A First Look at Related Website Sets

Stephen McQuistin
University of St Andrews
St Andrews, UK
sm@smcquistin.uk

Peter Snyder
Brave Software
San Francisco, USA
pes@brave.com

Hamed Haddadi
Imperial College London & Brave Software
London, UK
h.haddadi@imperial.ac.uk

Gareth Tyson
Hong Kong University of Science & Technology (GZ)
Guangzhou, China
gtyson@ust.hk

Abstract

We present the first measurement of the user-effect and privacy impact of "Related Website Sets," a recent proposal to reduce browser privacy protections between two sites if those sites are related to each other. An assumption (both explicitly and implicitly) underpinning the Related Website Sets proposal is that users can accurately determine if two sites are related via the same entity. In this work, we probe this assumption via measurements and a user study of 30 participants, to assess the ability of Web users to determine if two sites are (according to the Related Website Sets feature) related to each other. We find that this is largely not the case. Our findings indicate that 42 (36.8%) of the user determinations in our study are incorrect in privacy-harming ways, where users think that sites are not related, but would be treated as related (and so due less privacy protections) by the Related Website Sets feature. Additionally, 22 (73.3%) of participants made at least one incorrect evaluation during the study. We also characterise the Related Website Sets list, its composition over time, and its governance.

CCS Concepts

• **Security and privacy → Privacy protections**; **Domain-specific security and privacy architectures**; • **Information systems →** *Web applications.*

Keywords

Web privacy; website relatedness

ACM Reference Format:
Stephen McQuistin, Peter Snyder, Hamed Haddadi, and Gareth Tyson. 2024. A First Look at Related Website Sets. In *Proceedings of the 2024 ACM Internet Measurement Conference (IMC '24), November 4–6, 2024, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3646547.3689026

1 Introduction

Browser vendors increasingly implement site partitioning (sometimes called third-party storage partitioning) in their products to protect user privacy on the Web. While browser vendors seem to agree that that partitioned third-party state should be the default in all browsers, they largely *disagree* on what steps, if any, should be taken in the interim, while websites adjust. Many websites today were designed to run without storage partitioning, and so break (in the subjective evaluation of the site operator, the site user, or both) when storage partitioning is applied. Some browsers have decided to prioritise user privacy, while others apply heuristic and list-based approaches to determine when reduced privacy protections are acceptable.

Google has proposed one such list-based approach for deciding when to (or not to) apply storage partitioning, called Related Website Sets. The Related Website Sets proposal consists of two parts. First, a list of sets of sites that are related to each other, and second, a browser policy for allowing unpartitioned storage access between sites that the list indicates are related to each other.

Underlying the Related Website Sets proposal is the intuition that if users understand that two sites are affiliated to a common organisation, then there is less need for the browser to enforce a privacy boundary between those two sites. Under this view, enforcing a privacy boundary between two sites that a user knows are related to each other is harmful to users (*e.g.,* risk of sites breaking, needing to redundantly log into multiple related sites) without providing any privacy improvement: the user expected that their information was going to be shared by both sites anyway. Crucially, websites related to each other under the Related Website Sets proposal do not need to have a common owner, enabling data sharing that would otherwise not occur.

In this work, we evaluate whether the assumptions underlying the Related Website Sets proposal are accurate. Understanding whether users perceive site relationships in the same way that Related Website Sets list maintainers do is important for Web privacy. If user perceptions *do not* match the list maintainers' expectations, Related Website Sets will result in privacy (and user) harming behavior, just as the Web is seemingly about to adopt a new privacy-improving baseline.

We make the following contributions:

(1) A **user study** where participants are asked to subjectively evaluate whether pairs of sites are related to each other (§3);
(2) An **evaluation of the composition and management of the current Related Website Sets list** (§4); and
(3) A discussion of **how Related Website Sets relates to other proposals**, from both other browser vendors, and other privacy tools (§5).

IMC '24, November 4–6, 2024, Madrid, Spain

ACM ISBN 979-8-4007-0592-2/24/11
https://doi.org/10.1145/3646547.3689026

Reproducibility and data access. We make available our code for gathering, processing, and analysing the data discussed in this paper. This, and the data used in our survey, along with its anonymised results, is available from https://doi.org/10.17630/450a41e5-1f12-43ef-b909-2640dcd0fe50.

