A Systematic Study of the Consistency of Two-Factor Authentication User Journeys on Top-Ranked Websites

Sanam Ghorbani Lyastani[†,⋆], Michael Backes[†], Sven Bugiel[†]

[†]CISPA Helmholtz Center for Information Security, [⋆]Saarland University

Abstract—Heuristics for user experience state that users will transfer their expectations from one product to another. A lack of consistency between products can increase users' cognitive friction, leading to frustration and rejection. This paper presents the first systematic study of the external, functional consistency of two-factor authentication user journeys on top-ranked websites. We find that these websites implement only a minimal number of design aspects consistently (e.g., naming and location of settings) but exhibit mixed design patterns for setup and usage of a second factor. Moreover, we find that some of the more consistently realized aspects, such as descriptions of two-factor authentication, have been described in the literature as problematic and adverse to user experience. Our results advocate for more general UX guidelines for 2FA implementers and raise new research questions about the 2FA user journeys.

I. INTRODUCTION

Would you buy a car where the gas and brake pedals are interchanged? You would probably be able to learn to drive this car safely after some acclimatization period. Still, it would be an experience that is very inconsistent with what you are used to, and you would most likely not continue using such an unpleasant car. Like this everyday example, a consistent user experience is crucial for websites to fit the mental models that users built and avoid unnecessarily increasing the users' cognitive load and friction by forcing them to learn something new. This important best practice has been captured in *Jakob's Law of Internet User Experience* [37], [50], [51] as one of several heuristics for user experience [74], [75] and usability [49] that guide website design. Striving for consistent user experience has ruled website design for years, evident in the design of, e.g., online shopping, banking, forums, blogs, or streaming services. The same best practices also apply to user authentication as part of the user experience.

When it comes to the incumbent authentication scheme on the web today, text-based passwords, the user experience of passwords is highly consistent across different websites, although recent work [45] discovered inconsistent password policies for blocklists, strength meters, and composition when setting passwords on the top websites. Regardless of this inconsistency, text-based passwords are notorious for their security issues. Among the different solutions proposed to strengthen user authentication on the web, two-factor authentication (2FA) has been shown to have a very tangible positive effect on account security [40], [46], [68]. Nowadays, 2FA is frequently recommended to end users to improve their security hygiene [55]. Fortunately, many websites are starting to offer 2FA options to their users [4], [29]. However, previous work [13], [58] demonstrated that users struggled with 2FA when their 2FA journey did not match their expectations or previous experiences and advocated for more standardized procedures. In a survey with 2FA adopters (see [30] and the summary in Appendix A), we found corroborating evidence that inconsistent implementations of the 2FA user journey caused friction for users that lowered the usability of 2FA and led users to refuse 2FA or abandon websites. Unfortunately, up to today, *we have only very few insights about how consistent the user experience of 2FA is across different websites.*

To provide new insights about how websites offer 2FA to their users and how consistent this user experience is across websites, we systematically study the 2FA user journeys on 85 popular websites in this paper. More specifically, we want to determine whether these websites consistently follow the same design patterns and strategies to offer 2FA to their users. Or, in other words, we are interested in the external functional consistency of the 2FA user journeys across popular websites.

To approach our research question systematically, we need concrete factors based on which we can compare the different user journeys. Unfortunately, such a list of factors does not exist for two-factor authentication, and there is no common guideline or best practice on how to implement the 2FA user flow on websites. Furthermore, 2FA is a technology that has only started gaining wider adoption among websites in the last couple of years and was hence in many cases not part of the initial website design. Additionally, the 2FA ecosystem is fragmented into various options for 2FA, such as TOTP, WebAuthn, push notifications, SMS, or custom solutions, each with its own setup process, dependencies (e.g., hardware token or app), and benefits/drawbacks in terms of usability and security [9], [56]. For these reasons, it was not a priori obvious which exact comparison factors could describe potentially diverse user journeys on different websites.

To solve this challenge, we devised a methodology to derive a list of comparison factors from open and axial coding of existing user journeys on the 85 websites in our data set. As a result, we created a list of 22 comparison factors that describe the user journey from *discovery* of an offered (promoted) 2FA support during sign-in/registration, to the *education* of the user about the available second-factor options and their

setup processes, to *usage* and *deactivation* of the chosen 2FA option(s). Based on these factors, we then compared the 85 websites in our data set to identify common design patterns and differences and to highlight beneficial or detrimental patterns for user experience.

Our results show that there is no overarching design pattern for the user journey that most websites follow. Instead, we found the design space to be clustered into groups of websites with very similar patterns, some of those favored by the top websites and others by less popular sites. The only design aspects that almost all websites agree on about 2FA are that it is an optional feature, how it should be called and described, and where it should be found in the account settings. In contrast, for the crucial steps of setting up and using 2FA, we found that websites implement mixed strategies, such as varying numbers of simultaneously supported 2FA technologies, inconsistent presentation of device remembrance options, or varying degrees of feedback to users.

According to UX guidelines, this lack of consistency increases users' cognitive load and should be avoided. However, consistency alone does not guarantee a good user experience. We found that several of the more consistently used design patterns have been described in prior work as problematic for user experience, including non-encouraging descriptions or missing possibilities to personalize the 2FA. We also discovered that the journeys of top websites, like `icloud.com`, are outliers from the best practices in the academic literature. Therefore, our results create a call for action to reinvestigate what constitutes a good overall 2FA user experience, to study whether there is a "gold standard" for implementing 2FA user journeys, or to explore the motivations of website developers to implement specific design patterns.

II. BACKGROUND

A. Two-Factor Authentication

With two-factor authentication enabled on a website, a user must successfully provide two authentication factors to verify their identity. Almost always, the first factor is a traditional text-based password. For the second factor, there are different technical realizations of knowledge, possession, and inherence factors. Most common [4], [11] are *one-time codes* delivered via SMS text-message, phone call, or TOTP [47] apps, like Google Authenticator, Duo, or custom apps that the user registered with the website; *push notifications* by sending an alert message to a dedicated app on the user's phone that asks the user to confirm a login attempt; and hardware tokens via the *U2F* or *FIDO2/WebAuthn* [73] standards that rely on public key cryptography and challenge-response protocols.

Each of these comes with its own set of usability and security benefits and drawbacks [56]. Important for our work is that a website with 2FA support can offer one or multiple of those 2FA options, may even allow users to set one of those solutions up multiple times, or may enforce a particular order in which they can be set up or used.

A commonly acknowledged problem with two-factor authentication is account recovery when a user loses access to a factor (e.g., a mobile device with the TOTP app is unavailable). Often the strategy to avoid lockout from a 2FA-protected account is to set up a dedicated recovery option, such as printed-out one-time passwords that can replace another 2FA option, or to configure multiple 2FA options, when supported by the website, e.g., multiple hardware security keys.

B. User Experience

Unfortunately, providing an exact definition of "user experience" is very difficult, as there is no consensus on the exact definition [7], [36], [42], [53]. However, a common topic among the definitions is that UX encompasses the various aspects of user interaction with a product, such as a website. Cooper et al. [15] note that there exist three overlapping concerns for UX: form, content, and behavior. While form and content (e.g., UI design or phrasing) have an impact on usability, this work focuses on behavior (i.e., functionality) and only touches on some aspects of form and content.

To help designers provide the best possible user experience, various best practices and general guidelines have been developed (e.g., books [15], [39], [62], [69], [75] or online resources, such as *Laws of UX* [74], *Nielsen Norman Group* [2], or *Interaction Design Foundation* [1]). Among the earliest are Shneiderman's eight "Golden Rules" for interface design [61], [62] and Nielsen's "10 Usability Heuristics for User Interface Design" [49], [52]. Shneiderman's rules state, for instance, that one should strive for consistency and provide informative feedback to users. Of Nielsen's heuristics, heuristic nr. 4, also known as *Jakob's law of Internet user experience* [51], is the most important for this work and provides the motivation to study the consistency of 2FA user journeys across websites. This heuristic states that *"users spend most of their time on other sites"* and that *"users prefer a site to work the same way as all the other sites they already know."* As a consequence, one should *"design for patterns for which users are accustomed."* Having such conventions and consistency helps users build upon existing mental models and avoid cognitive friction by forcing them to learn something new [75]. If an unconventional website mismatches the user's mental model, the website will be difficult to learn, difficult to use, or even rejected [69]. One way to drive *external* consistency is to make ample use of guidelines. For instance, for apps there are Google's Material Design Guidelines [34] and Apple's Human Interaction Guidelines [8]. We are not aware of any general guidelines for implementers and designers of two-factor authentication on websites, although there exist case-specific guidelines (for example, FIDO2 [27]) or small collections of best practices (e.g., [22], [67]).

Although in this work we focus on external, *functional consistency*, some of the comparison factors for 2FA user journeys that we identified (see Section VI) also touch on other UX guidelines and best practices. Tesler's law [75] states that for any system there is a certain amount of complexity that cannot be reduced, and it is recommended that the product design ensures that as much as possible of the burden on the user is lifted. Krug [39] recommends that if a difficulty for the user cannot be avoided, the design should provide brief and timely guidance, and Cooper et al. [15] recommend contextual help and assistive interfaces without the need to break the user's flow. If it cannot be avoided that the user has to learn something new, users learn best from examples (e.g., pictures, screenshots, or short tutorial videos) [69]. In addition,

Hick's law [75] recommends breaking down complex tasks into smaller steps to decrease the cognitive load. Moreover, excise tasks, such as navigational excise, should be reduced, e.g., by reducing the number of places that a user must go and providing clear overviews [15]. Hereby, it is important to consider that users do not read but scan webpages [39] and that this scanning is based on the mental model they built from past experiences, which creates expectations of what they want to see and where [69]. Furthermore, part of Postel's law [75], similar to Shneiderman's third golden rule [61], [62], recommends providing clear feedback to users, and the Peak-End Rule [75] recommends paying attention to the final moments of the user journey because people judge an experience largely based on how they felt at its peak and recall negative experiences more vividly than positive ones. Lastly, personalization can enhance the user experience. Although we did not explicitly investigate websites for their quality of those additional guidelines, some of our comparison factors indicate if 2FA settings are found in common places, if additional information and instructions are provided, if user notifications are present, or if users can set preferences.

III. Related Work

Several works have studied two-factor authentication problems and focused on the usability component and user attitudes. Bonneau et al. [9] conducted a systematic expert assessment of various authentication solutions, including many of the solutions used for 2FA. They concluded that the usability of these solutions falls very often short compared to text-based passwords. In contrast to Bonneau et al., most other works relied on user studies to investigate 2FA problems.

