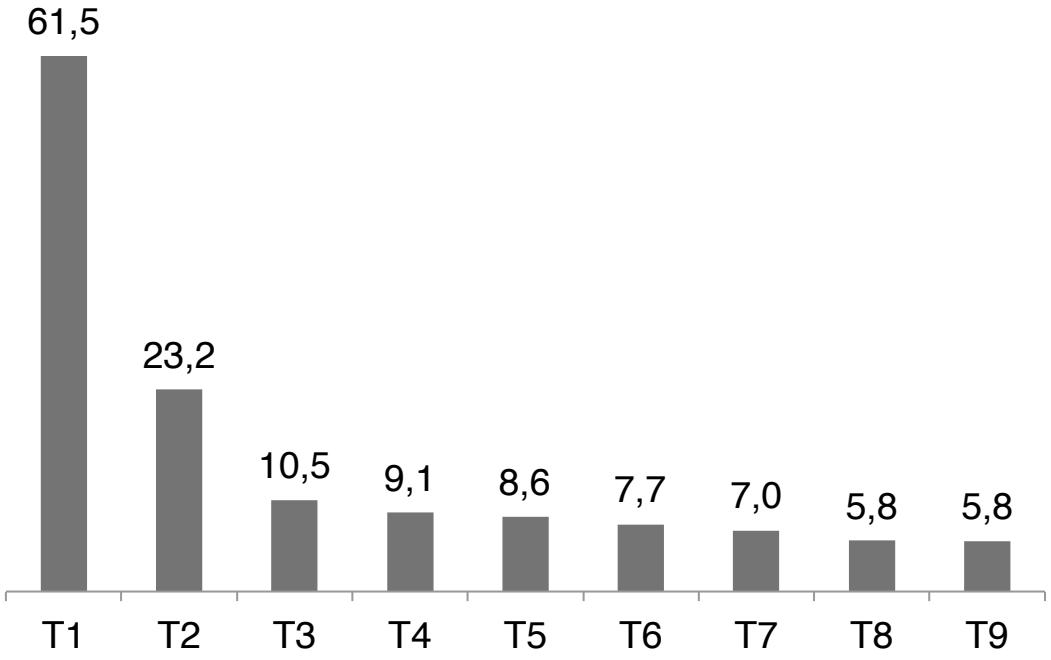
Unwanted side-effects:

- When several site-owners use the JavaScript code of a single tracking company, the tracker can learn a significant proportion of the user‘s browsing, potentially including very sensitive data*
- Web-sites have very limited or no control what is stored about their users or what trackers combine into what kind of profiles.
- Very often this is just an unwanted side-effect of a particular technical implementation choice, e.g., introducing a unique user ID („bklc“ in the example) across sites.

* Very often the information that the web-site itself receives is not critical on its own, however once combined across different web-sites (very often outside the scope of a single one published) it could potentially become very privacy sensitive profiling information (think e.g., health information on its own as not critical, but very sensitive if e.g., combined across web-sites with the personal profile page of a social network).

TRACKING IMPLEMENTATIONS CAN INTRODUCE SEVERE SIDE-EFFECTS WITH SOME ORGANIZATIONS MONITORING >60% OF THE PAGE LOADS.

ACROSS WEB-PAGE REACH FOR THE LARGEST TRACKING COMPANIES, DE IN %



Undesired effects

- User's are very often not aware about the tracking on web-pages (they expect to "just load the page").
- Being tracked across domains by a single organization can be a significant privacy invasion for the user, it potentially (willingly or accidently) leads to a comprehensive user-profile.
- Publishers (willingly or unwillingly) give away a huge amount of data about their users to 3rd parties.

Source and full results available at: Z. Yu, S. Macbeth, K. Modi and J.M. Pujol. "Tracking the Trackers". In Proceedings of the 25th ACM International Conference on World Wide Web. Montreal, Canada. 2016

USERS START USING BLOCKING TOOLS, BUT THIS INTRODUCES UNWANTED SIDE-EFFECTS FOR USERS AND PUBLISHERS

Selection of most-used anti-tracking services in Germany

Perceived threat drives anti-tracking services, however this can have unwanted side-effects for publishers, tracking companies and the users:

- A lot of services use curated blocklists of trackers, i.e. they do not send any (legitimate/ non-private) data and making e.g., measurements impossible.
- Many users turn to ad blocking to simply “stop the tracking” and thus reduce monetization of publishers.
- Preventing the JavaScript code from being executed often causes UX problems on the site for the users.

CLIQZ believes in anti-tracking that protects the users, but does not limit the comfort or reduce legitimate use cases

CLIQZ ANTI-TRACKING PROTECTS THE USERS' PRIVACY. BUT: IT ALLOWS SAFE, NON PRIVACY-INVADING TRACKING USE CASES

Users

- Are privacy protected.
- Are not profiled without giving their consent.
- Get the best web experience with as much functionality as possible, while preserving their right on their own data.