2 Web privacy boundaries

Site-as-privacy-boundary on the Web. All current Web browsers either use, or plan to use, the *site* as the default privacy boundary on the Web. The site, in this context, refers to "effective top level domain, plus one subdomain" (eTLD+1). The "effective top level domain" refers to the suffixes defined in the public suffix list.[1]

Web browsers use eTLD+1 domains as site boundaries and aim to keep activity on one site unlinkable to activity on any other site. For example, the browser aims to keep activity on `facebook.com` unlinkable from activity on `mayoclinic.com` (two different sites), but does not aim to keep activity on `eff.org` unlinkable to activity on `act.eff.org` (two domains on the same site).

Enforcing site-as-privacy-boundary. Storage partitioning is a general strategy of giving sites access to different storage areas (e.g., cookies, `localStorage`) depending on the context the site is loaded in, and is the primary way that browsers enforce (or plan to enforce) the site-as-privacy-boundary.

As an example, imagine that `tracker.example` is a domain operated by a tracking service. This site can be loaded in different contexts: for example, users can visit it directly, as the first party, or they can access it indirectly, via assets (such as adverts) loaded in an `<iframe>`. Without storage partitioning, `tracker.example` is able to set and access the same set of cookies in both of these contexts, allowing it to track users across the Web.

Storage partitioning prevents this scenario, and enforces the site's privacy boundary by giving `tracker.example` [3] access to a different set of cookies depending on the context it is loaded in. When the user visits `tracker.example` directly, it can access one set of cookies, and when `tracker.example` is being loaded as a third-party on another site, such as `site.example`, it can access a separate set of cookies.

Related Website Sets and creating exceptions to the site-as-privacy-boundary. Google's proposed *Related Website Sets* feature is a browser capability that creates exceptions to storage partitioning, weakening the site-as-privacy-boundary policy. The general idea behind the Related Website Sets proposal is that there is little benefit to users (and possibly, some inconvenience created) by enforcing a privacy boundary between different sites that are clearly affiliated to the same organisation.

To understand how the Related Website Sets would work in practice, consider the following real-world example. "Times Internet" is a company that operates multiple popular websites in India, including `https://timesinternet.in` and `https://www.indiatimes.com/`. Without Related Website Sets, the browser would not allow either site to know that the same user was interacting with both sites until the user took some additional step, such as logging into both sites with common user credentials. With Related Website Sets, either site can embed an iframe from the other site. If code in that iframe then calls the `requestStorageAccess` method, then, since the two sites are related by the Related Website Sets list,[2] the browser would allow code running in that iframe access to that site's unpartitioned storage area (*e.g.,* cookies, user identifiers, etc), despite being embedded as a third-party. This would allow both sites to link page visits on each site to the same user.

Related Website Sets are comprised of multiple subsets. *Service sites* are sites that cannot be the top-level domain in a storage access grant. Users must first interact with another member of the set; otherwise, storage access is automatically granted. Service sites must be under common ownership with the set primary, and are designed to support the functionality or security of other set members. *Associated sites* are sites that must be clearly affiliated with the set primary (*e.g.,* using common branding, an about page, or similar). However, they are *not* required to have common ownership. *ccTLD sites* are ccTLD variations of other set members, and must have common ownership with the domain that they are a variant of.

Status of site-as-privacy-boundary in browsers. Most browsers currently use the site as the Web's privacy boundary, or have plans to. Safari, Brave, and Firefox all enforce the site-as-privacy-boundary by default, though some make some temporary exceptions to avoid breaking websites. Firefox and Safari both include the Storage Access API[3] that allows embedded third-party sites to request unpartitioned storage (and so, for an exception to the site-as-privacy-boundary policy), but requires user consent via prompt in some (Firefox) or all (Safari) cases.

Chrome and Edge *do not* currently implement a default site-as-privacy-boundary policy. While Chrome will continue to allow third-party cookies for the immediate future, it has deployed Related Website Sets, and intends it to be a permanent method that sites can use to gain exceptions from the site-as-privacy-boundary policy where users have opted to enable it.[4] To the best of our knowledge, Microsoft has not yet announced if it plans to adopt Related Website Sets.