A focal point of these user studies was the setup and usage of different two-factor authentication solutions to understand users' attitudes toward 2FA, obstacles for its adoption, and how to improve the usability and user experience. Early works studied two-factor authentication in settings such as online banking [35], [38], [70], [71] or military [63] services. Like other studies of 2FA [10], [20], [21], [25] they found that users consider 2FA to be often burdensome and slow, that convenience trumps perceived security, and that users do not always understand the risks that 2FA tries to remedy. Several works have studied 2FA problems in organizational contexts [6], [14], [23], [57], [64], [66] where the use of MFA can be mandated. While these studies show that many of the problems overlap with non-organizational settings, they could also shed new light on the positive influence of features such as device remembrance [23], [57] or better help and instructions.

Several studies [38], [56], [70], [71] compared different options for the second factor to identify option-specific differences in user attitudes and usability, while other works specifically studied security keys [13], [16], [58] or authenticator apps [19]. An interesting aspect of these works [13], [56], [58] for our study is that they differentiated between 2FA setup and login, where users often struggled in the setup due to unclear instructions/workflows. Strong recommendations from those works were clearer instructions and guidance for the setup to avoid user frustration that often leads to non-adoption. Additionally, improved notification design patterns [32] have been shown to encourage users to adopt 2FA.

Lastly, recent works [26], [31], [41], [54] specifically studied FIDO2 *single*-factor authentication. They found similar user concerns as for 2FA. However, the weighting of the concerns shifted (e.g., loss of the authenticator device is ranked very high) or new concerns were added (e.g., misunderstanding biometric WebAuthn). Relevant to our work, the FIDO Alliance has recently published UX guidelines for security keys [28] and implementers of desktop authenticators [27] that, similar to our methodology, divide the user journey into different steps and provide recommendations for the design of each step; however, explicitly tailored to the technical details of FIDO2/WebAuthn with biometric authenticator devices or security keys. Nevertheless, those guidelines incorporate many of the UX guidelines explained in Section II-B.

The key difference of our work is that we do not study how concrete changes in form, content, or functionality affect the usability and concrete experience of 2FA, but that we are the first to systematically study how *consistent* the user experience is across existing popular websites. Our work, in contrast to previous works, strongly focuses on Jakob's law of Internet user experience which states that an inconsistent user experience across websites increases cognitive friction and can be detrimental to users' adoption. Providing first insights into how well 2FA user journeys adhere to this law is the core contribution of this work. Further, we are not aware of prior studies that measured Jakob's law across a larger number of websites, but instead, to the best of our knowledge, qualitative and quantitative testing of websites focuses on single websites or comparative user studies between a small set of websites based on general UX best-practices and guidelines. Therefore, we had to devise a methodology to measure the consistency of the 2FA user journeys on different websites.

IV. Methodology

To compare the 2FA user journeys of different websites and measure their consistency, we require concrete comparison factors that describe these journeys. Unfortunately, there is no existing list of such comparison factors, of which we are aware, or general guidelines for implementing 2FA on websites from which we could extract such factors. Therefore, a crucial challenge for our study is to create a list of relevant and representative factors. We used inductive research methods (e.g., [43, Chapter 11.4]) to solve this challenge. Figure 1 gives an overview of our methodology, whose data collection (Section IV-A) and identification of comparison factors (Section IV-B) we explain in the following. In a nutshell, we use open and axial coding from grounded theory on the screen-recorded 2FA user journeys of different websites to identify the list of comparison factors and to form an agreement about how each website matches each factor. Using the coding results, we then compare the different websites and study how consistently they implement the 2FA user journey and where they differ (Sections VI and VII).

A. Data Collection

The first part of our methodology is to collect a representative data set of user journeys recorded on different websites that we can analyze. Since we are building our knowledge about user journeys inductively, the screen recordings must have as high as possible coverage of all steps and choices
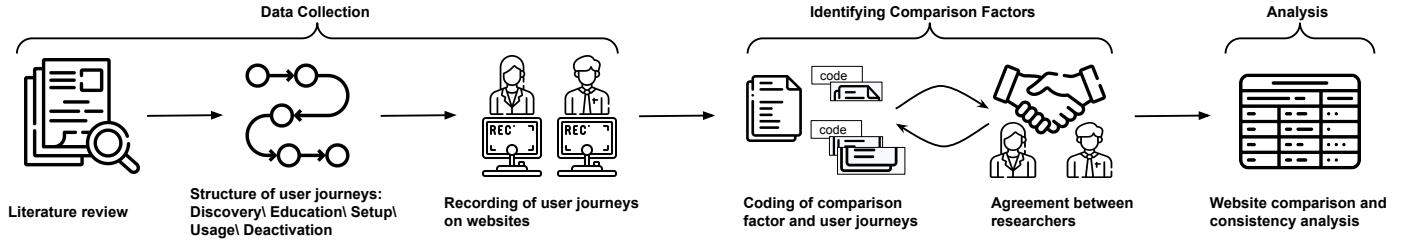
Fig. 1: Overview of our methodology

along each journey. To this end, an automated tool, such as a web crawler, could be used to explore various websites. Unfortunately, the need for a priori knowledge about how websites might implement their user journeys to guide the crawler and the need to use additional authentication devices (e.g., phone or security key) hamper an automated collection. Alternatively, we could use a crowd-sourced data collection, e.g., Amazon Mechanical Turk. Unfortunately, this was not possible in our setting for ethical reasons. We would need to ask our participants to use private accounts (or create fake accounts) on different websites and explore security settings for which they might need to provide a (personal) email address, phone number, or security key, and risk accidentally locking themselves out of an account as a result of a misconfiguration.

Instead, two researchers independently explored and screen-recorded the 2FA user journeys for our study. Their general instruction was to "thoroughly explore all aspects" of these journeys. However, this exploration could be informed beforehand from the literature, which discusses different aspects of 2FA user journeys (see also Section III). For example, recent works (e.g., [18], [32], [41]) and guidelines [27], [28] identify discovery of 2FA options and user education, different works studied 2FA setup and login (e.g., [56], [58]) or mandating 2FA (e.g., [6]), and account recovery is a commonly identified problem. Based on those insights, we structure the exploration of user journeys into five steps: The first step is *Discovery* of 2FA support on the website. We explore the landing pages, FAQ, and account registration for information on 2FA and follow all linked information. To further encourage users, there might also be nudges and messages about securing the account with 2FA, for which we scan the websites' interfaces. To use 2FA, the user must find the corresponding settings in their account settings, which we explore for the locations and options for authentication. In the next step, *Education*, we examine how a website introduces 2FA and if it gives further explanations, such as descriptions of how 2FA works and what it offers. Once the user has decided to use 2FA, they need to *Setup* their second factor(s). We explore the workflow of setting up all supported 2FA options (e.g., TOTP or Security Key). This exploration includes examining the websites' instructions, exploring the different settings choices (e.g., personalization choices), and feedback from the website on successful setup. After setting up two-factor authentication, we examine the *Usage* of 2FA on the website. We re-login and observe how the website prompts us to authenticate and whether it provides any options (e.g., device remembrance), which we explore. Finally, we explore the 2FA *Deactivation* procedure in the website settings and how the website communicates those changes.

For data collection, we maintained identical study conditions. All recordings were made on MacBooks running macOS 11 in the same network with the latest version of the Chrome browser when we started our data collection. Data collection was carried out between 06/2021 and 08/2021. This fixed setup should minimize the risk [72] of external factors (e.g., varying geolocation) and possible risk-based authentication to distort the data.

It is important to note that we focus *only* on the workflow for account creation, initial 2FA setup, and 2FA usage. We do not explore the workflows for account recovery or to change personal information relevant to 2FA after 2FA setup, such as a phone number or email address. We consider those follow-up problems to be studied after we have insights into the consistency of the fundamental steps that mint the users' first impressions about 2FA on a particular website.

B. Identifying Comparison Factors

Since there is no predefined set of factors to compare 2FA user journeys, we applied emergent coding [43, Chapter 11.4], in particular open and axial coding from grounded theory, to identify comparison factors from our recorded user journeys. These coding techniques are commonly applied in qualitative data analysis for text content. To still use those established methods, we treated the screen-recorded journeys like semi-structured interviews. Semi-structured interviews follow a set of predetermined questions, but the remaining questions are made up during the interview based on the interviewee's answers. We transferred this idea to our data collection (see Section IV-A): The exploration of user journeys follows a set of predetermined questions for discovery over usage to deactivation but allows the researcher to divert to individually explore a website in more detail and discover new or unique aspects of 2FA user journeys. Two researchers separately iterated through the set of recorded user journeys and segmented the observed journeys into meaningful parts to which they assigned concepts (i.e., codes). This is followed by axial coding, where the two researchers combined those concepts via induction and deduction into categories. For example, the codes "2FA advertised on the landing page" and "2FA recommended during account creation" can be combined into "Promotion of 2FA." These combined concepts can be used as comparison factors on all websites. The researchers also noted whether there exists a functional dependency between factors. After agreeing on the list of comparison factors, the researchers discussed how each website matches each comparison factor (e.g., fully, partially, or not at all). Since the matching of comparison factors might reveal that the list of factors is

too fine-grained, potentially weighting small differences too heavily, or too coarse-grained, potentially hiding important differences, the researchers repeated the axial coding process until a set of comparison factors and website matching was found to which all involved researchers agreed. The focus of coding was on the *functional* aspects of the websites, and less on the elements of the content or user interface since this study focuses on the consistency between websites and *not* rating the quality of each website's user journey.

V. Data Set

To gather a set of websites for our study, we relied on the open source project 2fa.directory [4], [5] that maintains a list of websites with 2FA support, which almost 1,000 contributors currently curate. The websites are assigned to different categories, such as social, communication, or retail. Since 2fa.directory distinguishes websites at the level of subdomains, we merged subdomains into their domain when we were aware that they use the same account for authentication. For example, drive.google.com, cloud.google.com, and mail.google.com are in different categories but rely on the same Google account, while amazon.com and aws.amazon.com have separate accounts. For merged entries, we chose the category we thought end users most likely knew the domain for (e.g., *mail* for google.com). Since we rely on a manual investigation of the user journeys of each website, we needed to reduce the set of all websites listed on 2fa.directory to a feasible number. First, we excluded categories for which we cannot create an account, for example, almost all websites in the *banking* and *government* categories. Second, we used the Tranco [44], [65] data set to rank websites according to their popularity. We selected the top websites from each category, where we selected the number of websites from each category based on the category's weight in the 2fa.directory data set. For example, there were only four *VPN provider* websites in the 2fa.directory set but 45 *Gaming* websites. This initially resulted in 120 websites. Unfortunately, we had to exclude 35 websites that we could not study for different reasons, such as language barriers, geo-restrictions, or the need for financial expenditures. In the end, we recorded the 2FA user journey on 85 websites with 2FA support from 26 categories.