Site-owner/Publishers

- Experience less site malfunctions compared to other full-blocking tools.
- Get analytics data from all site visitors (incl. the CLIQZ users).
- Can control how CLIQZ anti-tracking should send data to them and thus have more control about the trackers on their pages.
- No/reduced accidental user data leakage to 3rd parties
- Technically proves "Datenschutz-Konformität"

Tracking companies

- Receive data as long as it is not privacy invading.
- Can fully deliver against site-owners' legitimate use cases.
- Do not run into trust-issues with the users.

CLIQZ anti-tracking sends data only if it is technically guaranteed to not contain private data about the user. Thus it reduces the risk for

- the user,
- site-owners to accidently give away private data or unknowingly leak valuable business information to 3rd parties,
- trackers to accidently build profiles that might (in the worst case) not even comply with German profiling laws.

CLIQZ ANTI-TRACKING IS INSPIRED BY CONCEPTS LEVERAGED FROM K-ANONYMITY, IN WHICH USERS COLLECTIVELY IDENTIFY UNSAFE DATA

Example for tracker information

```
bklc=55f6ad4d
l=https://www.DOMAIN.de/
ua=f82610bef1d54776cde605b90b0c7949
t=1444203542439
m=020810a3483fc8307caa483fd192bc02
lang=07ef608d8a7e9677f0b83775f0b83775
sr=1440x900x24
cpu=4b4e4ecaab1f1c93ab1f1c93ab1f1c93
platform=6d44fad93929d59b3929d59b3929d59b
plugins=d4de4a68c91685d0ff4838ce3714359a
cn=df62ddfcfa96f717f2ee5a7d912e7102
```

… and how CLIQZ anti-tracking is treating it

- The core of CLIQZ' anti-tracking system is to detect values which most likely uniquely identify the user
- Such "unsafe" values are considered elements that are only and always sent by a small number of users, irrespective of their function
- These unsafe values get scrambled, while the remaining "safe" information is passed on unchanged to deliver against safe use-cases

- In the given example, the string 55f6ad4d has been seen only once in a large population
- This value will be anonymised before passing it on
- The remaining lines are sent on as-is and thus allowing some user analytics

IN DETAIL – HOW IT WORKS: UNSAFE DATA ELEMENTS ARE THOSE THAT ARE ONLY AND ALWAYS SENT BY A SMALL NUMBER OF USERS

1	**Classify potential trackers**	**Potential trackers are 3rd party domains in browser requests,** • that are present in various different domains, visited by the community of our users during a longer time period, and • have been seen at least several thousand times during this period, and • where a proportion of requests have been observed attempting canvas fingerprint, sending private browser data or cookie values, or frequently sending QueryString data
2	**Determine safe data elements**	**Among potential trackers, data elements are defined "safe" if and only if they reach the "safeness quorum", i.e. when** • they are seen at least by large enough group of users to guarantee anonymity • during a period of several days.
3	**Whitelist safe data elements**	**Safe data elements are shared with all browsers,** ▪ so that the browser can locally decide whether to allow the data element to be sent to a potential tracker ▪ through a fast update mechanism to minimize the transient state from first-seen to whitelisted*

* Never-seen data elements are declared unsafe per default. The transient stage for to-be safe date elements is the period to get to the safeness quorum plus the time to propagate that information to the community. To mitigate risks from mis-classified data, a local set of whitelisted data elements is created on a user's browser: data elements are added to this list, if a user has seen several different values for a given key during several days.CLIQZ' statistical analysis based on 200.000 users shows that only 0.07% of data elements are declared unsafe due to the transient state only. Losing this amount of data is – although inconvenient – statistically not significant, unless it was used to identify a small sub-population, in which case it was privacy-violating anyway.

What this means for publishers and tracking companies, and CLIQZ recommendations

PUBLISHERS CAN CONTROL HOW CLIQZ ANTI-TRACKING SHOULD BEHAVE

We are aware of small edge-cases where CLIQZ anti-tracking blocks legitimate tracking and/or is putting additional efforts on publishers analytic teams:

- CLIQZ is aware that there are data elements that are not likely to reach the safety quorum, but still represent valid tracking use cases.
- An example could be the basket-value for an e-commerce site.
- CLIQZ anti-tracking is foremost about protecting users‘ privacy, if in doubt we have to be too conservative.
- We’re however not deliberately breaking these use cases. To mitigate these occurrences, we are happy to collaborate with publishers and site-owners.

Recommended action

First of all we recommend to implement a tracking that does not identify single users or their privacy. Generally most standard use-cases can be implemented this way, even using 3rd parties (e.g., counting visits, page impressions, and unique users within one domain). We’re happy to provide examples how to do this.

However, for other trackers, to put as much control as possible in the hand of the publishers we have introduced a tracking.txt file that publishers can use to tell CLIQZ how to behave towards trackers:

- You can create a “tracking.txt“ file for your site, providing specific rules about how to treat data elements in various trackers (see appendix for the protocol).
- This will allow you to define per tracker how CLIQZ should behave and thus allow you better analytics while we do not have to compromise our users privacy.

If preferred, though we do not recommend this:

You can even tell CLIQZ anti-tracking to behave similar to traditional anti-tracking services (thus blocking all requests*, with the effect of reduced measurability at your end and potential side breakage for the user).