Governance of Related Website Sets. Related Website Sets are proposed by site owners using pull requests on GitHub.[5] Pull requests are subject to a series of automated and manual checks. The automated checks ensure that Google's Contributor Licence Agreement[6] has been completed by the contributor, before a series of technical validation checks are run (*e.g.,* ensuring that no non-HTTPS sites are present, that all sites are eTLD+1s, among other factors).[7]

3 Can Users Determine Relatedness?

Underpinning the Related Website Sets proposal is the motivation that it creates exceptions to the site-as-privacy-boundary where doing so removes inconvenience to users, and follows from their expectations of privacy. As described in Section 2, the proposal allows sites to be related to each other by common ownership *or* common affiliation. The associated subset is reserved for "*domains*

[1] https://publicsuffix.org/

[2] https://github.com/GoogleChrome/related-website-sets/blob/main/related_website_sets.JSON

[3] https://privacycg.github.io/storage-access/

[4] https://privacysandbox.com/intl/en_us/news/privacy-sandbox-update/

[5] https://github.com/GoogleChrome/related-website-sets

[6] https://opensource.google/documentation/reference/cla/

[7] https://github.com/GoogleChrome/related-website-sets/blob/main/RWS-Submission_Guidelines.md#set-validation-requirements

whose affiliation with the set primary is clearly presented to users".[8] Therefore, the efficacy of the privacy boundaries that the Related Website Sets approach constructs relies upon users being able to determine that set members are related to each other, regardless of common ownership, and therefore, that they could reasonably expect their browser to share data between them.

Can users accurately determine relatedness? To assess whether users could determine that websites were related to each other, we conducted a user study in May 2024. Users were presented with links to 20 pairs of websites, asked to open those links and to view the websites, and then asked to determine if the two websites were related to each other by an affiliation to a common company or organisation. Users were not asked to identify what, if any, the common affiliation was. Each question was timed to determine how long participants spent assessing the relatedness of each pair of sites. Finally, after answering all 20 questions, participants were asked to indicate which factors they considered in determining when websites were and were not related to each other. Participants were able to exit the survey at any time, and to skip individual questions. The survey was entirely anonymous, with no personally identifiable information collected about participants, ethical approval was obtained,[9] and best practice related to informed consent was followed in carrying out the study.

The study was advertised via social media and within the institutions of the authors. This may skew participation towards individuals that have a computer science background, and that are familiar with the Web. While we do not investigate the impact that this has on our results, we hypothesise that they represent a baseline, and that participants with less familiarity would be less able to determine the relatedness of websites.

The pairs of websites were drawn from 4 groups, with each participant asked about 5 pairs, at random, from each group:

(1) *Sites that are members of the same Related Website Set.* All combinations of set primaries and associated sites within each set ("RWS (same set)"). The combinations in this group *are* related under the RWS proposal.
(2) *Sites that are members of other Related Website Sets.* All combinations of set primaries and associated sites; each site from a different set ("RWS (other set)"). The combinations in this group *are not* related under the RWS proposal.
(3) *Sites from Related Website Sets and another site within the same Forcepoint category.* Pairs were formed from all combinations of set primaries and associated sites, and a list of 200 sites, drawn randomly from the Tranco Top 10K list [5], filtered to sites within the same Forcepoint[10] category ("Top Site (same category)"). The combinations in this group are not related under the RWS proposal, but may be similar to each other given that they fall within the same Forcepoint category.
(4) *Sites from Related Website Sets and another site in a different Forcepoint category.* All combinations of set primaries and associated sites, and the above list of 200 sites, filtered to sites in a different Forcepoint category ("Top Site (other category)"). The combinations in this group are not related under the RWS proposal, but may be dissimilar to each other given that they are in different Forcepoint categories.

Category	Related	Unrelated
RWS (same set)	72 (28.1s)	42 (39.4s)
RWS (other set)	5 (25.5s)	100 (32.5s)
Top Site (same category)	8 (32.6s)	104 (33.2s)
Top Site (other category)	7 (31.5s)	92 (26.5s)

Table 1: Website relatedness survey results summary.