VI. Comparison Factors

In this section, we explain the comparison factors that we identified in our analysis of 85 popular websites following the methodology of Section IV and describe informally how we categorize websites according to these factors. We apply the methodology of Bonneau et al. [9] by categorizing every website if it matches (●), partially matches (◐, ◑), or not matches (□) a factor. However, in our categorization, some factors are dependent on other factors, and we denote it explicitly when a conditional factor's prerequisite is not fulfilled (■) and this factor does not apply to a website. Further, in contrast to Bonneau et al., we do *not* use the categorization as a ranking to determine if a website is better than another website, but we use the categorization to identify patterns in how websites realize their 2FA user journey and to study whether websites realize this journey in a consistent way. Although, for some of the factors described below, this categorization overlaps with a scale from known best practices

to known poor practices from the literature. We found 22 comparison factors; 8 are conditional and depend on other factors to be applicable. Appendix D provides some examples of the different comparison factors and we provide additional examples in the appendix of [30].

A. Factors for Discovery

Promotion: The website promotes its 2FA support in a clear and obvious way during account creation or immediately after login (e.g., through a banner, pop-up, or highlighted message) (●). If the website does not clearly promote but only mentions the 2FA support in a way that could be easily missed by the user (for example, only a quick link in the footer of the landing page), we categorize this as *quasi*-promotion (◐). If the service does not promote its 2FA support and the user has to discover it themselves (e.g., browsing the settings pages), we categorize this as not matching (□).

Non-Optional: The website mandates setting up 2FA for user accounts (●). For instance, without setting a 2FA option up, the account registration cannot be completed; or after account registration, core functionality and features of the website are not available to the user until the user sets up 2FA for their account. Otherwise, using 2FA is optional and not mandatory for the website (□).

Common-Naming-and-Location: The website denotes its 2FA settings with a commonly used name, and the 2FA settings are in a commonly used location in the account settings (●). We identify commonly used names and locations in our analysis of our selected websites and summarize the results in Section VII-A. If either the name (◐) or the location (◑) is uncommon, we categorize this as *quasi-common-naming-and-location*. If the naming and location are uncommon, we categorize the website as not matching this factor (□).

B. Factors for Education

Descriptive-Notification: The website briefly describes what 2FA is in general or why it is important to users. The description is provided to the user *before* the user clicks to enable 2FA (●), e.g., located together with a notification about 2FA availability or within the settings page; or the description is only provided *after* the user starts the 2FA setup process (◐) at which point the user can still abort the setup. If the website does not present a description of 2FA, we categorize this website as not matching (□).

Additional-Information: The website provides more detailed information through a link (e.g. "learn more") to help users understand 2FA (●). If no such information is provided or the link is broken, the factor does not match (□).

C. Factors for Setup

Option-Specific-Information: The website provides specific information about all 2FA options it supports (●). For instance, it informs the user that TOTP or Push-notifications require the installation of an app or that WebAuthn requires a hardware authenticator. If the website does not provide this information but directly starts the setup process (e.g., asking users to scan a QR code or to use a security key without further explanation), this factor does not match (□).

Step-Wise-Instructions: The website gives an overview of the steps involved in setting up a specific 2FA option (e.g., linking a device or app, verifying the link, setting a recovery option) and/or details the instructions for each step for all 2FA options (●). Otherwise, this factor does not match (□).

Multiselection: The website offers multiple 2FA options (or setting up one method multiple times) and allows the user to set up multiple 2FA options (●), e.g., TOTP and Push-notification or multiple security keys. If the website supports multiple 2FA options but only allows the user to select one non-repeated option, we categorize this as *quasi-multiselection* (◑). This factor does not match (□) if the website only offers a single, one-time configurable option.

Grouped-Setting: The website's user settings present the 2FA options grouped, and users have a single setting location to manage all their 2FA options (●), e.g., all under the same settings tab. If the 2FA settings are split between different sections of the settings, we consider this to be not matching this factor (□). For instance, the management of security keys is organizationally separated from managing other 2FA options and, hence, might not be obvious to users. This factor depends on *Multiselection* being (quasi-)matched.

No-Enforced-Options: The website immediately presents all supported 2FA options to the user and allows them to choose their options themselves (●). If the website mandates the setup of specific 2FA options before the user can set up other options, we consider this not to match this factor (□). For example, the user must configure SMS-based 2FA before having the possibility to configure TOTP or WebAuthn options. This factor depends on *Multiselection* being (quasi-)matched.

Selectable-Primary-Option: If the website allows the configuration of multiple 2FA options and allows the user to select a primary option, which is the first option requested during login before falling back to other configured options (or recovery), we consider this a match (●). If the website does not support setting a user-selected primary 2FA option, we consider this not matching (□). This factor depends on *Multiselection* being matched.

Settings-Changed-Verification: The website requires the user to verify their identity before being able to change the 2FA settings (●). Otherwise, this factor does not match (□).

Settings-Changed-Notification: The website notifies the user about the changed 2FA settings via an out-of-band channel, e.g., by email or push notification (●). If there is no notification, this website does not match this factor (□).

Confirm-Successful-Setup: The website requires the user to confirm the 2FA authentication to complete the setup successfully and provides clear messaging about the successful setup for all options (●). For example, the user must enter the current TOTP or confirm a push notification to complete the setup, and the website shows a highlighted message in the settings. If the messaging is missing, but confirmation is required, we consider this as *quasi-confirm-successful-setup* (◑). If the website does not require confirmation (for all options), this website does not match this factor (□).

Informed-2FA-Recovery-Options: The website offers dedicated recovery options (such as one-time codes or asking to set up multiple 2FA options) and explains to the user why configuring dedicated 2FA recovery options is important for preparing for cases where the default 2FA options are not available, e.g., to prevent account lockout due to a lost or broken authentication device (●). If the website offers such recovery options but does not explicitly inform the user about their benefits and importance, we consider this as *quasi-informed-recovery-options* (◑). If the website does not offer explicit 2FA recovery options (e.g., it relies on a general account recovery or customer support), we consider this as not matching this factor (□).

Enforced-2FA-Recovery-Setup: Setting up recovery options is a mandatory step in setting up 2FA for this website (●), and the user cannot finish or continue setting up 2FA unless they set up the recovery option first. For example, the user has to confirm that they printed one-time backup codes to finish the 2FA setup or the website enforces setting up multiple 2FA options with a clear hint at account recovery. If setting up recovery options is not mandatory, but the website nudges users or strongly recommends them to set up a recovery option, we consider this *quasi-mandatory-recovery-setup* (◑). If setting up dedicated recovery options is at the user's discretion (without nudging or recommending), we consider this factor not matching (□). This factor depends on *Informed-2FA-recovery-options* being (quasi-)matched.

D. Factors for Usage

Device-Remembrance: The website offers a device remembrance during login, such that the user does not have to use 2FA on subsequent logins on the same device (e.g., "remember this device" checkbox). If the website automatically sets device remembrance without involving the user, e.g., during the first login after 2FA setup or during 2FA setup, we categorize this as ●. If device remembrance is at the discretion of the user and is stated as opt-*out* (e.g., an unchecked checkbox described as "ask me again on this device" or a pre-ticked checkbox "trust this device"), we categorize this as ◑. If device remembrance is stated as opt-*in* (e.g., "trust this device" checkbox that was not pre-checked), we categorize this as ◑. If device remembrance is not offered, we categorize as □.

No-Preselected-Option: If the website supports more than one active 2FA option at a time and no primary method is set (or could be set), how does the website present the configured 2FA options to their end users: the website shows all configured 2FA options at the same time during login (●), e.g., as a drop-down list. Alternatively, the website selected the primary option based on internal metrics (□), e.g., a security policy or the user's usage history. This preselection is usually intransparent to the user. This factor depends on *Multiselection* being matched and *Selectable-primary-option* not being matched.

E. Factors for Deactivation

Informed-Deactivation: The website allows the user to deactivate 2FA options and also explains to the user the potential risks associated with this (●) or does not provide any explanation or warning (◑). If the website does not allow the user to deactivate two-factor authentication, we consider this a mismatch for this factor (□).

Deactivation-Verification: The website requires the user to verify their identity before being able to deactivate a 2FA option (⬤). If a 2FA option can be disabled by the user without further authorization, we consider this website not to match this factor (□). This factor depends on *Informed-deactivation* being (quasi-)matched.

Deactivation-Notification: The website notifies the user about the deactivated 2FA option via an out-of-band channel, e.g., by email (⬤). If there is no notification, this website does not match this factor (□). This factor depends on *Informed-deactivation* being (quasi-)matched.

Communicate-Successful-Deactivation: The website communicates successful deactivation to the user as part of its user interface (⬤), e.g., highlighted message or pop-up. Otherwise, we consider this website not to match this factor (□). This factor depends on *Informed-deactivation* being (quasi-)matched.

VII. Results

We first provide an overview of the collected data (Section VII-A), followed by exploratory data analysis of the comparison factors (Section VII-B). Lastly, we discuss the results of qualitative data analysis of our observations (Section VII-C).

A. Overview of Website Data

Table I summarizes how each of the 85 websites in our data set matches the 22 comparison factors that we identified. We will explore these data further in the following sections. Table II summarizes the naming and location of the 2FA settings, the type of 2FA description, and the forms of device remembrance. Table III provides the codebook for device remembrance descriptions. More details per website are provided in [30], including Tranco [44], [65] rank and website category according to 2fa.directory [4], [5]. In summary, we found that 73 (86%) of the websites use a combination of "two-factor"/"two-step"/"multiple-factor" with "authentication"/"verification" for the naming, and on 78 (92%) websites, the 2FA settings are located in the security settings of the account settings under similar paths (e.g., "Security," "Login security," or "Authentication"). We considered those names and locations the common naming and location during our evaluation of the *Common-Naming-and-Location* factor. Of the 75 websites that describe 2FA in their settings, 69 (92%) describe 2FA in the form of "an additional layer of security," while 6 websites describe the 2FA mechanism with a focus on the user device (e.g., "we ask for additional authentication when logging in from a device that we do not know"). Only 31 websites in our data set offered a device remembrance feature, and half of those (16; 52%) describe this feature in terms of remembering the device or client (e.g., "Do not require OTP on this browser" or "Do not ask again on this device"). Almost a third (9; 29%) describe it in terms of trust (e.g., "Trust this device for {duration}"), and only four websites (13%) phrase it as skipping the additional step (e.g., "We won't ask for the next {duration}"). We also noticed that websites have a mixed strategy for phrasing the user's choice (i.e., opt-*in* versus opt-*out*), which we encoded in our factor *Device-remembrance* in Table I (i.e., ◑ vs. ◑).