* Please note: we explicitly decided to not make this CLIQZ anti-tracking default behaviour. It assumes that 3rd parties are malicious, which is obviously not true. CLIQZ strongly believes that 3rd parties have valid use-cases and we want them to continue their operations as normal as long as they only use data that is considered safe for the user.

FAQ

Does CLIQZ block adds?
No. We understand that adds can be an important part of the revenue-stream. Some of our users might decide to install an additional ad-blocker. CLIQZ however does currently neither come by default with an ad-blocker nor by default blocks ads.

Can my tracker be whitelisted manually?
No. Manual whitelisting is generally not possible. Our algorithms determine automatically whether a tracker sends private data. If your tracker is not sending private information it will automatically be whitelisted by the algorithm. Please note: We have no manual way to check or verify that your tracker has good “intentions“ only. Thus we have to rely on technical measurability.

I don‘t want CLIQZ entries in my tracking – what can I do?
Sometimes you might prefer not to see our users at all (instead of anonymised data). As a publisher you can always block all CLIQZ-Anti-Tracking via a tracking.txt (see appendix for examples). Note: This will not deactivate anti-tracking as the privacy of our users has the highest priority. It will however change the setting so that you will not get any data from our users (i.e., also not anonymised data). CLIQZ will then basically act like most other anti-tracking software and fully block requests. We do not recommend this option as it will disable even simple statistics like counting users at your end.

NEXT STEPS

After an intensive phase of testing in 2015,
CLIQZ anti-tracking will go live for our users March 8, 2016.

Questions or comments? Please reach out to us:

www:	https://cliqz.com/tracking
Email:	anti-tracking@cliqz.com
Twitter:	@cliqz
Facebook:	https://www.facebook.com/cliqzde
Phone:	**+49 89 9250 1055**

CLIQZ

Appendix: “tracking.txt” protocol

PROVIDING “TRACKING.TXT”

Tracking.txt gives publishers more control about the behavior of anti-tracking and trackers. Publishers can decide how the privacy protection should happen, and e.g., tell CLIQZ (and other trackers) to e.g, not send any data at all. The format of a tracking.txt file is simple:

- It must be a UTF-8 text file containing multiple lines,
- lines starting with # are treated as comments and ignored, lines starting with ! are treaded as directives and ignored, although they will place a role for versioning,
- lines starting with R are basic domain-level rules that apply to a domain that is considered a potential tracker. The format is:
    - R DOMAIN_PATTERN ACTION
    - for instance: “R adclick.com block” would block the connection to any 3rd party whose domain ends with adclick.com (for instance: cd1.adclick.com, adclick.com, cdnadclick.com). The matching is a case insensitive strong suffix match. Besides matching the domain patterns the action will only be applied to those domains that are considered potential trackers by Cliqz tracking protection. If the domain was not considered a potential tracker, Cliqz anti-tracking will not intervene regardless of the rules defined on tracking.txt.
- You can define multiple rules one per line, for instance:

```
R xyx.adclick.com replace
R .adclick.com block
```

- For any 3rd party domain considered a tracker, the Cliqz tracking protection will evaluate the 3rd party domain against the rules in the order they appear. If the domain ends with the DOMAIN_PATTERN the ACTION will be applied to that connection. If the domain does not end with the DOMAIN_PATTERN it will continue to the next rule. When no more rules are available Cliqz tracking protection will default to its default ACTION. (initially placeholder, later on replace).
- Note that rules have precedence, it is the responsibility of the tracking.txt creator to build a valid ruleset. For instance, the ruleset below might not behave as the user might thing,

```
R adclick.com replace
R xyx.adclick.com block
```

- The second rule will be never applied, since any domain that matches "xyx.adclick.com" will also match "adclick.com”.
- Possible actions are**:**
    - Block: will block the request, no data will be sent to the 3rd party
    - placeholder: unsafe data elements will be replaced by a fixed string (will be a shortened url pointing to , e.g. http://cliqz.com/tracking
    - replace: unsafe data elements will be replaced by a randomized string
    - empty: unsafe data elements will be replaced by an empty string
- The length of the tracking.txt will be limited to 4KB (including comments), any tracking.txt bigger than this will be ignored and the default behavior of Cliqz Tracking Protection will apply.
- The tracking.txt must be reachable as a GET request on endpoint /tracking.txt for any FULL domain that you want to have the custom rules that overrule the Cliqz Tracking protection default behavior. For instance, a tracking.txt present, on: http://mysite.com/tracking.txt will NOT be automatically applied to: http://sales.mysite.com/ or http://www.mysite.com/
- Typical aliases like "www" are NOT implicitly.

##
Example of tracking.txt
##
R xyx.adclick.com block
R adclick.com empty
R .googleanaylitcs.com block

This example will block anything send to 3rd party domains like "*.googleanalytics.com" and "*xyz.adclick.com". If the 3rd party domain is "*adclick.com" will the exception of "*xyz.adclick.com" the unsafe data elements will be replaced by empty strings. For the remainder of the domains the action will be the default action of Cliqz Tracking protection.