Factor used	Related	Unrelated
Domain name	12 (57.1%)	11 (52.4%)
Branding elements	14 (66.7%)	13 (61.9%)
Header text	9 (42.8%)	11 (52.4%)
Footer text	13 (61.9%)	11 (52.4%)
"About" pages or similar	10 (47.6%)	7 (33.3%)
Other	4 (19%)	5 (23.8%)

Table 2: Website relatedness survey: factors used to determine relatedness and unrelatedness.

Further manual filtering was performed to check that the websites on the Related Website Sets list were live, and that they were primarily English-language. As the survey was advertised in English-speaking regions, this manual filtering was performed to ensure that participants could reasonably assess relatedness. A large proportion of the sites in the Related Website Sets are not primarily English-language, and so this filtering reduced the number of sites on the Related Website Sets list from 146 sites to 31 sites. 822 pairs were generated, comprised of 39 *RWS (same set)* pairs; 426 *RWS (other set)* pairs; 141 *Top Site (same category)* pairs; and 216 *Top Site (other category)* pairs. The full set of generated pairs, alongside the anonymised data gathered by the study, is contained in the dataset released alongside this paper, and described in Section 1.

A total of 30 participants[11] provided 430 responses. Figure 1 and Table 1 summarize the results. Of the 114 responses to pairs within the same RWS set (*i.e.,* those pairs that are related), 36.8% incorrectly identified the websites as being unrelated. Across the 316 responses for pairs drawn from the other 3 categories, 93.7% indicate that the websites are unrelated.

Performing a two-sample Kolmogorov-Smirnov test pair-wise across the timing distributions for responses within each of the categories, we find no statistical significance between them ($p < 0.05$). However, looking only at the split of responses to pairs within the *RWS (same set)* category, as shown in Figure 2, we find a statistically significant difference in the time taken to determine relatedness vs. unrelatedness. This suggests that participants were more quickly

[8] https://github.com/GoogleChrome/related-website-sets/blob/main/RWS-Submission_Guidelines.md#set-formation-requirements
[9] Approved by the University Teaching and Research Ethics Committee (UTREC) at the University of St Andrews, with approval code CS17715.
[10] The Forcepoint ThreatSeeker (https://www.forcepoint.com/product/feature/threatseeker) database classifies URLs into broad categories (*e.g.,* news and media, business and economy).
[11] Due to the anonymous nature of the survey, "participants" here means individual sessions of the survey.

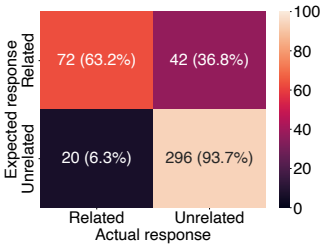

Figure 1: Website relatedness survey results matrix; percentages and heatmap color are within *Expected response.*

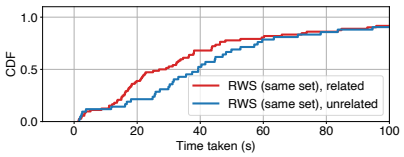

Figure 2: Website relatedness survey timing distributions; for pairs within the *RWS (same set) category*, split by response.

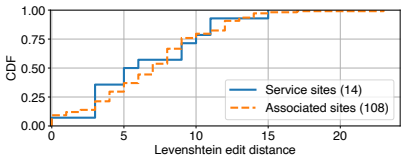

Figure 3: CDFs of the Levenshtein edit distance between service/associated sites and their primary domain, per the Related Website Set list at 26 March 2024.

able to determine relatedness, and that they spent longer evaluating the websites before concluding that they were unrelated.

Key takeaway — In 36.8% of pairs, participants incorrectly identified that websites drawn from the same Related Website Set were unrelated, and, when presented with such websites, spent longer determining their relatedness. Users would be unlikely to expect data to be shared in those instances where they were unable to determine relatedness.

How do users determine relatedness? Table 2 summarizes the factors that participants used to determine website relatedness. Of the 21 participants that responded to this question, "Branding elements" (*e.g.,* logos, colors, and similar) were most frequently used. The domain name itself was also frequently used, with 57.1% of respondents using this to determine that websites were related.