B. Exploratory Data Analysis

Our factors allow us to compare the 2FA user journeys of different websites. We first explore our collected data (in Table I) through similarity analysis and clustering to gain insight into the overall consistency of those journeys on different websites and to identify potential clusters of websites that follow similar design patterns for their 2FA user experience.

1) Website similarity and factor consistency: To get a general impression of how similar the journeys are on the 85 websites, we compared them pairwise. Since our comparison factors are feature vectors of nominal (i.e., categorical) variables for each website, there is no intrinsic ordering and no equal space between variable values to measure the distance between values. We used the Hamming distance between pairs of websites as a measure of similarity. Since our variables have only values between 2–4, Hamming distance (i.e., "overlap without weights") is the most efficient measure of similarity for our data to obtain an overall impression of consistency between websites instead of measures that consider the number and/or frequency of values per variable, such as (Inverse) Occurrence Frequency, Goodall [33], or Eskin et al. [24]. To avoid artificial inflation of similarity from unfulfilled conditional factors, we calculate the Hamming distance only for the 14 non-conditional factors. We find that the average website in our data set differs in 6–7 of those 14 factors from the other websites, indicating that from a bird's-eye view, the user journeys are not very consistent across those websites. Further details about the frequency distribution of the pairwise Hamming distances are provided in Appendix B.

Furthermore, we measured the consistency of individual, non-conditional factors across all websites using Shannon entropy. High entropy means high inconsistency, whereas low entropy implies high consistency. The results are summarized in Table IV. Since some factors can also *quasi*-match (◑/◑) and, thus, have a different maximum entropy from binary (two-point scale) factors with only ⬤ and □, we distinguish between the point scales for each factor. The maximum possible entropy for each scale is indicated in column *Max ent*. Noticeable outliers with high entropy, i.e., low consistency, are *Multiselection* and most of the two-scale factors, which are close to the highest possible entropy. For example, *Multiselection* is almost evenly split (34×⬤, 28×◑, 23×□). In contrast, *Non-optional* is very consistent (6×⬤, 79×□) and *Common-Naming-and-Location* shows a strong tendency (67×⬤, 6×◑, 11×◑, 1×□). In summary, we found that none of the factors exhibit high consistency, except for *Non-optional* 2FA and *Common-Naming-and-Location*.

2) Factor clusters: Since our data do *not* indicate a "global consistency," we explore further whether there exist clusters of websites that have close similarities to each other but are more dissimilar from others. We applied a two-stage clustering process: first, we cluster websites based on their non-conditional factors (*inter-cluster*) and, additionally, assign each website to a subcluster based on the conditional factors (*intra-cluster*). Our intention for this two-stage process was that inter-clusters based on non-conditional factors provide the primary view of the different strategies for the 2FA user journeys across all websites, while additional intra-clusters based on conditional factors could support a more differentiated discussion of the overall strategies. Since our comparison factors are nominal

TABLE I: Comparison of popular websites based on the factors introduced in Section VI and clustering described in Section VII.

Website	Subcluster	Category	Discovery			Education			Setup													Usage		Deactivation			
			Promotion	Non-optional	Common-Naming-and-Location	Descriptive-notification	Additional-information	Option-specific-information	Stepwise-instructions	Multiselection	Grouped-setting	No-enforced-option	Selectable-primary-option	Settings-changed-verification	Settings-changed-notification	Confirm-successful-setup	Informed-2FA-recovery-options	Enforced-2FA-recovery-setup	Device-remembrance	No-preselected-option	Informed-deactivation	Deactivation-verification	Deactivation-notification	Communicate-successful-deact.

Cluster 1 ($n = 30$)

Website	Subcluster	Category
airvpn.org	1	VPN Providers
booking.com	1	Hotels/Accom.
clickup.com	1	Task Management
clio.com	1	Legal
digicert.com	1	Security
instagram.com	1	Social
laravel.com	1	Cloud Computing
mega.io	1	Backup and Sync
orcid.org	1	Identity Management
runsignup.com	1	Health
teamviewer.com	1	Remote Access
toodledo.com	1	Task Management
1password.com	2	Identity Management
airtable.com	2	Task Management
arlo.com	2	IoT
easydns.com	2	Domains
gitlab.com	2	Developer
roboform.com	2	Identity Management
bitdefender.com	3	Security
blockchain.info	3	Cryptocurrencies
coned.com	3	Utilities
facebook.com	3	Social
hover.com	3	Domains
join.me	3	Remote Access
jottacloud.com	3	Backup and Sync
kraken.com	3	Cryptocurrencies
logmein.com	3	Remote Access
mailchimp.com	3	Communication
namecheap.com	3	Domains
xero.com	3	Finance

Cluster 2 ($n = 29$)

Website	Subcluster	Category
bitwarden.com	1	Identity Management
blizzard.com	1	Gaming
callcentric.com	1	Utilities
clubhouse.io	1	Task Management
icloud.com	1	Backup and Sync
keepersecurity.com	1	Identity Management
kickstarter.com	1	Crowdfunding
realvnc.com	1	Remote Access
reddit.com	1	Social
roblox.com	1	Gaming
synology.com	1	Backup and Sync
virustotal.com	1	Security
adobe.com	2	Other
backblaze.com	2	Backup and Sync
bybit.com	2	Cryptocurrencies
docusign.com	2	Legal

Discovery | **Education** | **Setup** | **Usage** | **Deactivation**

Column factors (left to right):
Promotion, Non-optional, Common-Naming-and-Location, Descriptive-notification, Additional-information, Option-specific-information, Stepwise-instructions, Multiselection, Grouped-setting, No-enforced-option, Selectable-primary-option, Settings-changed-verification, Settings-changed-notification, Confirm-successful-setup, Informed-2FA-recovery-options, Enforced-2FA-recovery-setup, Device-remembrance, No-preselected-option, Informed-deactivation, Deactivation-verification, Deactivation-notification, Communicate-successful-deact.

Cluster 2 (continued)

Website	Subcluster	Category
dropbox.com	2	Backup and Sync
ea.com	2	Gaming
evernote.com	2	Backup and Sync
lastpass.com	2	Identity Management
norton.com	2	Security
stripe.com	2	Payments
twitch.tv	2	Entertainment
unity.com	2	Developer
vk.com	2	Social
zoho.com	2	Email
zoom.us	2	Communication
cloudflare.com	3	Security
tumblr.com	3	Social

Cluster 3 (n = 4)

Website	Subcluster	Category
23andme.com	1	Health
discord.com	1	Communication
opera.com	1	Other
gandi.net	3	Domains

Cluster 4 (n = 9)

Website	Subcluster	Category
atlassian.com	1	Developer
basecamp.com	2	Communication
binance.com	2	Cryptocurrencies
bitfinex.com	2	Cryptocurrencies
digitalocean.com	2	Cloud Computing
google.com	2	Email
newegg.com	2	Retail
epicgames.com	3	Gaming
id.me	3	Identity Management

Cluster 5 (n = 8)

Website	Subcluster	Category
github.com	1	Developer
meistertask.com	1	Task Management
youneedabudget.com	1	Finance
ifttt.com	2	IoT
playstation.com	2	Gaming
twitter.com	2	Social
etsy.com	3	Retail
va.gov	3	Government

Cluster 6 (n = 5)

Website	Subcluster	Category
amazon.com	2	Retail
ebay.com	2	Retail
linkedin.com	2	Social
paypal.com	2	Payments
yahoo.com	2	Email

● = factor matches; □ = factor does not match; ■ = does not apply

◐ = factor quasi matches (for factors with 3-point and 4-point scales); ◑ = factor quasi matches (only for factors with 4-point scales)

Gray column headers indicate conditional factors dependent on other factors to be applicable

TABLE II: Naming, location, descriptions of 2FA, and description of remembrance (where applicable).

Name of 2FA in website settings	
Two-Factor Authentication (2FA)	42 (49.41%)
Two-Step Verification (2SV)	24 (28.24%)
Other	12 (14.12%)
Multi-Factor Authentication (MFA)	4 (4.71%)
Two-Step Authentication (2SA)	3 (3.53%)
Location of 2FA settings	
Security / Account	78 (91.76%)
Other	7 (8.24%)
Focus of 2FA description	
Security	69 (92.00%)
Device	6 (8.00%)
Description of device remembrance	
Remember	16 (51.61%)
Trust	9 (29.03%)
Skip	4 (12.90%)
Other	2 (6.45%)

TABLE III: Codebook for device remembrance

Code	Examples
Remember	Remember verification for this computer
	Recognize this device in the future
	Do not require OTP on this browser
	Skip two-factor authentication on this device
	Save browser
	Do not ask again on this device
	Remember this device
	Remember this computer for {duration}
	Do not ask for code on this device
Trust	Trust this device (opt-in)
	Trust this device (opt-out)
	Do not trust this device (opt-out)
	Do not trust this device (opt-in)
	Trust this device for {duration}
	Untrust this device
Skip	Require code to login for {duration}
	We won't ask for the next {duration}
	Skip this for {duration}
Other	Stay signed/logged in

opt-out: checkbox is pre-checked; *opt-in*: checkbox is not pre-checked
duration: a number of days, weeks, or logins

variables, we apply k-modes [12] clustering in both stages. For inter-clustering, Silhouette testing [59] indicated that 2, 5, or 6 clusters fit the data best, and we decided on 6 clusters due to the best descriptive performance of those clusters. For the intra-clustering of the conditional factors, we found 3 clusters to best describe the data. The result of the final clustering is noted in Table I. Appendix C provides less noisy views of the cluster structures.