Key takeaway — Common branding elements, along with the domain names of the sites, are often used by participants to determine website relatedness. These factors should be used by the maintainers of the Related Website Set list when determining whether associated sites should be included in a set.

How similar are the second-level domains of set members? Given that participants (57.1% of respondents) indicate that they assessed relatedness using the domain names of sites, we evaluate whether automated tooling could be used to determine relatedness. To estimate how feasible this is, Figure 3 shows CDFs of the Levenshtein edit distance between each service or associated site's SLD, and its set primary's SLD.

A small proportion (9.3%) of associated site SLDs are identical to that of their set primary (*e.g.,* `poalim.xyz` is associated with `poalim.site`). This is likely to be impacted by the ability for site owners to declare an unlimited number of ccTLD variants of domains within the same set, limiting exact-match SLDs to those that have different gTLDs. Associated site SLDs have a median edit distance of 7 from that of their set primary's SLD. In such cases, it is unlikely a user could easily identify that the sites are related. Within this, there are sites that share common components (*e.g.,* `autobild.de` associated to `bild.de`) and others that are entirely distinct (*e.g.,* `nourishingpursuits.com` associated to `cafemedia.com`). In addition, though not present in the Related Website Sets list, domain squatting means that using SLD similarity as a measure of relatedness is risky; it does not confirm common ownership in itself.

Key takeaway — The similarity of SLDs is not a reliable way of determining relatedness between an associated site and a set primary, with half of associated site SLDs having an edit distance of 6 or more from that of their set primary.

How similar in structure and style are set members? Next, with 66.7% of respondents indicating that they used common branding elements to determine relatedness, we assess whether the *content* of the sites is similar, by computing the HTML similarity of each service and associated site when compared to its set primary. We use a well-known library[12] that, for a pair of websites, can compute the *style similarity* (based on CSS classes), *structural similarity* (based on HTML tags), and *joint similarity* (a weighted sum of both). From Figure 4 we observe that a significant proportion of service and associated sites are *dissimilar* to their set primaries, with a median joint HTML similarity score of 0.04.

Key takeaway — HTML similarity metrics show that service and associated sites are largely dissimilar to their set primaries. This means that manual validation of the common affiliation of associated sites is necessary, given that it is difficult to assess automatically.

4 The Related Website Set List

Having explored the question of the extent to which users can determine website relatedness, we next characterise the current Related Website Sets list, and how it is managed using GitHub. These are important research questions: if widely adopted, the composition and management of the Related Website Sets list may have significant implications on user privacy.

How are set members distributed across subsets? We first consider the composition of sets in terms of the subset types defined in Section 2. Associated sites are the most potentially privacy-impacting, given that common ownership is not required, and that users often fail to determine relatedness. Figure 7 shows the count of sites per subset category. As of the most recent RWS list in the dataset (26 March 2024), there were 41 sets; of these, 22% had one or more service sites; 14.6% had one or more ccTLD sites; and 92.7% had one or more associated sites. This shows that the overwhelming use case for the Related Website Sets mechanism is to incorporate associated sites, with a mean of 2.6 associated sites per set.

Key takeaway — 92.7% of Related Website Sets include one or more associated sites, where common ownership to the set primary is not required.

[12] https://github.com/matiskay/html-similarity

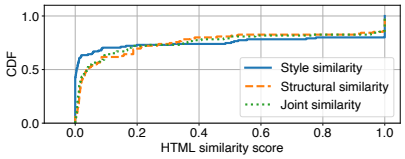

Figure 4: CDFs of HTML similarity scores of set primaries and their service/associated sites, per the RWS list at 26 March 2024.

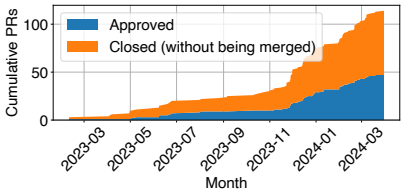

Figure 5: Cumulative count of PRs that propose a new set, split by final state.

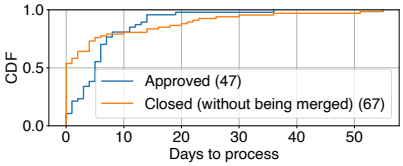

Figure 6: CDF of days taken to process PRs that propose a new set.