When comparing the characteristics of the *inter-clusters*, we find three aspects that differentiate the clusters the most: how they inform and instruct their users, how they offer support for multiple 2FA options, and whether they offer device remembrance. In terms of informing and instructing users, the six clusters can be combined into two larger clusters. Websites in *Clusters 1, 2* and *3* generally do not verify or notify about changes in 2FA settings (except for *Cluster 2*), omit

TABLE IV: Shannon entropy of each non-conditional factor

	Comparison Factor	$H(X)$	Max ent.
Two-point scale	Non-optional	0.37	1.0
	Additional-information	0.90	
	Option-specific-information	0.99	
	Stepwise-instructions	0.87	
	Settings-changed-verification	0.99	
	Settings-changed-notification	1.00	
Three-point scale	Promotion	1.12	1.57
	Descriptive-notification	1.11	
	Multiselection	1.57	
	Confirm-successful-setup	1.24	
	Informed-2FA-recovery-options	1.26	
	Informed-deactivation	1.05	
Four-point scale	Common-Naming-and-Location	1.00	2.0
	Device-remembrance	1.60	
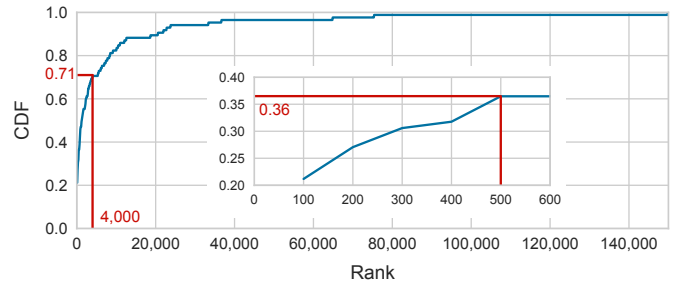
Fig. 2: Cumulative frequency distribution of Tranco [44], [65] rankings of our 85 websites. Percentiles for 0.36 (31 websites) and 0.71 (60 websites) are marked, which correspond to websites in the top-500 and in the top-4000 in Tranco.

additional information, do not give step-wise instructions (with the exception of *Cluster 3*), and often do not provide specific information about 2FA options. In contrast, *Clusters 4, 5* and *6* provide this information and instructions more regularly and, in addition, the websites in *Cluster 4* warn users about the deactivation of 2FA. Alternatively, the six *inter-clusters* could be combined into two groups based on their strategy to support multiple 2FA options. Websites in *Clusters 1, 5* and *6* usually allow only one option to be activated simultaneously, although they usually offer multiple options. In contrast, websites in *Clusters 2, 3* and *4*, when supporting multiple 2FA options, usually allow users to choose between multiple activated 2FA options for login. Lastly, regarding device remembrance, the websites in *Clusters 1, 3, 4* and *5* have in common that they mostly do not offer device remembrance for future logins. *Clusters 2* and *6* usually offer this.

Regarding *intra-clusters*, *Subcluster 1* websites do not usually provide a selection of multiple 2FA options, and when they do, they enforce certain 2FA options. The websites in *Subclusters 2* and *3* support multiple 2FA options but differ in their strategy to enforce certain 2FA options and verify 2FA deactivation. Websites in *Subcluster 2* almost always verify 2FA deactivation, while websites in *Subcluster 3* do not enforce certain 2FA options. Unfortunately, *Clusters 3* to *6* are too small to reliably comment on the relationship between *inter-clusters* and *intra-cluster*.

TABLE V: Contingency table for cluster vs. rank category

Cluster	Top-500	Top-4000	Long tail	$\sum$
1	6	10	14	30
2	13	9	7	29
3	2	1	1	4
4	2	6	1	9
5	3	3	2	8
6	5	0	0	5
$\sum$	31	29	25	85

TABLE VI: Opinionated separation of comparison factors

Category	Conditional	Factors
User Experience	No	Promotion
	No	Common-Naming-and-Location
	No	Descriptive-Notification
	No	Option-Specific-Information
	Yes	Grouped-Setting
	Yes	Selectable-Primary-Option
	No	Confirm-Successful-Setup
	No	Informed-2FA-Recovery-Options
	Yes	Enforced-2FA-Recovery-Setup
	No	Device-Remembrance
	Yes	No-Preselected-Option
	Yes	Communicate-Successful-Deactivation
Security	No	Settings-Changed-Verification
	Yes	Deactivation-Verification
Both Security and	No	Non-optional
User Experience	No	Step-Wise-Instructions
	No	Multiselection
	Yes	No-Enforced-Options
	No	Settings-Changed-Notification
	No	Informed-Deactivation
	Yes	Deactivation-Notification
Neither	No	Additional-Information

3) Clusters vs. Website Ranks: We divide the websites in our data set into three roughly equal-sized groups through the 36th and 71st percentiles of the websites' Tranco ranks. Figure 2 illustrates the CFD of the Tranco ranking in our data set. Based on this CFD, the first group of websites ($n = 31$) is in the *Top-500* of Tranco, the second group of websites ($n = 29$) ranks between 501 and 4,000 (denoted as *Top-4000*), and the third group ($n = 25$) is the *"long tail"* with a rank greater than 4,000. Since we initially selected the most popular websites in each category, this distribution is naturally heavily skewed toward the top ranks. We then used the *inter-*cluster to describe each website's underlying 2FA user flow, which we analyzed for an association with the website Tranco rank group. Table V shows the contingency table for cluster vs. rank. Fisher's exact test ($p = 0.04388$) shows that this association is statistically significant.

We also considered the association between website categories and clusters, but unfortunately, the website categories are too diverse, and the number of websites per category is too small to derive a meaningful connection between cluster and category. We are also unaware of any reliable, more coarse-grained website categorization that could be used.

4) Opinionated Separation of Comparison Factors: Our analysis considered all the comparison factors at once and did not differentiate between different categories of factors.

To provide a different view on the consistency of 2FA user journeys, we conducted an expert evaluation of our factors to create an opinionated separation of factors by their impact on security, user experience, both, or neither. The entire evaluation process is described in [30] and Table VI summarizes the results. We split our comparison factors into four disjoint sets: *Non-conditional-UX* (7 factors), *Non-conditional-Security* (6 factors), *Conditional-UX* (5 factors), and *Conditional-Security* (3 factors). Only the factor *Additional-information* was considered irrelevant for UX and security. We repeated the data analysis of Sections VII-B1 and VII-B2 for those four sets.

Pairwise Hamming distance: Considering only *non-conditional-UX* factors, the average website differs in 3–4 of the 7 factors from other websites, and considering only *non-conditional-security* factors, the average website differs in 2–3 of the 6 factors from other websites. Thus, with this distance metric, the websites in our data set do not show better consistency when considering separated sets of factors.

Factor clusters: For each set of factors, we calculate the mean Silhouette coefficient for different numbers of clusters with KModes to determine the best number of clusters to describe our data set. Compared to clustering with all factors, we found that the best-fitting number of clusters is larger when considering our separated factors. For *Non-conditional-UX* comparison factors, we found 5 clusters, and for *Conditional-UX* comparison factors, 10 clusters. For *Non-conditional-security* comparison factors, we calculated 9 clusters as the best number of clusters. For *Conditional-security* comparison factors, Silhouette testing showed 8 to be the best number of clusters. As a result, considering sets of separated factors, we found more diverse strategies for how websites in our data set implement their 2FA user journeys with regard to purely UX or security. We illustrate the corresponding clusters in [30].

C. Qualitative Data Analysis

We discuss the consistencies and inconsistencies we observed during our analysis of the 2FA user journeys.

1) Consistent Discovery for Self-Motivated Users: Our analysis shows that the vast majority of websites in our data set did not *immediately* promote 2FA to their end users in any form before/during sign-up and login—a website might promote 2FA only at a later point (e.g., an account existed for some time or the user takes actions that increase the severity of an account compromise), which our recording of journeys does not cover. The few websites that immediately promoted their 2FA support did this with mixed strategies, where most of them promoted 2FA during or immediately after account creation. In contrast, the remaining websites mentioned it only on their landing page, where users could easily miss it. However, we discovered that some websites' nudging to 2FA merely redirected the user to the account settings' security section, where the user has to pick up the journey themselves. Furthermore, six websites in our data set mandated 2FA, most of those sites in the cryptocurrency category. However, for two websites that mandate 2FA, we found that the intention to use the phone number or verified email address as a second factor was not clearly communicated to the user during account creation (e.g., Figure **??** in Appendix D).

Our analysis showed that users looking for 2FA settings have a consistent experience across websites. Almost all websites used a common location for their 2FA settings. Therefore, users who once went through the 2FA workflow can find the 2FA settings more easily on other websites. Most websites also use similar descriptions of 2FA (e.g., "second layer of security," "prevent unauthorized access," or "ask for authentication on new devices"), which helps the user to recognize the 2FA settings despite variations in the naming. Examples of clear exceptions to this pattern are illustrated in Figure 5 and Figure 7 in Appendix D.

2) Consistent Lack of Informing and Educating Users: We found that only a minority of the websites provided additional information (e.g., "learn more" link to detailed information including pictures and tutorials), and even fewer websites educate the user about the benefits and drawbacks of the 2FA options that they support, but instead immediately start the setup process. During this setup, only about a third of all websites guided the user with step-by-step instructions for setting up a chosen 2FA option. Most websites require the user to verify their identity to change their 2FA settings and inform them about such changes (e.g., by email). Very noticeable exceptions are the websites in Cluster 1, which almost entirely omit both verification and notification of settings changes.

The most consistent behavior we have observed to inform users is the confirmation of a successful setup. More than four-fifths of the websites required a successful confirmation from the user (e.g., the user had to enter the current TOTP code to complete the setup) and, in most cases, also provided some visual feedback to the user to inform them about the successfully concluded 2FA setup.

3) Mixed Strategies for 2FA Setup and Configuration: We observed the most inconsistent behavior when it comes to the setup of 2FA options and their possible configurations by the user. First, there is an almost even split between three basic strategies: "offering only one 2FA option," "offering multiple 2FA options but only one can be active at a time," and "offering multiple 2FA options and multiple can be active at the same time." Unfortunately, we could not find an explanation on any of the websites that let their users select only one 2FA option about why they implement 2FA this way. Second, among the websites that support multiple 2FA options, all but six websites show the 2FA options grouped in the same settings location, while those six exceptions, for instance, differentiate between 2FA and security keys in their security settings. However, thirdly, half of the websites with support for multiple 2FA options enforce a particular 2FA option to be set up before they offer the other options to the user. For example, only after providing their phone number can the user set up security keys or TOTP as an alternative. Fourth and last, websites are very consistent in proposing the 2FA option that should be used to login. Very few websites allow the user to select the 2FA option that should be used primarily for the login. Only a single website asked the user upfront during login which 2FA option they would like to use for the current login (see Figure 6 in Appendix D). The vast majority of websites used internal metrics to determine which 2FA option should be used for login, and the user could only navigate through the "use a different method" or "do you have difficulties" menus to select another 2FA option.

4) Mainly Optional Recovery: Three-quarters of all websites offer recovery options, and most of those websites also explain to the user the importance of setting up recovery options or the risks of neglecting to set up a recovery option. The preferred recovery option among these websites was printable one-time codes. Also, websites are very consistent in enforcing the setup of a recovery option. Almost three-quarters of the websites with a recovery option nudge the user to set up the recovery, and only six websites enforce this during the 2FA setup. This low number of websites with mandatory recovery also means that there is no fail-safe account recovery strategy by websites that support at most one active 2FA option. It would be intuitive that such websites would enforce a recovery option to prevent account lockout in case this single option is unavailable. Still, our data do not support this. For the usage of recovery options, we found only one website that, although supporting one-time codes, does not offer an obvious way to use them (i.e., there was no link to a recovery page and no instructions to use the recovery codes as input to the regular OTP form field).

5) Mixed Strategies for Device Remembrance: More than half of the websites in our data set do not support device remembrance, i.e., the user cannot explicitly select to skip the second-factor authentication during future logins. For the websites that support this feature, we found that not only do they describe it in different ways but also that their remembrance logic differs. Almost two-thirds of the websites need the user to opt-*in* to this feature, a fifth of the websites needs the user to explicitly opt-*out* from remembering the device, and another fifth of the websites automatically places a device remembrance cookie without asking the user.

6) Consistent Support for Deactivation: All but five websites in our data set support the deactivation of 2FA (and only two of those exceptions mandate the setup of 2FA). However, only a minority of websites communicate to the user the risk of deactivating 2FA (e.g., easier account hijacking). Furthermore, similar to the previously mentioned lack of consistently informing and educating users about 2FA, we find that only about half of all websites verify the user identity before deactivation, notify the user about deactivation, or communicate a successful deactivation in the website settings.

VIII. Discussion and Future Work

Are the websites in our data set consistently following the same design patterns and strategies in offering 2FA? Although some factors individually show high consistency, we did not find a single start-to-end design pattern for the user journey that is consistently followed by the majority of websites in our data set. Instead, we found that websites are clustered into smaller groups with similar 2FA user journeys. Separating the comparison factors by their impact on UX and security did not indicate a more consistent strategy for pure usability or security-related steps along the user journey. In fact, we found that the websites were more clustered when considering separated sets of factors. Taking into account the rankings of websites, our results indicate that the design represented by clusters 2, 3, and 6 is more popular among the top-ranked websites, while more than 50% of the websites from the long-tail ranks are in cluster 1.

Implications for developers and users: UX guidelines state that users prefer a site to work the same way as all other sites they already know. This heuristic has been shown to be successful on the Web, for example, when it comes to online shopping or banking experiences. To follow this heuristic vis-a-vis two-factor authentication as a factor for the overall user experience on the Web, developers could follow the 2FA experience on the majority of websites or on the most popular websites, which likely minted the users' mental models. Our results show that such a majority does not exist among the popular websites and that even the leading websites (e.g., Google and Apple) do not agree in their user journeys. Therefore, a recommendation for influential industry associations and consortia would be to draft recommendations for website developers on how to achieve a consistent strategy for 2FA user journeys. Possible avenues for the community and future work to contribute to this endeavor could be to create guidelines that foster consistent strategies for implementing the best possible 2FA UX on the Web.

A crucial consideration when striving for consistency is that consistency in itself does not guarantee a good UX; a bad design could be consistently implemented, but users have learned to live with it. Krug [39] even advises to sacrifice consistency in favor of more clarity for end-users. As a concrete example from our data set, Apple's icloud.com is an outlier in various comparison factors: it mandates a phone number as the only 2FA option without clearly informing the user during account creation and without the option to add other options later or to deactivate it (see Appendix **??**). But do users perceive icloud.com's 2FA user journey as a bad or good experience? Our study design does not attempt to assign a quality measurement to individual factors, and it does not measure the quality of user experiences attached to the different clusters of user journeys. But clearly, our results motivate that the impact of the different strategies for the 2FA user journey on the perceived usability by users has to be thoroughly investigated in an effort to make the best strategy consistent across websites. Our data indicated a connection between the rank of a website and the site's strategy, but it is unclear to what extent the guidelines for 2FA journeys need to be contextualized. For example, certain comparison factors may be dictated or recommended by regulations, such as PCI-DSS or the EU Revised Payment Services Directive (PSD2) with strong customer authentication (SCA), different types of websites may have different security policies, or specific user groups [17], [48]) require different support. Thus, it is unclear whether consistency between *all* comparison factors is required or even desirable.

Indications for the external validity of prior works: Although we did not study the usability of individual instruments, comparison factors, or steps in the user journey, we can provide a new perspective on some aspects of previous work and UX guidelines we observed during our data analysis. We noted that the discovery of 2FA and the initial education of users are very consistent and that there is a common naming and description of 2FA in place. But neither of the two types of 2FA descriptions that we have noted in our analysis (see Table II) complies with recent results by Golla et al. [32] and Lassak et al. [41] on how users should be nudged and educated to encourage the adoption of 2FA. Furthermore, Ciolino et al. [13] conducted a user study of 2FA setup and login "in

the wild." Their participants encountered some of the patterns we identified in our work and described them as problematic. For instance, enforcing the SMS 2FA option while not communicating that additional 2FA options become available after registering the phone number confused participants that were explicitly looking for registration of security keys; an opt-out device remembrance, which we found on several websites (7×⬤, 8×◑), frustrated participants that were expecting to be prompted for a second factor on login but missed that they had to take explicit action for that; and their participants expressed the desire for personalization by being able to select the preferred 2FA option for logins, which we found is not a widespread feature but, on the contrary, the 2FA option is in most cases chosen by the website (only 8 out of 34 websites with fully matched *Multiselection* allowed setting a primary option, and only one website of the remaining 26 sites did *not* pre-select the option). Lastly, from our clustering, we noticed that recommendations by UX guidelines to provide adequate contextual help and break down complex tasks, in this case for setting up 2FA, were ignored on many websites that did not offer additional or option-specific information or simply step-wise instructions. Also, the recommendation to provide clear feedback to users was not realized on many websites that did not notify users or communicate a successful 2FA setup or deactivation. Thus, our study provides indications for the external validity of prior results. In our opinion, measuring to what extent each pattern we detected matches the recommendations and settings of related work would be an interesting follow-up study to provide better insights into the external validity of previous studies (e.g., taking textual content and UI designs into more consideration). Those indications also emphasize the need to establish more general UX guidelines for implementers of 2FA user journeys to improve the usability of 2FA. The first option-specific guidelines [27], [28] or collections of best-practices [22], [67] are a good starting point.

FIDO UX Guidelines [27], [28]: The FIDO Alliance UX guidelines also consider similar steps in the FIDO2 user journey (promotion, invitation, registration, and login). They recommend the promotion of biometric awareness or security keys at sign-in and registration, educating users about the FIDO value proposition of a "simple and secure sign-in without password" or about authentication with security keys, providing a "learn more" link and giving concrete statements based on user studies, confirming successful registration with a clear indication to users, encouraging users to register multiple keys for recovery and backup [28], and explicitly promoting "Security and Privacy settings" to manage 2FA options. Unfortunately, these guidelines are not suitable as a general guideline and, at some points, conflict with recent recommendations from research (e.g., the promotion message [32], [41] or automatically setting FIDO2 as the default sign-in option [13]). The guidelines [27] are strongly tailored to promote biometric authentication as a convenient alternative to passwords or to promote 2FA with security keys to consumers on regulated industry websites [28], such as banking or healthcare. They do not target a 2FA setting [27], and the guidelines do not address the setup and UX of multiple authentication options, or limit themselves to only security keys as the second factor [28]. For desktop authenticators [27], the password is considered a fallback option; therefore, these guidelines omit explicit recovery steps.

Limitations: Like any other qualitative study, our work also has some limitations. Despite our best efforts, we cannot exclude a subjective bias by the involved researchers, e.g., in identifying the comparison factors or selecting a clustering with the best descriptive power. We aimed to study 120 popular websites, but only 85 were possible due to various restrictions and obstacles. Thus, our study is skewed toward top websites in English language and from specific categories. We fixed the conditions for data collection to increase the internal validity of our data, but we cannot exclude that our setting is considered high-risk or low-risk by a website and that we experienced a different user journey than other users of the same site. Moreover, we collected our data only from desktop computers; thus, our comparison factors may differ on mobile devices. Furthermore, with the adoption of new technologies (e.g., Passkeys) and changes in website policies (e.g., Google plans to mandate 2FA for an increasing number of its users [3]), our comparison might not capture the most recent picture. However, we believe that our general results remain valid. Lastly, we did not continue to monitor the websites, nor did we explore the user flow for going through account recovery or changing 2FA-relevant information (e.g., phone number or email address), since we focused on the steps of the user journey that mint the users' initial impressions of 2FA.

Future work: Conducting user and developer studies is an obvious way to follow up on our results. While we detected a lack of consistency in the 2FA user journeys that can increase users' cognitive friction, it is unclear whether this contributes to the notoriously low adoption rate of 2FA among end-users. Our survey [30] indicated that several users did indeed refrain from setting up 2FA or deactivated it due to differences in the user experience between websites. Furthermore, it is unclear whether a "gold standard" for journeys exists or to what extent journeys need to be contextualized (e.g., website category, regulations, or specific user groups). Comparative studies of different design patterns could answer those questions and others, such as a weighting of comparison factors by their impact on, e.g., the UX or 2FA security. Moreover, we consider it worthwhile to explore developers' reasons for choosing a particular design pattern to understand the reasons behind those inconsistent journeys. In addition to human-centered studies, extending our methodology to user journeys for account recovery, to other device form factors, such as mobile devices, or to entirely new solutions, such as Passkey, would complement our results. Lastly, we think that studying the 2FA user journeys can provide insights into the external validity of (previous) studies of individual aspects of 2FA and shed new light on what constitutes a good 2FA user experience.

IX. CONCLUSION

This work contributes a methodology for comparing 2FA user journeys on websites and presents the first systematic study of the consistency of those journeys on top-ranked websites. Our results show a lack of consistency for the various steps along those journeys. We find that even the more consistent design patterns were described as problematic for usability in the literature. We strongly believe that our results motivate different future works that can lead to the creation of more general user experience guidelines for implementers of two-factor authentication.

REFERENCES

[1] "Interaction design foundation," https://www.interaction-design.org.

[2] "Nielsen norman group," https://www.nngroup.com/.

[3] "Making sign-in safer and more convenient," https://blog.google/technology/safety-security/making-sign-safer-and-more-convenient/, Oct. 2021.

[4] 2FA Directory, https://2fa.directory/.

[5] 2factorauth, https://github.com/2factorauth/twofactorauth.

[6] J. Abbott and S. Patil, *How Mandatory Second Factor Affects the Authentication User Experience*. ACM, 2020.

[7] All about UX, "User experience definitions," http://www.allaboutux.org/ux-definitions.

[8] Apple Inc., "Human interface guidelines," https://developer.apple.com/design/human-interface-guidelines/.