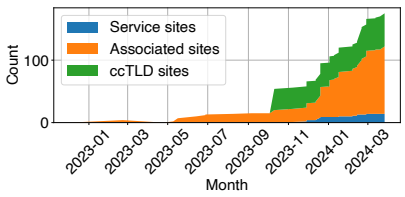

Figure 7: Set composition over time.

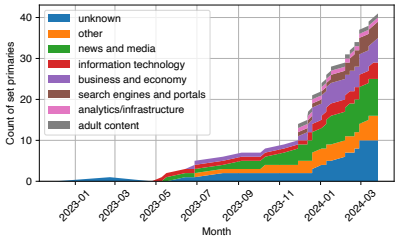

Figure 8: Forcepoint ThreatSeeker categories of set primaries

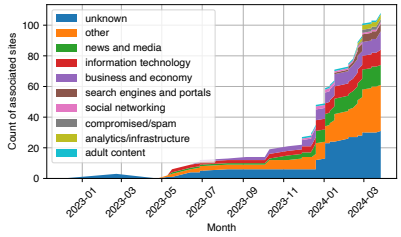

Figure 9: Forcepoint ThreatSeeker categories of associated sites.

What categories of sites are sets made up of? Next, we explore the composition of sets in terms of the Forcepoint category that they fall into, to characterise those sites that are seeking an alternative to third-party cookie functionality. Figures 8 and 9 show the categories of set primaries and associated sites, drawn from Forcepoint ThreatSeeker classifications. Note, similar categories are merged together, while smaller categories are grouped into "Other". As shown, the largest individual category for set primaries is *News and media*; these contain associated sites in other categories (*e.g.,* the set for `bild.de`, a German news site, includes `computerbild.de`, its related IT news website). The RWS mechanism allows data to be shared across these sites, enabling, for example, common ad tracking and profiling. Additionally, some sets contain analytics or tracking infrastructure explicitly: `ya.ru` (Yandex, a Russian Internet company) includes `webvisor.com`, a Web analytics service.

Key takeaway — A significant portion of sites fall into categories, such as "News and media", that are likely to benefit from existing third-party cookie functionality, making their early adoption of alternatives intuitive.

How often are set proposals rejected? Next, we turn to the governance of the Related Website Sets list. Given the potential privacy implications of the list, it is crucial that it is well maintained, and that the rules governing set submissions are clear. Figure 5 shows the cumulative count of pull requests on the Related Website Sets list over time, through to March 30th 2024, comprising 114 requests. As shown, the rate at which pull requests are submitted has grown over time, as the Related Website Sets proposal has developed. Additionally, the split between approved and closed (without merging) requests has shifted, with 58.8% of all pull requests closed without being merged, suggesting a significant volume of invalid submissions.

GitHub bot comment	Count
Unable to fetch .well-known JSON file	202
Associated site isn't an eTLD+1	65
Service site without X-Robots-Tag header	19
PR set does not match .well-known JSON file	12
Alias site isn't an eTLD+1	10
Primary site isn't an eTLD+1	9
Other	8
No rationale for one or more set members	5

Table 3: RWS GitHub bot validation messages.

A set may be proposed across multiple pull requests. Site owners will often propose a new set, receive the results of the automated checks, close the pull request, and then open a new pull request. Across the 114 pull requests in the dataset, only 60 set primaries are represented, giving a mean of 1.9 pull requests per set primary.

Key takeaway — 58.8% of pull requests on the Related Website Sets list are rejected, suggesting that the automated checks are successful in enforcing the technical set-level requirements of the Related Website Sets proposal.

What are the common validation errors? The set-level technical validation checks are carried out automatically, and the results are reported back via a GitHub bot, which adds a comment to the pull request; understanding the reasons why set proposals are rejected may help to streamline the process, and identify common misunderstandings. Table 3 shows the number of occurrences of each of the validation errors observed in the pull request dataset. While validation is performed at the set-level, some error messages are produced per site; additionally, the validation is performed

again if the pull request is updated. This results in a one-to-many mapping between pull requests and the validation errors that are observed.

The most frequent error is that the `.well-known` file is not able to be fetched. This is a JSON file (to be publicly accessible via each set member) that contains the set (*i.e.,* the same data that is available in the Related Website Set list). This ensures that proposers have administrative access to the domains that they are submitting. This is likely to be an oversight on the part of the submitter.