[9] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *Proc. 33rd IEEE Symposium on Security and Privacy (SP '12)*. IEEE, 2012.

[10] C. Braz and J.-M. Robert, "Security and usability: The case of the user authentication methods," in *Proc. 18th Conference on L'Interaction Homme-Machine (IHM '06)*. ACM, 2006.

[11] E. Bursztein, "The bleak picture of two-factor authentication adoption in the wild," https://elie.net/blog/security/the-bleak-picture-of-two-factor-authentication-adoption-in-the-wild/, Dec. 2018.

[12] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes clustering," *Journal of classification*, vol. 18, no. 1, pp. 35–55, 2001.

[13] S. Ciolino, S. Parkin, and P. Dunphy, "Of two minds about two-factor: Understanding everyday FIDO u2f usability through device comparison and experience sampling," in *Proc. 15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, 2019.

[14] J. Colnago, S. Devlin, M. Oates, C. Swoopes, L. Bauer, L. Cranor, and N. Christin, *"It's Not Actually That Horrible": Exploring Adoption of Two-Factor Authentication at a University*. ACM, 2018.

[15] A. Cooper, R. Reimann, D. Cronin, and C. Noessel, *About Face: The Essentials of Interaction Design*, 4th ed. Wiley John + Sons, 2014.

[16] S. Das, A. Dingman, and L. J. Camp, "Why johnny doesn't use two factor: A two-phase usability study of the fido u2f security key," in *Financial Cryptography and Data Security*, 2018.

[17] S. Das, A. Kim, B. Jelen, L. Huber, and L. J. Camp, "Non-inclusive online security: Older adults' experience with two-factor authentication," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2020.

[18] S. Das, S. Mare, and L. J. Camp, "Smart storytelling: Video and text risk communication to increase mfa acceptability," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, 2020.

[19] S. Das, B. Wang, and L. J. Camp, "Mfa is a waste of time! understanding negative connotation towards mfa applications via user generated content," in *Proceedings of the Thriteenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2019)*, 2019.

[20] S. Das, B. Wang, A. Kim, and L. J. Camp, "MFA is A necessary chore!: Exploring user mental models of multi-factor authentication technologies," in *53rd Hawaii International Conference on System Sciences, HICSS*, 2020.

[21] S. Das, B. Wang, Z. Tingle, and L. J. Camp, "Evaluating user perception of multi-factor authentication: A systematic review," in *roceedings of the Thirteenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2019)*. Springer, 2019.

[22] D. Dias, "9 best practices and ux improvements for the two-factor authentication (2fa)," https://thedaviddias.dev/blog/9-best-practices-ux-for-two-factor-authentification/.

[23] J. Dutson, D. Allen, D. Eggett, and K. Seamons, "Don't punish all of us: Measuring user attitudes about two-factor authentication," in *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2019.

[24] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, *A Geometric Framework for Unsupervised Anomaly Detection*. Boston, MA: Springer US, 2002, pp. 77–101.

[25] M. Fagan and M. M. H. Khan, "Why do they do what they do?: A study of what motivates users to (not) follow computer security advice," in *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

[26] F. M. Farke, L. Lorenz, T. Schnitzler, P. Markert, and M. Dürmuth, ""you still use the password after all" – exploring fido2 security keys in a small company," in *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, 2020.

[27] FIDO Alliance, "Fido desktop authenticator ux guidelines (v1)," Jun. 2021.

[28] ——, "Fido security key ux guidelines," Jun. 2022.

[29] S. Frazier, "The 2019 state of the auth report: Has 2fa hit mainstream yet," https://duo.com/blog/the-2019-state-of-the-auth-report-has-2fa-hit-mainstream-yet, Dec. 2019.

[30] S. Ghorbani Lyastani, M. Backes, and S. Bugiel, "A systematic study of the consistency of two-factor authentication user journeys on top-ranked websites (extended version)," 2022. [Online]. Available: https://arxiv.org/abs/2210.09373

[31] S. Ghorbani Lyastani, M. Schilling, M. Neumayr, M. Backes, and S. Bugiel, "Is fido2 the kingslayer of user authentication? a comparative usability study of fido2 passwordless authentication," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020.

[32] M. Golla, G. Ho, M. Lohmus, M. Pulluri, and E. M. Redmiles, "Driving 2fa adoption at scale: Optimizing two-factor authentication notification design patterns," in *Proc. 30th USENIX Security Symposium (SEC' 21)*. USENIX Association, 2021.

[33] D. W. Goodall, "A new similarity index based on probability," *Biometrics*, vol. 22, p. 882, 1966.

[34] Google, "Material design," https://material.io/design.

[35] N. Gunson, D. Marshall, H. Morton, and M. Jack, "User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking," *Computers & Security*, vol. 30, no. 4, pp. 208 – 220, 2011.

[36] International Organization for Standardization, "Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems," International Organization for Standardization, Standard ISO 9241-210:2019, Jul. 2019.

[37] R. Krause, "Maintain consistency and adhere to standards (usability heuristic #4)," https://www.nngroup.com/articles/consistency-and-standards/.

[38] K. Krol, E. Philippou, E. D. Cristofaro, and M. A. Sasse, ""they brought in the horrible key ring thing!" analysing the usability of two-factor authentication in uk online banking." in *Workshop on Usable Security and Privacy (USEC'15))*. The Internet Society, 2015.

[39] S. Krug, *Don't Make Me Think: A Common Sense Approach to Web Usability*. New Riders, 2013.

[40] J. Lang, A. Czeskis, D. Balfanz, M. Schilder, and S. Srinivas, "Security keys: Practical cryptographic second factors for the modern web," in *Financial Cryptography and Data Security*, 2017.

[41] L. Lassak, A. Hildebrandt, M. Golla, and B. Ur, ""it's stored, hopefully, on an encrypted server": Mitigating users' misconceptions about fido2 biometric webauthn," in *Proc. 30th USENIX Security Symposium (SEC' 21)*. USENIX Association, 2021.

[42] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, *Understanding, Scoping and Defining User Experience: A Survey Approach*. ACM, 2009.

[43] J. Lazar, J. H. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction*, 2nd ed. Morgan Kaufmann, 2017.

[44] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *Proc. 26th Annual Network and Distributed System Security Symposium (NDSS '19)*. The Internet Society, 2019.

[45] K. Lee, S. Sjöberg, and A. Narayanan, "Password policies of most top websites fail to follow best practices," in *Proc. 18th Symposium on Usable Privacy and Security (SOUPS'22)*. USENIX Association, 2022.

[46] A. Mirian, "Hack for hire," *Commun. ACM*, vol. 62, no. 12, p. 32–37, Nov. 2019.

[47] D. M'Raihi, S. Machani, M. Pei, and J. Rydell, "Totp: Time-based one-time password algorithm," Internet Requests for Comments, RFC Editor, RFC 6238, May 2011, http://www.rfc-editor.org/rfc/rfc6238.txt. [Online]. Available: http://www.rfc-editor.org/rfc/rfc6238.txt

[48] D. Napoli, K. Baig, S. Maqsood, and S. Chiasson, ""i'm literally just hoping this will work:'' obstacles blocking the online security and privacy of users with visual disabilities," in *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, 2021.

[49] J. Nielsen, "10 usability heuristics for user interface design," https://www.nngroup.com/articles/ten-usability-heuristics/.

[50] ——, "End of web design," https://www.nngroup.com/articles/end-of-web-design/.

[51] ——, "Jakob's law of internet user experience," https://www.nngroup.com/videos/jakobs-law-internet-ux/.

[52] ——, "Enhancing the explanatory power of usability heuristics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, 1994, p. 152–158.

[53] D. Norman and J. Nielsen, "The definition of user experience (ux)," https://www.nngroup.com/articles/definition-user-experience/.

[54] K. Owens, O. Anise, A. Krauss, and B. Ur, "User perceptions of the usability and security of smartphones as fido2 roaming authenticators," in *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, 2021.

[55] E. M. Redmiles, N. Warford, A. Jayanti, A. Koneru, S. Kross, M. Morales, R. Stevens, and M. L. Mazurek, "A comprehensive quality evaluation of security and privacy advice on the web," in *Proc. 29th USENIX Security Symposium (SEC' 20)*. USENIX Association, 2020.

[56] K. Reese, T. Smith, J. Dutson, J. Armknecht, J. Cameron, and K. Seamons, "A usability study of five two-factor authentication methods," in *Proc. 15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, 2019.

[57] J. Reynolds, N. Samarin, J. Barnes, T. Judd, J. Mason, M. Bailey, and S. Egelman, "Empirical measurement of systemic 2fa usability," in *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020.

[58] J. Reynolds, T. Smith, K. Reese, L. Dickinson, S. Ruoti, and K. Seamons, "A tale of two studies: The best and worst of yubikey usability," in *Proc. 39th IEEE Symposium on Security and Privacy (SP '18)*. IEEE, 2018.

[59] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[60] SciPy Developer Documentation, "Statistical functions: normaltest," https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html.

[61] B. Shneiderman, "The eight golden rules of interface design," https://www.cs.umd.edu/~ben/goldenrules.html.

[62] ——, *Designing the user interface : strategies for effective human-computer interaction*, 4th ed. Pearson/Addison Wesley, 2004.

[63] D. Strouble, G. m. Shechtman, and A. S. Alsop, "Productivity and usability effects of using a two-factor security system," in *SAIS*, 2009.

[64] A. Tetlay, H. Treharne, T. Ascroft, and S. Moschoyiannis, "Lessons learnt from a 2fa roll out within a higher education organisation," *CoRR*, vol. abs/2011.02901, 2020.

[65] Tranco, "A research-oriented top sites ranking hardened against manipulation," https://tranco-list.eu/.

[66] J. Weidman and J. Grossklags, "I like it, but i hate it: Employee perceptions towards an institutional transition to byod second-factor authentication," in *Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC '17)*. ACM, 2017.

[67] J. Weiler, "Two-factor authentication – better user experience through security," https://amiconsult.de/en/better-user-experience-through-security/, Dec. 2020.

[68] A. Weinert, "Your pa$$word doesn't matter," https://techcommunity.microsoft.com/t5/azure-active-directory-identity/your-pa-word-doesn-t-matter/ba-p/731984, Jul. 2019.

[69] S. Weinschenk, *100 Things Every Designer Needs to Know about People*, 2nd ed. New Riders, 2020.

[70] C. S. Weir, G. Douglas, M. Carruthers, and M. Jack, "User perceptions of security, convenience and usability for ebanking authentication tokens," *Computers & Security*, vol. 28, no. 1, pp. 47 – 62, 2009.

[71] C. S. Weir, G. Douglas, T. Richardson, and M. Jack, "Usable security: User preferences for authentication methods in ebanking and the effects of experience," *Interacting with Computers*, vol. 22, no. 3, pp. 153 – 164, 2010.

[72] S. Wiefling, L. Lo Iacono, and M. Dürmuth, "Is This Really You? An Empirical Study on Risk-Based Authentication Applied in the Wild," in *34th IFIP TC-11 International Conference on Information Security and Privacy Protection (IFIP SEC 2019)*. Springer International Publishing, 2019.

[73] World Wide Web Consortium, "Web authentication: An api for accessing public key credentials level 2 — w3c recommendation, 8 april 2021," https://www.w3.org/TR/webauthn/, Mar. 2021.

[74] J. Yablonski, "Laws of ux," https://lawsofux.com/.

[75] ——, *Laws of UX: Design Principles for Persuasive and Ethical Products*. O'Reilly UK Ltd., 2020.

Appendix A
Summary of our survey among 2FA users

We conducted a survey among 2FA users to gather user experiences with different 2FA journeys and gain insights into whether users had negative experiences transferring their 2FA experiences between websites and whether this has stopped them from enabling or continuing to use 2FA. We recruited our participants through the *Prolific*[1] platform. Prolific collects basic demographic information[2] about their participant pool, to which we added a pre-screening question to select only participants that stated that they use 2FA on at least two different websites. We used Prolific to create a pre-screened participant pool whose demographics are representative of the US population and that has an approval rate of more than 90% on Prolific. In the end, 309 participants successfully completed the survey and only one participant had to be excluded from the final data set.

Regarding the question of whether users had negative experiences transferring their 2FA experiences between websites and whether this has stopped them from enabling or continuing to use 2FA, 28 (9.1%) of our 308 participants mentioned for at least one website, which differed in their opinion from others in its 2FA experience, that they use this website less due to this 2FA experience. Furthermore, 41 (15.9%) of the participants recalled a concrete situation with 2FA that was challenging because the 2FA experience differed from what they were used to and, as a result, they abandoned the website or refused to adopt a (specific) 2FA option. Taken together, 60 (19.5%) of our participants reported using a website less, abandoning a website, or refusing adoption of (a specific) 2FA option. Of these, 28 (9.1% of all participants) refused to adopt due to differences in the usability of 2FA in contrast to other websites, undesired/unfamiliar/custom 2FA options, or in one case due to an inconsistent device remembrance policy.

---

[1] https://www.prolific.co/

[2] https://researcher-help.prolific.co/hc/en-gb/articles/360009221093-How-do-I-use-Prolific-s-demographic-prescreening-
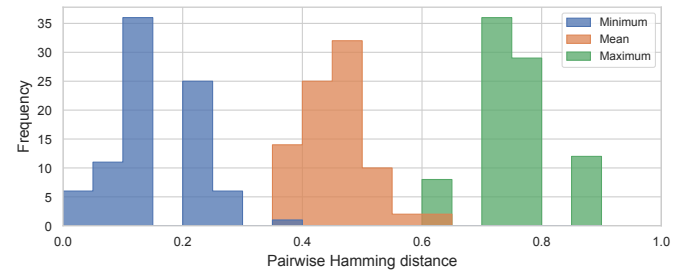
Appendix B
Details on Pairwise Hamming Distances

As explained in Section VII-B1, we compare the 85 websites in our dataset using pairwise Hamming distance between the 14 non-conditional factors of each website. Figure 3a depicts the (cumulative) frequency distribution of the pairwise Hamming distances between all websites in our data set, where a distance of 0 means equality in all factors and a distance of 1 means complete inequality of all 14 factors. This distribution is very symmetrically (skew=0.062) and only slightly heavy-tailed (kurtosis=-0.112), but an omnibus test of normality [60] ($p > .05$) indicates that it is not Gaussian. Further, Figure 3b shows each website's frequency distribution for minimum, mean, and maximum distance. The average website in our data has a minimum Hamming distance of 0.16 ± 0.02 (for a confidence interval of 95%), a mean distance of 0.46 ± 0.01, and a maximum distance of 0.75 ± 0.01. In other words, the average website in our data set differs on average in 6–7 out of 14 factors from the other websites and differs on average in at least 2–3 factors. Nevertheless, no pair of websites has a distance larger than 0.86, i.e., there are always two factors identical for each pair of websites.
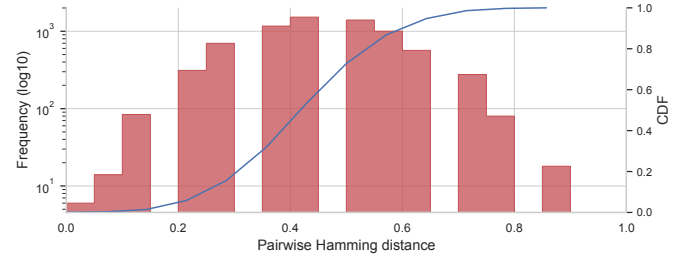

(a) Overall distribution of distances.

(b) Min, max, and median distances.

Fig. 3: Frequency distributions of pairwise Hamming distances of non-conditional comparison factors between all websites in our dataset.

Appendix C
High-level View of Clusters

Figure 4a provides a less noisy view of the clusters depicted in Table I to see the clusters' structure easily. Similarly, Figure 4b depicts only the *non-conditional* factors for which we identified six *inter-clusters*, and Figure 4c depicts only the *conditional* factors for which we found three *subclusters* or *intra-clusters*.
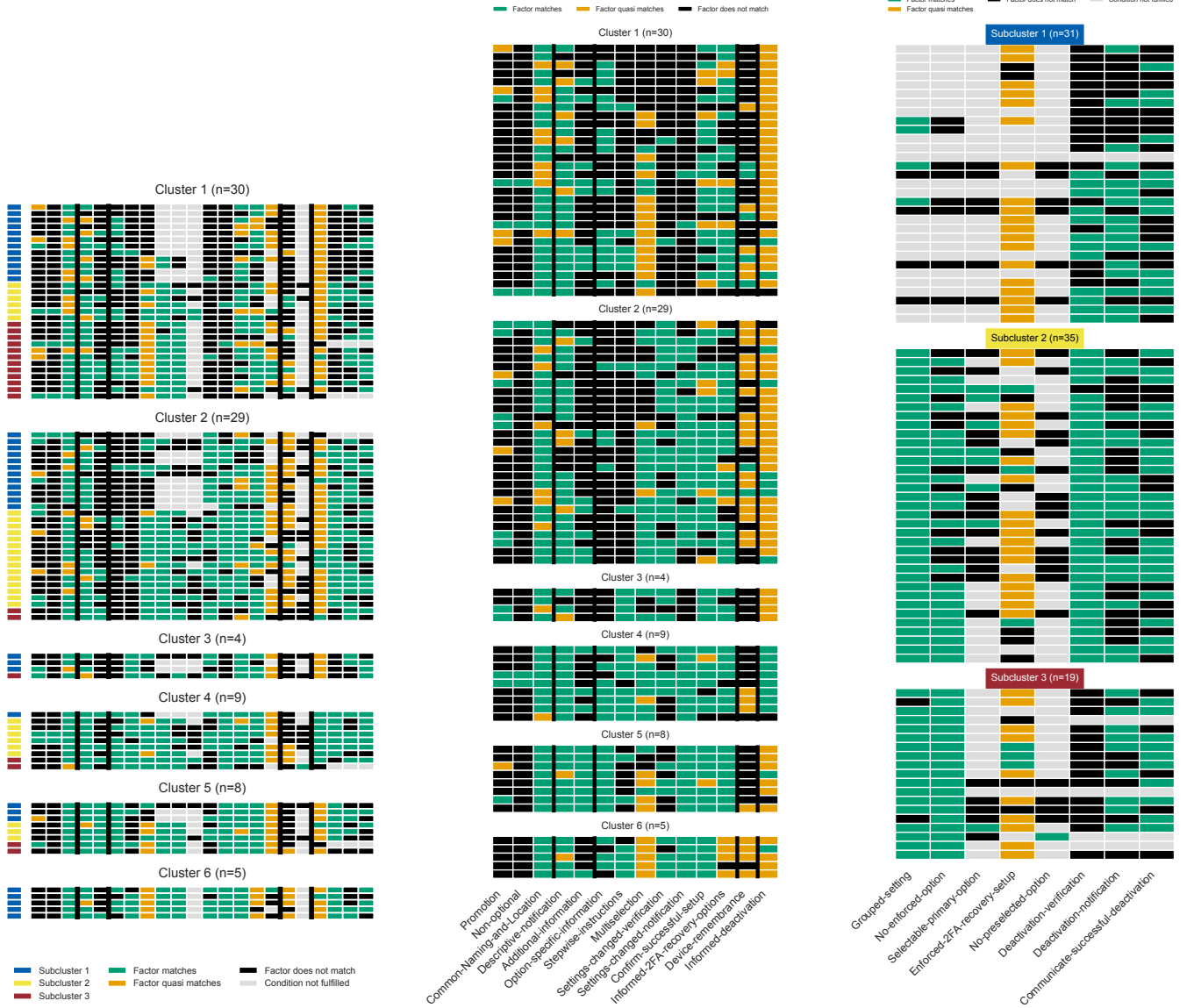
(a) Clusters of websites based on comparison factors. Subclusters based on the conditional factors are indicated in the first column. Thick lines separate factors from different steps in the user journey (see also Table I).

(b) Clusters of websites based on non-conditional factors. Only non-conditional factors are shown.

(c) Subclusters of websites based on conditional factors. Only conditional factors are shown.

Fig. 4: Clusters for all factors, only non-conditional factors, and only conditional factors.

APPENDIX D
EXAMPLES FROM WEBSITES

Figures 5, 6, and 7 illustrate examples of different aspects of the 2FA user journey from different websites. For each example, we note how it matches certain comparison factors that we identified in our study of 2FA user journeys (see Section VI). Additional examples for further comparison factors are provided in [30]

(a) tumblr.com before email verification

(b) tumblr.com after email verification

Fig. 5: tumblr.com has a *Common-Naming-and-Location* (⬤), but the security settings are initially hidden until the user verifies their email address.

Fig. 7: callcentric.com has uncommon name ("Two Point Authentication") for 2FA that places the 2FA settings at an uncommon location ("General" tab) despite the dedicated "Security" tab. (*Common-Naming-and-Location*: ☐)

Fig. 6: id.me allows users to choose their 2FA option upfront during login (*No-preselected-option*: ⬤) and supports multiple simultaneously activated 2FA options (*Multiselection*: ⬤).