The next most frequent error is that the set contains an associated site that is not an eTLD+1. For example, a set proposer might have `example.com` as the set primary, and `a.example.com` as an associated site. Assuming that `example.com` is not an eTLD (*i.e.,* present in the Public Suffix List), then `a.example.com` is not a third-party site, with respect to `example.com`. In these cases, this represents a fundamental misunderstanding of the privacy boundaries that already exist, and that Related Website Sets are reshaping.

*Key takeaway — The most frequent validation errors suggest that the Related Website Sets proposal is complex, both in terms of the technical requirements (*e.g., *for the* `.well-known` *file), and the privacy boundaries that are constructed. This suggests that documentation and tooling (for validating a proposed set before submission) could be improved.*

How long does it take for a pull request to be processed? Having looked at the rate that PRs are submitted, and the common automated validation errors, Figure 6 shows the time taken for set proposals to be processed, either successfully (*i.e.,* approved and merged in) or unsuccessfully (*i.e.,* closed, without being merged in).

We see that 54.3% of unsuccessful pull requests are closed within the day that they are opened. This is likely to result from the automated validation process described above: submitters frequently close their pull requests after receiving output from the GitHub bot. However, we observe a long-tail in the time taken to close unsuccessful requests.

The median time to process a successful request is 5 days. Only 1 of the 47 merged pull requests fail any of the automated checks, suggesting that the time to process successful requests is driven primarily by their manual validation by the maintainers of the Related Website Sets list.

Key takeaway — The automated validation checks provide quick feedback to submitters, while the manual validation checks contribute to a median time to process successful requests of 5 days. Given the low rate of requests, it is unclear how the manual component of the process will scale should this mechanism become more widely adopted.

5 Related Work & Discussion

List-based Web privacy. Related Website Sets is most similar to the Disconnect list,[13] an expert-curated commercial product. The Disconnect list consists of two sub-lists: the *services* list, a list of domains determined to be related to privacy-harming (or otherwise undesirable), and the *entities* list, a list of domains that are run by the same organisations. This second list is similar to Google's Related Website Sets list in several ways. First, both list sets of domains that are controlled by the same organisation. Second, both are used by popular Web browsers (*e.g.,* Firefox and Edge) to decide whether privacy protections should be relaxed. Third, both lists are curated by a small group of experts (*e.g.,* Disconnect employees, Google employees). A crucial difference is that the Related Website Sets list, through associated sites, relaxes the requirement that sites be operated by the same company, and only requires that they have a common affiliation that is clearly presented to users. Our work has shown that this relaxation is often at odds with users' ability to determine website relatedness.

There are other popular list-based approaches to Web privacy that differ from the Related Website Sets list significantly. Filter lists, such as EasyList,[14] and supplementary lists like uBlock Origin[15] and AdGuard,[16] are primarily crowd-sourced. While the Related Website Sets list defines rules over domains, filter lists define rules over URLs. Filter lists are much larger than expert curated lists; Related Website Sets and Disconnect describe hundreds or thousands of domains (respectively), while EasyList alone includes tens-of-thousands of rules [9].

Stateful third-party Web tracking. The Related Website Sets proposal, and this work examining it, relates to the well-studied area of stateful third-party tracking on the Web. Browser state could be abused to violate user privacy on the Web, both through browser capabilities intended to store application level information (*i.e.,* cookies [7, 8], and user IDs), but also other indirect capabilities (*e.g.,* browser caches [2], Flash [11], cached `ETag` headers [1], the DNS cache [4], the "favicon" cache [10], "Alternative-Service" (`Alt-SVC`) headers [13], and "HTTP Strict Transport Security" (`HSTS`) headers [12]). Other work attempted to distinguish between cases where third-party state is used to track users, and cases where third-party state is used for more benign purposes [6]. Our work has shown that the RWS may open up users to greater tracking.

6 Conclusions

This paper has evaluated whether the assumptions underlying the RWS proposal are accurate. Browser vendors have enabled, or plan to enable, third-party storage partitioning, an effective protection against the most pervasive kinds of privacy harm on the Web. However, there are significant compatibility risks. Many sites on the Web were designed for how Web browsers worked when they were less privacy-protected, and efforts to find ways to improve privacy *without* breaking "legacy" sites are essential.

While Related Website Sets proposes an appealing solution to this problem, we find that its underlying assumptions about *relatedness* do not hold, and we find that Web users cannot accurately evaluate whether two sites are in fact affiliated to the same organisation. This suggests that exceptions made to the site-as-privacy-boundary, on the basis of relatedness, need to be explicitly indicated to the user (*e.g.,* via the browser UI itself); we leave study of the efficacy of such an approach to future work.

We hope these findings are useful to those developing techniques to improve Web privacy without breaking compatibility, and demonstrate the importance of thoroughly testing any assumptions about user behaviour or knowledge that such techniques rest on.

[13] https://github.com/disconnectme/disconnect-tracking-protection

[14] https://easylist.to/

[15] https://github.com/gorhill/uBlock

[16] https://adguard.com/en/welcome.html

References

[1] Mika D Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. 2011. Flash cookies and privacy II: Now with HTML5 and ETag respawning. *Available at SSRN 1898390* (2011). https://doi.org/10.2139/ssrn.1898390

[2] Collin Jackson, Andrew Bortz, Dan Boneh, and John C. Mitchell. 2006. Protecting browser state from web privacy attacks. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland) *(WWW '06)*. Association for Computing Machinery, New York, NY, USA, 737–744. https://doi.org/10.1145/1135777.1135884

[3] Jordan Jueckstock, Peter Snyder, Shaown Sarker, Alexandros Kapravelos, and Benjamin Livshits. 2022. Measuring the Privacy vs. Compatibility Trade-off in Preventing Third-Party Stateful Tracking. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 710–720. https://doi.org/10.1145/3485447.3512231

[4] Amit Klein and Benny Pinkas. 2019. DNS Cache-Based User Tracking. In *Proceedings of the Network and Distributed System Security Symposium (NDSS) 2019*. Internet Society, USA. https://doi.org/10.14722/ndss.2019.23186

[5] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the Network and Distributed System Security Symposium (NDSS) 2019*. Internet Society, USA. https://doi.org/10.14722/ndss.2019.23386

[6] Tai-Ching Li, Huy Hang, Michalis Faloutsos, and Petros Efstathopoulos. 2015. TrackAdvisor: Taking Back Browsing Privacy from Third-Party Trackers. In *Passive and Active Measurement*. Springer International Publishing, Cham, 277–289. https://doi.org/10.1007/978-3-319-15509-8_21

[7] Jonathan R. Mayer and John C. Mitchell. 2012. Third-Party Web Tracking: Policy and Technology. In *2012 IEEE Symposium on Security and Privacy*. IEEE, USA, 413–427. https://doi.org/10.1109/SP.2012.47

[8] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. USENIX Association, USA, 155–168.

[9] Peter Snyder, Antoine Vastel, and Ben Livshits. 2020. Who filters the filters: Understanding the growth, usefulness and efficiency of crowdsourced ad blocking. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4, 2 (2020), 1–24. https://doi.org/10.1145/3392144

[10] Konstantinos Solomos, John Kristoff, Chris Kanich, and Jason Polakis. 2021. Tales of favicons and caches: Persistent tracking in modern browsers. In *Proceedings of the Network and Distributed System Security Symposium (NDSS) 2021*. Internet Society, USA. https://doi.org/10.14722/ndss.2021.24202

[11] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. 2010. Flash cookies and privacy. In *Proceedings of the AAAI Spring Symposium Series 2010*. Association for the Advancement of Artificial Intelligence, USA.

[12] Paul Syverson and Matthew Traudt. 2018. HSTS Supports Targeted Surveillance. In *Proceedings of the 8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*. USENIX Association, USA. https://www.usenix.org/conference/foci18/presentation/syverson

[13] Trishita Tiwari and Ari Trachtenberg. 2019. Alternative (ab)uses for HTTP Alternative Services. In *Proceedings of the 13th USENIX Workshop on Offensive Technologies (WOOT 19)*. USENIX Association, USA. https://www.usenix.org/conference/woot19/presentation/tiwari